Mètodes gràfics i estadístics per a la detecció de valors extrems

Tesi doctoral de

Maria Padilla Cozar

sota la supervisió del

Dr. Joan del Castillo i Franquet

Doctorat en Matemàtiques
Departament de Matemàtiques
Universitat Autònoma de Barcelona

Bellaterra, Desembre de 2015

A tots aquells que em van precedir: Joan del Castillo, Julià Cufí, Joaquim Bruna, Joaquín María Cascante, Joan Lluís Cerdà, Joan Augé, Rafael Aguiló, Ricardo San Juan, Julio Rey, Eduardo Torroja, Felix Klein, Karl Georg Christian von Staudt, Julius Plücker, Rudolf Otto Sigismund Lipschitz, Carl Friedrich Gauss, Christian Ludwig Gerling, Gustav Peter Lejeune Dirichlet, Martin Ohm, Johann Friedrich Pfaff, Simeon Denis Poisson, Jean-Baptiste Joseph Fourier, Karl Christian von Langsdorf, Abraham Gotthelf Kästner, Johann Elert Bode, Joseph Louis Lagrange, Pierre-Simon Laplace, Christian August Hausen, Johann Georg Büsch, Leonhard Euler...

Als meus pares, al meu germà,
i a en Kiko

En l'organització ideal hi ha un lloc per a cada cosa i cada cosa és al seu lloc.
Com cal escriure en matemàtiques (Paul R. Halmos)

No sé si cada cosa serà al seu lloc en
aquesta tesi, però espero haver-me apropat a
l'organització ideal tant com hagi estat possible.

Agraïments

En primer lloc, vull demanar disculpes a totes aquelles persones que no apareixen en aquest apartat però que han contribuït a que jo sigui com sóc i que hagi arribat a finalitzar aquesta tesi. Sempre es pot aprendre de la gent que t'envolta si dediques una mica del teu temps a escoltar-la.

No tinc paraules per poder expressar tot el meu agraïment al meu director de tesi, en Joan del Castillo i Franquet. Són moltes les hores que he passat al seu despatx aprenent matemàtiques, des de que era estudiant de la llicenciatura, i compartint l'experiència que suposa endinsar-se en el món de la investigació. Sempre troba la frase adequada per resumir les nostres converses i els seus consells són enriquidors.

A tot el personal del Departament de Matemàtiques de la Universitat Autònoma de Barcelona, mil gràcies. A la gent d'administració i serveis per la seva gestió, acompanyada sempre d'una bona cara, a les professores i professors per la seva dedicació, ha estat un plaer estudiar en aquesta casa i poder treballar-hi posteriorment, i als companys i companyes de doctorat.

Una menció especial a en Pere Puig, l'Alejandra Cabaña, la Isabel Serra, en Chainarong Kesamoon, l'Amanda Fernández i en Walter Andrés Ortiz, tots els records associats a vosaltres estan plens de somriures.

No em puc oblidar de la Pilar Soriano, el meu pas pel Departament d'Economia i d'Història Econòmica de la Universitat Autònoma de Barcelona i la docència impartida al Campus de Sabadell no hagués estat el mateix sense la seva gestió i els seus consells.

A totes les professores de matemàtiques que em van fer classe a l'EGB i a l'ESO i, en especial, a la Mari Luz i en Rafa, els professors de matemàtiques que vaig tenir a batxillerat. El camí ha estat dur de vegades, però no he deixat de gaudir en cap moment, tal com en Rafa em va avisar en els passadissos de l'institut quan li vaig dir que, tot i haver fet el batxillerat de ciències socials, volia estudiar matemàtiques.

Tampoc puc deixar d'esmentar als amics i amigues que sempre estan al meu costat quan els/les necessito i gràcies als quals la vida és més divertida. Les meves "floretes" de Lloret, la Mireia C., la Mireia M. i l'Eva, i els amics de la colla de la "uni", Sandra, Marc, Encarni, Raúl G-O., Carlos , Pili, Raúl G.R. i Nuria.

Per últim, però no menys important, donar les gràcies a la meva família, la de Lloret i la de Sabadell. Als meus pares, en Cristóbal i la Mari Cruz, per donar-me tot l'amor i el temps que han pogut (i més), al meu germà Joan, per fer-nos riure constantment i ampliar la nostra gran família amb la Mireia i la Júlia, a en Juan i la Lola, per obrir-me les portes de casa seva com a una filla més, i a en Kiko, perquè mai es queixa de les meves bogeries i sempre està al meu costat quan necessito una abraçada.

Maria Padilla Cozar
Bellaterra, Desembre de 2015

Índex

Introducció

La teoria de valors extrems (EVT) és l'única disciplina estadística que desenvolupa tècniques i models per descriure el comportament inusual en comptes del comportament habitual, és a dir, s'encarrega d'estudiar aquells successos que poden succeir amb freqüències d'una de cada mil vegades, o més petites, en contra del que es pot observar cada vint o cent vegades.

L'objectiu principal de la modelització d'aquestes situacions és l'estimació dels quantils corresponents a successos molt extrems. En moltes àrees d'aplicació, com el control estadístic de qualitat, les assegurances o les finances, un requisit típic és estimar el valor en risc a un cert nivell (VaR), prou alt per a que la possibilitat de superació d'aquest valor sigui menor a una quantitat donada. En camps com el de la hidrologia, la probabilitat d'una innundació inusualment gran es calcula per a períodes de 100 anys, per exemple.

La teoria probabilista associada a l'EVT es troba ben establerta, gràcies als resultats de Fisher i Tippet (1923), Balkema i de Haan (1974) i Pickands (1975). Dos enfocs principals van ser desenvolupats a partir dels resultats anteriors: el mètode per blocs de màxims, el primer en aparèixer, i el mètode d'excessos per sobre d'un llindar. L'aplicació pràctica d'aquests enfocs, però, presenta dificultats a l'hora d'aplicar eines estadístiques, veure Diebold *et al.* (1998).

La determinació del llindar a partir del qual la distribució límit pot utilitzar-se i quin és el comportament d'aquesta distribució són els problemes principals als quals s'ha de fer front. Un llindar massa alt exclou dades vàlides i comporta imprecisió, per contra, un llindar massa baix inclou valors inapropiats per explicar el comportament en el límit i comporta l'aparició de biaix, de manera que s'ha de tenir especial cura amb el balanç entre variància i biaix, veure Coles (2001).

Per distingir les dades generals de les que són objecte d'estudi en l'EVT, es farà servir

el concepte de cua, el qual fa referència a aquells valors que es troben per sobre d'un valor suficientment alt. Per als excessos per sobre d'un llindar, la distribució asimptòtica que caracteritza el comportament dela cua és la distribució de Pareto Generalitzada (GPD); el seu paràmetre de forma ξ, també anomenat índex del valor extrem, permet classificar les cues: pesades quan $\xi > 0$, exponencials per a $\xi = 0$ i lleugeres si $\xi < 0$.

L'aplicació del model GPD es pot trobar extensament detallada a McNeil *et al.* (2005), Embrechts *et al.* (1997), Reiss i Thomas (2007) o Beirlant *et al.* (2004), en els quals es tracta el tema de la selecció de llindar i l'estimació de l'índex del valor extrem amb detall, però es poden trobar situacions en les quals el model presenta limitacions amb les eines conegudes, com per exemple, quan hi ha falta d'existència de moments o la subjectivitat que pot sorgir quan es fan servir mètodes gràfics.

L'objectiu d'aquesta tesi és presentar noves eines per a l'EVT, que serveixen per a la selecció de llindar i l'estimació de l'índex del valor extrem i solucionen alguns dels problemes existents.

En el Capítol 1 es fa un repàs de la teoria estadística per a valors extrems, des dels teoremes clàssics fins a les aportacions a l'EVT més recents. Es recorden els dos mètodes gràfics més utilitzats, el *mean excess plot*, Davison i Smith (1990), i el *Hill-plot*, Dress *et al.* (2000), i també els mètodes d'estimació disponibles per a la GPD, finalment es presenten un nou mètode gràfic anomenat *CV-plot*, Castillo *et al.* (2014), i un enfoc aparegut recentment per a cues pesades i exponencials, Castillo *et al.* (2013).

En el Capítol 2 s'utilitza el fet que el coeficient de variació residual caracteritza distribucions, veure Gupta i Kirmani (2000), per trobar el *CV-plot* teòric per a algunes distribucions concretes, el qual es podria fer servir per a bondat de l'ajust, i ampliar la teoria asimptòtica del cas exponencial a qualsevol GPD, sempre que hi hagi existència de moments d'ordre quatre.

Per solucionar el problema de la falta d'existència de moments, es presenta una solució en el Capítol 3. Mitjançant una transformació adequada, és possible convertir cues pesades, les quals poden tenir falta d'existència de moments, en cues lleugeres i viceversa. D'aquesta manera, el *CV-plot* es podrà aplicar en qualsevol situació.

En el Capítol 4 es presenten uns estadístics, anomenats T_m, que permeten treballar a diferents nivells; estimant l'índex del valor extrem, contrastant la hipòtesi de GPD i com a part d'un algoritme automàtic de selecció de llindars que també retorna l'estimació de ξ. Aquest tercer nivell suposa un gran avenç respecte a l'aplicació habitual del mètode conegut com a *Peak over threshold*, PoT, el qual requereix de l'estimació del llindar, habitualment a partir d'un mètode gràfic, i l'estimació de l'índex del valor extrem, mitjançant màxima versemblança, veure Coles (2001), ja que desapareix la subjectivitat de l'investigador a l'hora d'estimar el llindar fent servir un mètode gràfic.

El Capítol 5 es troba dedicat a l'estudi de 16 conjunts de dades de temps real en sistemes informàtics encastats. El *CV-plot* i els estadístics T_m, a nivell de contrast

i d'algoritme, han estat utilitzats, obtenint bons resultats per a la majoria dels casos. Aquest estudi s'ha realitzat gràcies a un Projecte de Recerca conjunt entre la Universitat Autònoma de Barcelona i el *Barcelona Supercomputing Center* - Centro Nacional de Supercomputación (BSC), l'objectiu del qual és desenvolupar l'EVT en ciències de la computació.

En el Capítol 6 es tornen a aplicar les noves eines a dades estudiades anteriorment. En el primer cas, s'utilitzaran les dades daneses sobre assegurances de focs, veure McNeil (1997), les quals es fan servir com a exemple al llarg de tota la tesi i s'acaben d'estudiar en aquest apartat, en el segon, s'analitzaran les dades financeres estudiades a Gomes i Pestana (2007).

Per finalitzar, en el Capítol 7 es presenten les conclusions que es dedueixen de l'aplicació dels nous mètodes gràfics i estadístics proposats i la comparació amb els ja existents. També s'exposen les noves línies de recerca que poden ser objecte d'estudi.

Cal destacar que les aportacions principals d'aquesta tesi han donat lloc a quatre articles, alguns encara pendents de publicar, en els àmbits de l'Estadística i les Ciències de la Computació:

Castillo, J., Padilla, M. i Serra, I. (2014) *Comparison of techniques for extreme values using financial data*, Proceedings of the 21st International Conference on Computational Statistics hosting the 5th IASC World Conference, COMPSTAT 14, pp.45-52.

Abella, J., Padilla, M., Castillo, J. i Cazorla, F.J. (2015). *Relating Processor Design, Execution Time Distributions and Extreme Value Theory.*

Padilla, M., Castillo, J., Abella, J. i Cazorla, F.J. (2015). *Execution time distributions in embedded safety-critical systems using extreme value theory.*

Castillo, J. i Padilla, M. (2015). *Modeling extreme values by the residual coefficient of variation*, arXiv:1510.00179 [math.ST].

CAPÍTOL 1

Teoria estadística per a valors extrems

La teoria dels valors extrems (EVT) s'ha anat consolidant des de mitjans del segle passat com una branca de l'estadística que es pot aplicar en diversos camps com els de les finances (McNeil *et al.*, 2005), assegurances (Embrechts *et al.*, 1997), meteorologia (Thompson *et al.*, 2001), hidrologia (Reiss i Thomas, 2007) o metal·lúrgia (Beirlant *et al.*, 2004), entre d'altres.

L'objectiu principal consisteix en l'estimació de quantitats relacionades amb esdeveniments més grans que qualsevol dels que s'han observat prèviament, com poden ser quantils molt elevats o la probabilitat d'un esdeveniment per sobre d'un valor inusualment gran. En altres paraules, es realitza una extrapolació dels valors observats a valors no observats, la qual ve donada per arguments asimptòtics.

En aquest capítol es presentaran els enfocs clàssics que s'apliquen a l'estudi dels valors extrems, es comentaran els mètodes gràfics més utilitzats en aquest camp i es parlarà sobre l'estimació de paràmetres i les darreres novetats en aquesta teoria.

1.1 Teoria clàssica per a màxims

Existeixen dos enfoncs clàssics en la teoria dels valors extrems, el primer, i més antic, es basa en la distribució del màxim, mentre que el segon utilitza la distribució dels excessos per sobre d'un llindar. És possible trobar models asimptòtics per a cada un dels enfocs que, a més a més, es troben estretament relacionats mitjançant el paràmetre de forma de cada una de les distribucions límit.

Els dos primers apartats es troben dedicats a presentar tots els elements necessaris per enunciar els dos teoremes clàssics de la teoria i descriure la relació existent entre els

dos enfocs.

Els models asimptòtics que donen lloc a l'estudi dels màxims per blocs provenen del primer teorema clàssic de Fisher i Tippett (1928), del qual Gnedenko (1948) va donar una demostració més acurada.

Sigui

$$M_n = max(x_1, ..., x_n),$$

on $x_1, ..., x_n$, és un seqüència de variables aleatòries independents i idènticament distribuides (i.i.d.) amb funció de distribució F. La distribució de M_n es pot deduir fàcilment de la següent manera:

$$\begin{aligned} P(M_n \leq z) &= P(x_1 \leq z, ..., x_n \leq z) \\ &= P(x_1 \leq z)P(x_2 \leq z) \ ... \ P(x_n \leq z) = (F(z))^n. \end{aligned} \tag{1.1}$$

A la pràctica, donat que la funció de distribució F és desconeguda, també ho serà la distribució del màxim. Una opció per estimar-la és trobar primer una estimació de la funció F i llavors substituir aquesta dins de (1.1), però aquesta via no és gaire adequada perquè petites variacions en F poden produir grans variacions en F^n. Una alternativa a aquest procediment és buscar models per a F^n directament.

L'estudi de la distribució asimptòtica del màxim és un problema semblant al de la convergència per a la suma, caracteritzat pel teorema central del límit. Així doncs, sota una apropiada normalització, és possible trobar una distribució límit no degenerada per al màxim, ja que, sense aquesta normalització, qualsevol valor z del domini degenera a un únic punt donat que $F^n(z) \to 0$ quan $n \to \infty$.

Els dos elements claus per a aquesta normalització són la distribució generalitzada de valors extrems i el concepte de domini d'atracció del màxim, els quals es definiran a continuació.

Definició 1.1. Distribució generalitzada de valors extrems (GEV). La funció de distribució d'una GEV estàndard ve donada per

$$H_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-1/\xi}), & \xi \neq 0, \\ \exp(-e^{-x}), & \xi = 0, \end{cases} \tag{1.2}$$

on $\xi \in \mathbb{R}$ és el paràmetre de forma i $1 + \xi x > 0$. També és possible obtenir una família triparamètrica de distribucions definint

$$H_{\xi,\mu,\sigma}(x) = H_\xi((x-\mu)/\sigma) \tag{1.3}$$

on $\mu \in \mathbb{R}$ i $\sigma > 0$ són els paràmetres de localització i escala, respectivament.

El paràmetre de forma ξ, també conegut com a *índex del valor extrem*, determinarà el tipus de distribució, dels tres possibles que hi ha, dins d'aquesta família. Per a $\xi > 0$

la distribució és Fréchet, quan $\xi = 0$ la distribució és Gumbel i, finalment, si $\xi < 0$ la distribució obtinguda és una Weibull. Cal remarcar la continuïtat en ξ de la funció definida en (1.2), és fàcil veure que $\lim_{\xi \to 0} H_\xi(x) = H_0(x)$ per a qualsevol x fixat, fet que facilita l'ús d'aquesta distribució per a la seva implementació.

Les funcions de distribució i de densitat per a exemples dels tres tipus de distribucions anteriors es troben representades a la Figura 1.1. Es pot observar que el suport de la densitat de la distribució de Weibull és acotat superiorment mentre que les distribucions de Fréchet i Gumbel tenen densitats amb suport no acotat a la dreta.

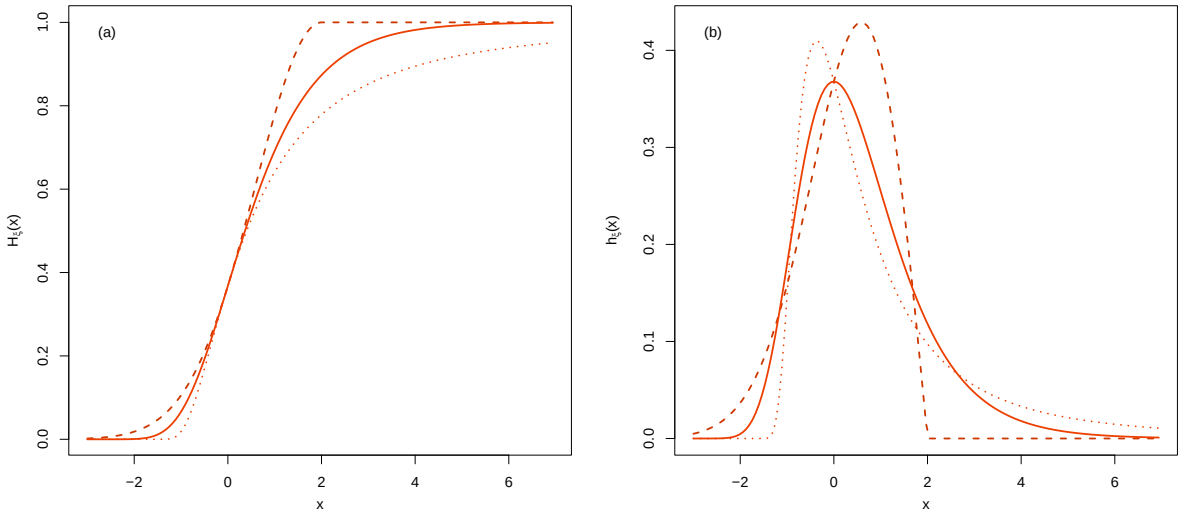

Figura 1.1: (a) Funció de distribució per a la GEV estàndar en tres casos: la línia sòlida correspon a $\xi = 0$ (Gumbel), la línia puntejada a $\xi = 0.5$ (Fréchet) i la línia discontínua a $\xi = -0.5$ (Weibull). (b) Funcions de densitat corresponents. En tots tres exemples $\mu = 0$ i $\sigma = 1$.

Definició 1.2. Domini d'atracció del màxim. Sigui F una funció de distribució i M_n el seu bloc de màxims. Si existeixen successions de nombres reals (d_n) i (c_n), on $c_n > 0$ per a qualsevol n, tals que

$$\lim_{n\to\infty} P((M_n - d_n)/c_n \leq x) = \lim_{n\to\infty} F^n(c_n x + d_n) = H(x) \tag{1.4}$$

per a alguna funció de distribució no-degenerada, llavors es diu que F pertany al domini d'atracció del màxim de H, denotat per $F \in D(H)$.

Per comoditat, quan es treballi amb la GEV, es farà servir la notació de domini d'atracció del màxim $F \in D(H_\xi)$, on ξ correspon a l'índex del valor extrem d'aquesta distribució.

Es poden trobar condicions suficients per a que una certa distribució F pertanyi al domini d'atracció del màxim de les distribucions Fréchet o Weibull; s'introduiran alguns conceptes previs, com el de funció de variació regular, per poder enunciar-les.

Definició 1.3. ***Tail function.*** Donada una funció de distribució F, es defineix la seva *tail function* com

$$\overline{F}(x) = 1 - F(x), \tag{1.5}$$

i s'entèn que la cua de la distribució és el comportament de $\overline{F}(x)$ quan $x \to \infty$. Es parlarà de cua esquerra de la distribució quan es faci referència al mateix concepte però per a valors de x tendint cap a l'inici del domini de definició.

Definició 1.4. Funció de variació regular. Es diu que una funció mesurable h entre $(0, +\infty)$ i $(0, +\infty)$ és de variació regular a infinit amb índex $\rho \in \mathbb{R}$, i es denotarà per $h \in RV_\rho$, si $\forall x > 0$,

$$\lim_{t\to\infty} \frac{h(tx)}{h(t)} = x^\rho. \tag{1.6}$$

Si $\rho = 0$, llavors h s'anomena funció de variació lenta (a infinit). Les funcions de variació lenta són denotades generalment per L. Si $h \in RV_\rho$, llavors $h(x)/x^\rho \in RV_0$. Per tant, agafant $L(x) = h(x)/x^\rho$ es pot veure que la funció $h \in RV_\rho$ es pot representar com $h(x) = x^\rho L(x)$.

Una variable aleatòria no negativa es diu que és de variació regular si la seva funció de distribució F satisfà $\overline{F} \in RV_{-\alpha}$ per a cert $\alpha \geq 0$.

Definició 1.5. Funció quantil. Sigui F una funció de distribució. La funció quantil Q es defineix com

$$Q(p) = inf\{x \,:\, F(x) \geq p\},\ 0 < p < 1. \tag{1.7}$$

Definició 1.6. ***Punt final*** **d'una distribució.** Donada una funció de distribució F, el seu *punt final* es denota per

$$x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}. \tag{1.8}$$

Amb tots aquests conceptes, ja es poden presentar les condicions que s'han de satisfer per a que una distribució pertanyi al domini d'atracció del màxim d'una distribució Fréchet o d'una distribució Weibull.

Proposició 1.1. *(Domini d'atracció del màxim de la distribució Fréchet). La funció de distribució F pertany al domini d'atracció del màxim de $H_\xi(x)$ $(\xi > 0)$ si i només si $\overline{F} \in RV_{-1/\xi}$. Si $F \in D(H_\xi),\ \xi > 0$, llavors*

$$a_n^{-1} M_n \xrightarrow{d} H_\xi, \tag{1.9}$$

amb $a_n = Q(1 - n^{-1})$.

La demostració de la Proposició 1.1 es pot trobar a l'apartat 3.3.1 de Embrechts *et al.* (1997), amb la notació Φ_α per a la distribució Fréchet ja que

$$F \in D(H_\xi) \text{ amb } \xi > 0 \text{ si i només si } F \in D(\Phi_\alpha) \text{ amb } \alpha = 1/\xi > 0.$$

Les distribucions que donen lloc al cas Fréchet són aquelles amb densitats no acotades i cues que són funcions de variació regular amb índex negatiu. Les cues decreixen polinomialment amb una taxa de decreixement $\alpha = 1/\xi$, valor que es coneix habitualment com a *índex de la cua* de la distribució.

Proposició 1.2. *(Domini d'atracció del màxim de la distribució Weibull). La funció de distribució F pertany al domini d'atracció del màxim de $H_\xi(x)$ ($\xi < 0$) si i només si $x_F < \infty$ i $\overline{F}(x_F - x^{-1}) \in RV_{1/\xi}$. Si $F \in D(H_\xi)$, $\xi < 0$, llavors*

$$a_n^{-1}(M_n - x_F) \xrightarrow{d} H_\xi, \tag{1.10}$$

amb $a_n = x_F - Q(1 - n^{-1})$.

La demostració de la Proposició 1.2 es pot trobar a l'apartat 3.3.2 de Embrechts *et al.* (1997), amb la notació Ψ_α per a la distribució Weibull ja que

$$F \in D(H_\xi) \text{ amb } \xi < 0 \text{ si i només si } F \in D(\Psi_\alpha) \text{ amb } \alpha = -1/\xi > 0.$$

Les distribucions que donen lloc al cas Weibull són aquelles amb densitats acotades superiorment i que tenen cues que són funcions de variació regular amb índex negatiu quan són avaluades sota la transformació $x_F - x^{-1}$, on x són els valors del domini.

Per al cas de la distribució Gumbel no hi ha una relació directa entre el domini d'atracció del màxim i la variació regular, però és possible trobar extensions per a una caracterització completa, per a més detalls sobre aquesta caracterització, veure Embrechts *et al.* (1997).

Ara que ja es tenen caracteritzats els dominis d'atracció del màxim de les distribucions de la família GEV, la pregunta que sorgeix és si hi ha distribucions diferents a aquestes tres que es puguin caracteritzar com a domini d'tracció del màxim d'altres distribucions. El primer teorema clàssic conté la resposta.

Teorema 1.1. ***(Fisher-Tippett, Gnedenko).*** *Si una certa funció de distribució F pertany al domini d'atracció del màxim d'una distribució H no-degenerada, llavors H ha de ser del tipus $H_\xi(x)$, és a dir, una GEV.*

La característica principal d'aquest resultat és que els tres tipus de distribucions inclosos dins de la GEV són els únics límits possibles per a la distribució del màxim normalitzada. És en aquest sentit que el paper de la GEV és per a valors extrems

anàleg al de la distribució normal com a límit de la suma normalitzada. Per consultar una demostració formal del Teorema 1.1, veure Leadbetter *et al.* (1983).

A la pràctica, la modelització d'extrems d'una succesió de variables independents $x_1, x_2, \ ..., x_{nm}$ es realitza de la següent manera. Es generen blocs de mida n, per a algun valor gran de n, i es genera una nova sèrie de blocs de màxims, $M_{n,1}, ..., M_{n,m}$, la qual s'ajustarà per una GEV. Normalment, els blocs són escollits per a un cert període de temps fixat, per exemple un any o un mes, de manera que el valor n correspon a les observacions dins d'aquest període, obtenint els màxims anuals o mensuals.

L'argument que permet fer servir la mostra de màxims sense haver-los de normalitzar és el següent; suposem que es satisfà (1.4),

$$P((M_n - d_n)/c_n \leq x) \approx H(x)$$

per a un n suficientment gran. Equivalentment,

$$\begin{aligned} P(M_n \leq x) &\approx H((x - d_n)/c_n) \\ &= H^*(x), \end{aligned} \tag{1.11}$$

on H^* torna a ser una GEV.

Per tant, com que els paràmetres per a una GEV s'hauran d'estimar de totes maneres, no cal perdre temps en buscar els paràmetres de normalització i es pot passar directament a estimar els corresponents paràmetres dels blocs de màxims.

Un punt feble de la modelització per blocs de màxims és que es perden molts valors, de manera que, si la mida dels blocs és gran i la de la mostra inicial petita, la mostra final pot arribar a ser poc representativa. En aquest sentit, l'elecció de la mida del bloc és un punt clau, una mida de bloc petita presenta una situació lluny del límit i, si la mida del bloc és gran, la variància en les estimacions dels paràmetres també ho és.

De manera semblant, es pot realitzar l'estudi asimptòtic de la distribució del mínim. En casos com temps mínim de fallida d'un sistema o rècords esportius, aquesta és la distribució d'estudi, la qual es pot deduir a partir de la distribució del màxim de les dades oposades.

Sigui

$$\widetilde{M_n} = min(x_1, ..., x_n) = max(-x_1, ..., -x_n),$$

on $x_1, ..., x_n$, és una mostra i.i.d. amb funció de distribució F. La distribució de $\widetilde{M_n}$ es pot deduir fàcilment de la següent manera:

$$P(\widetilde{M_n} \leq z) = P(-M_n \leq z) = P(M_n \geq -z) = 1 - P(M_n \leq -z). \tag{1.12}$$

A la pràctica, ajustant per una GEV el màxim de les dades oposades queda definida la distribució del mínim aplicant (1.12).

1.2 Models sobre un llindar

El segon enfoc utilitzat per a l'estudi de valors extrems es basa en l'estudi dels valors a partir d'un cert llindar. Van ser Balkema i de Haan (1974) i Pickands (1975) els que van desenvolupar la teoria asimptòtica que permet l'estudi de les probabilitats per sobre d'un llindar.

Tot i que aquesta metodologia també elimina dades de la mostra inicial, en general, no són tantes com quan s'utilitza el mètode del màxim per blocs, fet que fa atractiu l'ús d'aquest enfoc enfront del descrit a l'apartat anterior.

Considerem una variable aleatòria (v.a.) X amb funció de distribució F, la distribució sobre els excessos a partir d'un llindar t, F_t, es pot obtenir a partir de la probabilitat condicionada de la següent manera:

$$F_t(y) = P(X - t \leq y \mid X > t) = \frac{F(t+y) - F(t)}{1 - F(t)},\ 0 \leq y \leq x_F - t. \tag{1.13}$$

Com en la secció anterior, la distribució F serà desconeguda a la pràctica, en cas contrari només caldria substituir-la dins de l'equació (1.13) i la distribució condicionada quedaria definida. Per resoldre el problema, es buscaran aproximacions per a valors alts del llindar t, és en aquest punt on la distribució de Pareto generalitzada serà crucial.

Definició 1.7. Distribució de Pareto generalitzada (GPD). La funció de distribució d'una GPD ve donada per

$$G_{\xi,\psi}(x) = \begin{cases} 1 - (1 + \xi x/\psi)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-x/\psi), & \xi = 0, \end{cases} \tag{1.14}$$

on $\xi \in \mathbb{R}$ i $\psi > 0$ són els paràmetres de forma i escala, respectivament, $0 \leq x \leq -\psi/\xi$ quan $\xi < 0$ i $x \geq 0$ quan $\xi \geq 0$.

El valor del paràmetre ξ de la GPD determinarà el tipus de distribució, dels tres possibles que hi ha, dins d'aquesta família, tal com passa amb la GEV. Per a $\xi > 0$ la distribució és la ben coneguda distribució de Pareto, quan $\xi = 0$ la distribució és exponencial i, finalment, si $\xi < 0$ la distribució obtinguda és de suport compacte. Cal remarcar la continuïtat en ξ de la funció definida en (1.14); es té $\lim_{\xi \to 0} G_{\xi,\psi}(x) = G_{0,\psi}(x)$ per a qualsevol x fixat, fet que facilita l'ús d'aquesta distribució per a la seva implementació.

L'esperança d'una GPD és $\psi/(1-\xi)$ i la variància $\psi^2/[(1-\xi)^2(1-2\xi)]$ donats $\xi < 1$ i $\xi < 1/2$, respectivament. Hi ha una limitació d'existència de moments quan $\xi > 0$, més concretament, per a la distribució de Pareto hi ha existència de moments finits fins a ordre n sempre que $\xi < 1/n$. Aquest fet pot provocar no poder aplicar algun dels mètodes gràfics que es presentaran més endavant, però no suposarà cap problema un

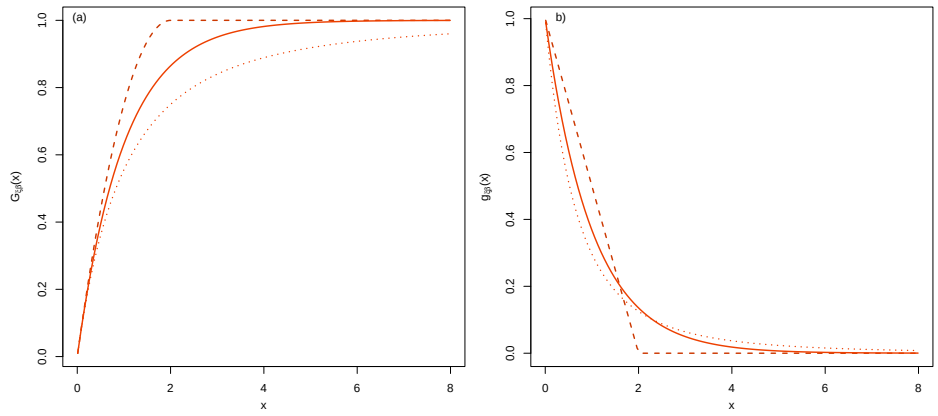

Figura 1.2: (a) Funció de distribució per a la GPD en tres casos: la línia sòlida correspon a $\xi = 0$ (exponencial), la línia puntejada a $\xi = 0.5$ (Pareto) i la línia discontínua a $\xi = -0.5$ (suport compacte). (b) Funcions de densitat corresponents. En tots tres exemples $\psi = 1$.

cop definides, en capítols posteriors, les transformacions que permeten obtenir varibales aleatòries amb moments finits de tots els ordres.

Direm que una distribució té cua pesada quan la seva cua no està acotada superiorment per la cua de la distribució exponencial i que té cua lleugera quan la seva cua sí que ho està. Per al cas de la GPD, quan es té $\xi > 0$, les cues són pesades i, quan $\xi < 0$, les cues són lleugeres.

Les funcions de distribució i de densitat per a exemples dels tres tipus de distribucions anteriors es troben representades a la Figura 1.2. Es pot observar que la distribució de suport compacte té cua lleugera i un punt final finit. Les distribucions exponencial i Pareto tenen punts finals infinits, però el decreixement de la distribució de Pareto, que és polinomial, és molt més lent que el de la distribució exponencial. La distribució de Pareto té, clarament, una cua pesada.

Teorema 1.2. ***(Pickands-Balkema-de Haan).*** *Existeix una funció $\psi(t)$ tal que*

$$\lim_{t \to x_F} \sup_{0 \leq x < x_F - t} |F_t(x) - G_{\xi,\psi(t)}(x)| = 0, \tag{1.15}$$

si i només si $F \in D(H_\xi)$. A més a més, $H_\xi = H_{\xi,\mu,\sigma}(y)$, $y \in \mathbb{R}$, amb $\psi(t) = \sigma + \xi(t - \mu)$.

El principal resultat que es desprèn d'aquest teorema és que el límit per als excessos sobre un llindar es pot aproximar per una GPD a partir d'un llindar suficientment gran. Per consultar una demostració formal del Teorema 1.2, veure Leadbetter *et al.* (1983).

El mateix teorema ens indica també que, si els blocs de màxims es poden aproximar per una GEV amb paràmetre de forma ξ, llavors els excessos per sobre d'un llindar es

poden aproximar per una GPD amb el mateix paràmetre de forma. Existeix, doncs, una dualitat entre les famílies GEV i GPD donada pel paràmetre de forma ξ.

Es pot deduir fàcilment que la família GPD és tancada per condicionament dels excessos sobre un llindar, és a dir, quan una GPD es trunca i es trasllada a l'origen, la nova distribució torna a ser GPD.

Proposició 1.3. *Sigui F la distribució d'una GPD, és a dir, $F = G_{\xi,\psi}(x)$ per a algun $\xi \in \mathbb{R}$ i $\psi > 0$, llavors*

$$F_t(x) = G_{\xi,\psi+\xi t}(x) \tag{1.16}$$

on $0 \leq x < \infty$ si $\xi \leq 0$ i $0 \leq x \leq -\psi/\xi$ si $\xi < 0$.

Demostració. El resultat es dedueix a partir de l'equació (1.13). □

A la pràctica, s'escull un llindar prou gran i els excessos per sobre d'aquest s'ajusten per una GPD. Un dels mètodes més habituals per a l'elecció del llindar i l'ajust és el conegut com a *Peak over Threshold*, el qual es descriu a continuació.

1.3 Peak over Threshold (PoT)

El mètode conegut com a *Peak over Threshold* (PoT) ofereix un enfoc per modelitzar la cua d'una distribució. Es basa en els resultats obtinguts a partir del Teorema (1.2), del qual es dedueix que la GPD és la distribució que s'ha d'utilitzar per modelitzar valors per sobre d'un llindar suficientment gran, sempre que es disposi d'una mostra prou representativa per poder fer prediccions precisses després d'agafar els valors per sobre d'aquest.

La primera part del mètode consisteix en escollir un llindar adequat, el qual es determina habitualment mitjançant un mètode gràfic, com poden ser el *mean excess plot*, Davison i Smith (1990), o el *Hill-plot*, Dress *et al.* (2000). Aquest és un punt feble del model, ja que un mateix gràfic pot ser interpretat de manera diferent per cada persona i això afecta a l'elecció del punt òptim del llindar.

La segona part consisteix en agafar tots els valors de la mostra original que superen el llindar escollit, traslladar-los a l'origen calculant $(X - t \mid x \geq t)$ i ajustar per una GPD, que es denotarà per $G_{\xi,t,\psi}(x)$. D'entre els mètodes disponibles per poder ajustar una GPD destaquen el de màxima versemblança (ML) i el conegut com a *probability weighted moments* (PWM), veure Castillo i Serra (2015).

El darrer pas consisteix en donar l'expressió per a la distribució de la cua. Per a punts de la cua de la distribució es satisfà

$$F(x) = P\{X \leq x\} = (1 - P\{X \leq t\})F_t(x - t) + P\{X \leq t\}. \tag{1.17}$$

Com a estimació de $F_t(x-t)$ es farà servir $G_{\xi,t,\psi}(x)$ i per a $P\{X \leq t\}$ el valor de la funció de distribució empírica avaluat en t, $F_n(t)$. Per tant, per a $x \geq t$ l'estimació de la cua que aproxima $F(x)$ és

$$\widehat{F(x)} = (1 - F_n(t))G_{\xi,t,\psi}(x) + F_n(t). \tag{1.18}$$

Per a valors $x \leq t$, l'aproximació que es farà servir per a la seva distribució és la de la pròpia distribució empírica, $F_n(x)$.

Per veure com funciona amb més detall la primera part del mètode, a continuació es descriuran els dos mètodes gràfics més utilitzats, el *mean excess plot* i el *Hill-plot*.

El *mean excess plot* és una eina gràfica, molt utilitzada en l'estudi de valors extrems, que serveix per escollir el valor del llindar t a partir del qual es considerarà la cua de la distribució i per determinar l'adequació del model GPD a la pràctica. Es fa servir normalment en el primer pas del mètode PoT per aquest motiu, com ja s'ha comentat abans.

Considerem la funció de la mitjana dels excessos, *mean excess* (ME), sobre una variable aleatòria X definida com

$$t \to ME(t) := E(X - t \mid X > t), \tag{1.19}$$

sempre que $E(X \mid X > 0) < \infty$.

Donada una mostra i.i.d. $x_1, x_2, ..., x_n$ amb distribució F, un estimador de $ME(t)$ és la funció ME empírica, $me(t)$, definida com

$$t \to me(t) = \sum_{i=1}^{n} (x_i - t) I_{[x_i > t]} / \sum_{i=1}^{n} I_{[X_i > t]},\ t \geq 0. \tag{1.20}$$

Ja que el cas que interessa estudiar és la GPD, anem a veure quina informació es pot extreure en aquest cas. Per a una variable $X \sim G_{\xi,\psi}$, la condició necessària d'esperança finita per a valors positius només es compleix si $\xi < 1$ i, a més a més, la funció ME és lineal en t:

$$ME(t) = \frac{\psi}{1-\xi} + \frac{\xi}{1-\xi} t, \tag{1.21}$$

on $0 \leq t < \infty$ si $0 \leq \xi < 1$ i $0 \leq t \leq -\psi/\xi$ si $\xi < 0$.

Aquesta linealitat de la funció ME és en la que es van basar Davison i Smith (1990) per crear una comprovació gràfica per al model GPD, anomenada *ME-plot*. Aquest mètode consisteix en dibuixar el conjunt de punts $\{(x_{(k)}, me(x_{(k)})) : 1 \leq k < n\}$, on $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ corresponen als estadístics d'ordre de la mostra. Si el gràfic és prou lineal a partir d'un cert punt, s'escollirà aquest com a llindar t i es podria extreure informació del paràmetre ξ fent servir el pendent de la gràfica a partir d'aquest lloc.

A la Figura 1.3 es pot veure l'*ME-plot* de 2.156 dades sobre pèrdues en assegurances per focs per sobre d'un milió de corones daneses des de 1980 fins a 1990, inclós. Com

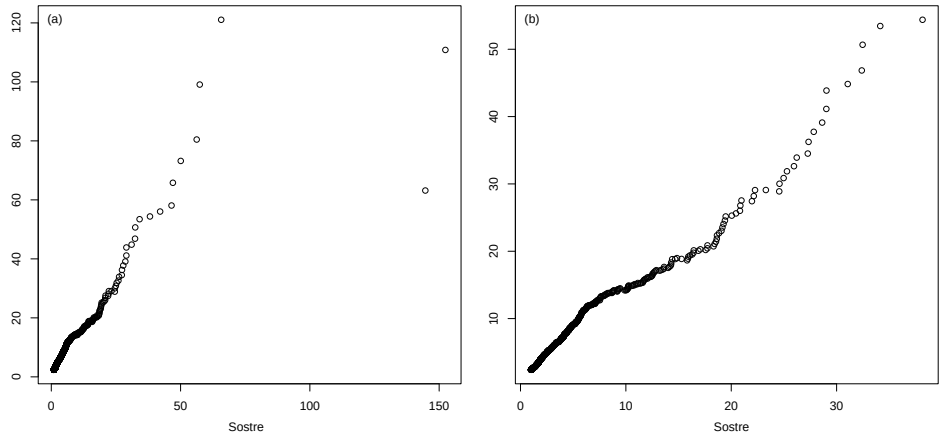

Figura 1.3: (a) *Mean excess plot* de les dades daneses sobre pèrdues en assegurances de focs. (b) *Mean excess plot* per a les mateixes dades sense considerar els 10 estadístics d'ordre més grans de la mostra.

que per a mostres petites el valor de la mitjana residual pot donar valors extranys, és recomanable dibuixar l'*ME-plot* treient aquests valors per poder observar millor el gràfic, tal i com es pot veure en comparar les dues imatges de la figura. Els dos gràfics s'han realitzat amb la funció `meplot` del paquet `evir` de R, veure R Development Core Team (2010).

Aquestes dades s'estudiaran amb detall més endavant, però una primera ullada a aquest gràfic pot indicar que a partir dels 10 milions de pèrdues es podria considerar una cua GPD amb índex del valor extrem positiu, ja que s'observa una recta creixent a partir d'aques punt.

El *Hill-plot* és l'altre mètode gràfic utilitzat habitualment en l'estudi de valors extrems. Serveix per decidir quin és el punt de tall a partir del qual comença la cua i per estimar l'índex del valor extrem que determina el seu comportament, de la mateixa manera que passa amb l'*ME-plot*.

L'estimador de Hill basat en els $k+1$ estadístics d'ordre majors es defineix com

$$H_{k,n} := \frac{1}{k}\sum_{i=1}^{k} log\left(\frac{X_{(n-i+1)}}{X_{(n-k)}}\right),\ k = 1, ..., n-1, \tag{1.22}$$

el qual és un estimador del paràmetre ξ.

Cal observar que els sumands que apareixen a la fórmula anterior són sempre positius, de manera que aquest estimador només té sentit per a l'estudi de cues pesades, és a dir, quan $\xi > 0$. Per tant, aplicat fora d'aquesta situació, els resultats poden no ser coherents, per exemple, les cues de dades financeres es consideren pesades habitualment, però es poden trobar casos en els quals no és evident, com en alguns tipus de canvi. Serà

recomanable realitzar, com a mínim, un test per decidir entre exponencialitat i GPD per contrastar l'adequació del model.

Per decidir quin és el valor k òptim, es representa el *Hill-plot* definit com

$$\{(k, H_{k,n}^{-1}), 1 \leq k \leq n-1\} \tag{1.23}$$

i a continuació s'escull com a llindar el valor a partir del qual el gràfic s'estabilitza, podent estimar el valor de ξ també a través d'aquesta regió on el gràfic sembla més o menys constant.

Aquest mètode presenta algunes dificultats degut a que el gràfic no sempre mostra una regió prou gran a partir de la qual hi ha estabilitat i, principalment, a que és més efectiu només quan la distribució d'estudi és Pareto o molt propera a la Pareto. Es pot observar també un fort biaix per a trajectòries de la mostra quan el model no és estrictament Pareto. Per a més detalls sobre els inconvenients del *Hill-plot* veure Drees *et al.* (2000) i Beirlant *et al.* (1999).

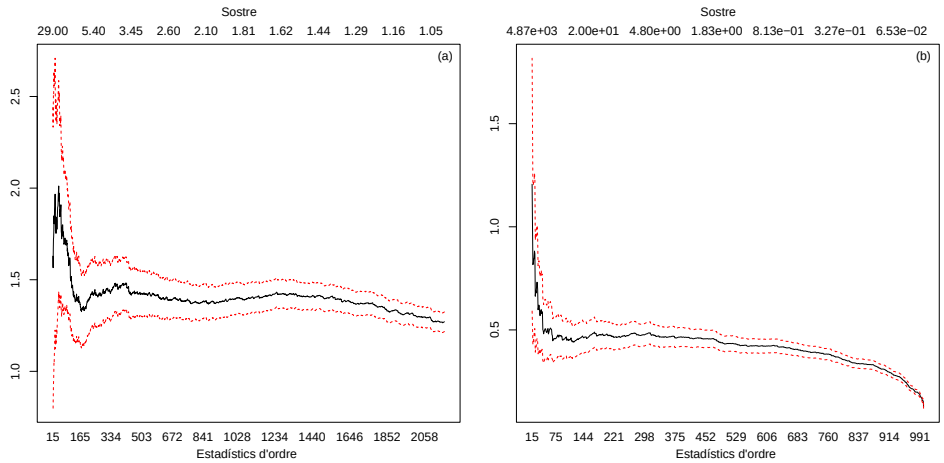

Figura 1.4: (a) *Hill-plot* de les dades daneses sobre pèrdues en assegurances de focs. (b) *Hill-plot* sobre una mostra de mida 1000 d'una GPD amb paràmetre $\xi = 2$. En tots dos casos les línies vermelles puntejades indiquen l'interval de confiança del 95%.

A la Figura 1.4, a l'apartat (a), es pot veure el *Hill-plot* per a les dades daneses, que indicaria una cua amb índex del valor extrem igual a l'invers de 1,4 (0,71 aproximadament), ja que la gràfica s'estabilitza al voltant d'aquest valor. A l'apartat (b), s'ha representat el *Hill-plot* d'una mostra simulada de mida 1.000 d'una GPD amb índex del valor extrem $\xi = 2$, s'observa com la gráfica s'estabilitza al voltant del valor 0,5, que és l'invers de 2, cal recordar que aquest mètode funciona molt bé quan la distribució és Pareto. Els dos gràfics s'han realitzat amb la funció `hill` del paquet `evir` de R.

Diferents alternatives han aparegut per intentar millorar aquest mètode, algunes d'elles basades en el desenvolupament de segon ordre de la cua de la funció $1 - F$.

Sigui $U(t)$ la inversa de la funció $1/(1-F)$. Suposem que aquesta funció pertany a la classe de models Hall-Welsh, és a dir,

$$U(t) = Ct^{\xi}\left(1 + \frac{A(t)}{\rho} + o(t^{\rho})\right),\ A(t) = \xi\beta t^{\rho} \tag{1.24}$$

quan $t \to \infty$ amb $C > 0, \xi > 0, \rho < 0$ i $\beta \neq 0$.

Un nou estimador per a l'índex de la cua es pot definir de la següent manera

$$\overline{H}(k) \equiv \bar{H}_{\widehat{\beta},\widehat{\rho}}(k) := H(k)\left(1 - \frac{\widehat{\beta}}{1-\widehat{\rho}}\left(\frac{n}{k}\right)^{\widehat{\rho}}\right), \tag{1.25}$$

on $(\widehat{\beta}, \widehat{\rho})$ són estimadros consistents per a (β, ρ), trobats a partir dels mateixos estadístics d'ordre que es fan servir per a l'estimació de l'índex de la cua.

Aquest nou estimador corregeix el biaix que presenta l'estimador de Hill. Si s'aplica a les dades daneses, s'obté una estimació de 0,72 per a ξ, molt propera al resultat que es pot extreure del *Hill-plot* visualment. Per a més detalls sobre l'estimació dels paràmetres de segon ordre i l'aplicació del mètode veure Gomes i Pestana (2007).

Un cop s'ha escollit el punt a partir del qual comença la cua de la distribució, cal estimar els paràmetres de forma, ξ, i d'escala, ψ. Els mètodes gràfics presentats a la secció anterior es podrien fer servir per a aquest propòsit; amb l'*ME-plot* es poden estimar els dos paràmetres mitjançant una regressió a partir del punt de tall, per exemple, però amb el *Hill-plot* només es podria estimar el paràmetre ξ.

El més habitual en la modelització de distribucions a partir d'un llindar és escollir aquest mitjançant el gràfic de l'*ME-plot* i ajustar llavors els paràmetres per ML, veure Coles (2001), apartat (4.3). En el següent apartat es descriurà amb detall l'estimació ML per a una GPD.

1.4 Estadística en EVT

A Diebold *et al.* (1998) es pot llegir la frase «*Una lectura de la literatura sobre EVT revela un contrast nítid i desafortunat entre la teoria de la probabilitat i la teoria estadística. La primera és elegant, rigorosa i voluminosa, mentre que la teoria estadística segueix sent primitiva i esquelètica en molts aspectes*». Tot i que aquesta aquesta reflexió té uns quants anys, les coses no han canviat gaire; la teoria probabilística presenta uns resultats aparentment senzills, però els problemes que van apareixent no tenen una resolució fàcil quan es tracta d'analitzar les dades, com poden ser l'existència de l'estimador màxim versemblant o la falta de moments de la GPD quan es volen utilitzar mètodes gràfics.

Gràcies als avenços informàtics, alguns d'aquests problemes que no tenen una solució teòrica s'han pogut resoldre numèricament. En aquest sentit, és important identificar els problemes existents, entendre'ls des del punt de vista teòric i buscar possibles solucions.

En aquest apartat es presentaran resultats recents, sorgits per intentar entendre i resoldre algunes de les problemàtiques típiques que poden aparèixer quan es planteja un problema de valors extrems, com pot ser l'existència de l'estimador màxim versemblant o les limitacions dels mètodes gràfics existents.

1.4.1 Versemblança del model GPD

Un dels mètodes habituals per estimar paràmetres és el de màxima versemblança; en molts casos l'optimització és complicada, no sempre es pot garantir l'existència del màxim i cal recórrer a l'ús de mètodes numèrics. El cas de la GPD presenta dificultats especials.

Sigui X una variable aleatòria GPD, la funció de densitat de X ve donada per

$$f(x;\xi,\psi) = \frac{1}{\psi}\left(1+\frac{\xi x}{\psi}\right)^{-(1+\xi)/\xi}, \tag{1.26}$$

i la corresponent funció de log-versemblança per l'equació

$$l(\xi,\psi) = \begin{cases} -n\log(\psi) - (1+1/\xi)\sum_{i=1}^{n}\log(1+\xi x_i/\psi), & \xi \neq 0 \\ -n\log(\psi) - \psi^{-1}\sum_{i=1}^{n} x_i, & \xi = 0 \end{cases} \tag{1.27}$$

on $x_1, x_2, ..., x_n$ és una mostra i.i.d. d'una GPD i $(1+\xi x_i/\psi) > 0$ per a $i = 1, ..., n$.

No és possible trobar una expressió analítica per al màxim de la log-versemblança d'una GPD, de manera que es requereixen mètodes numèrics, tenint cura d'evitar inestabilitats numèriques quan ξ es troba a prop de zero. Malgrat aquest inconvenient, recentment han aparegut resultats que permeten l'ús de la màxima versemblança per a la GPD, veure Castillo i Daoudi (2009) i Castillo i Serra (2015).

L'estudi de l'estimació per ML es pot trobar tractat a Davison (1984), Smith (1985) i Davison i Smith (1990). En particular, l'estimador ML existeix per a mostres grans donat $\xi > -1$, per a $\xi < -1$ no hi ha màxim local amb probabilitat tendint a 1, i és asimptòticament normal i eficient per a $\xi > -0{,}5$, però s'observa un comportament anòmal de la funció per a mostres de mida petita o moderada.

Els valors que es consideren habitualment són $-0{,}5 < \xi < 0{,}5$, per motius teòrics i pràctics. Les GPD amb $\xi < -1$ són aquelles que tenen suport compacte però amb funció de densitat creixent tendint a infinit a mida que s'apropen al punt final, poques situacions reals es troben on els valors de la cua tenen més probabilitat d'aparèixer que

els centrals. Per a $\xi > 0{,}5$ no hi ha existència de variància i el cas $-1 < \xi < -0{,}5$ no té bones propietats asimptòtiques de la versemblança, però cal remarcar que no s'han de descartar models propers a una cua uniforme ($\xi = -1$).

Per al cas de mostres petites, Castillo i Daoudi (2009) han donat arguments que expliquen aquest comportament inestable, caracteritzant la *profile-likelihood* en termes del coeficient de variació empíric de la mostra.

Reparametritzant la funció de log-versemblança en termes de ξ i σ expressant $l(\xi, \psi) = l(\xi, \xi\sigma)$, on $\sigma > 0$ per a $\xi \geq 0$ i $\sigma > max\{x_i\}$ per a $\xi < 0$, derivant aquesta expressió respecte ξ i igualant a 0, es pot escriure

$$\xi(\sigma) = \frac{1}{n} \sum_{i=1}^{n} log(1 + \ x_i/\sigma), \tag{1.28}$$

i la *profile-likelihood* es denota llavors com

$$l_p(\sigma) = -n \left(log[\xi(\sigma) \ \sigma] + \xi(\sigma) + 1 \right). \tag{1.29}$$

Per a una GPD el coeficient de variació, que només depèn del paràmetre de forma, es pot expressar com $\sqrt{1/(1-2\xi)}$, mentre que el coeficient de variació empíric es pot escriure com

$$cv = \sqrt{m_2 - m_1^2}/m_1, \tag{1.30}$$

on $m_k = \sum_{i=1}^{n} x_i^k/n$ són els moments mostrals.

Per a $\xi > 0$ el coeficient de variació és major que 1, per a la distribució exponencial ($\xi = 0$) val exactament 1 i és més petit que 1 en la resta de casos; ara bé, el coeficient de variació empíric no ha de satisfer aquestes condicions, tot i que és d'esperar que es trobi a prop del valor real.

El resultat principal presentat per Castillo i Daoudi (2009) diu que donada una mostra $\{x_i\}$ de nombres positius amb coeficient de variació empíric $cv > 1$, la funció de versemblança per a la distribució de Pareto té un màxim global en un punt finit. També és possible trobar un resultat similar per a les distribucions de suport compacte de la GPD.

Per una altra banda, Castillo i Serra (2015) presenten dues grans novetats, un nou algoritme per trobar l'estimador ML i resultats teòrics sobre les solucions d'aquest estimador. L'algoritme, basat en la *profile-likelihood*, és simple, ràpid i estable, a més a més, sempre retorna solució. Està programat en R i el seu codi és el següent:

```
#Estimador ML d'una mostra x d'una GPD(xi,psi)
eGPD<-function(x){
fxi<-function(sigma) -mean(log(1-x/sigma))
fp<-function(sigma){
```

```
 length(x)*(-log(fxi(sigma)*sigma)+fxi(sigma)-1)
}
int<-c(-100*max(x),100*max(x))
sigma<-optimize(fp,interval=int,maximum=T)$maximum
list(xi=-fxi(sigma),psi=fxi(sigma)*sigma)}
```

Sobre l'estimador ML, demostren que, per al cas $-1 \leq \xi \leq 0$, existeix un màxim global i, per al cas $\xi < -1$, la versemblança sempre és infinita. Un altre punt interessant que es tracta en aquest article és l'enfoc metodològic a l'hora d'escollir el model i estimar els paràmetres, el qual es comentarà més a fons en capítols posteriors.

1.4.2 Nou mètode gràfic

En la Secció 1.3 s'han presentat els dos mètodes gràfics més utilitzats en l'àmbit dels valors extrems, el *Hill-plot* i l'*ME-plot*; en aquest apartat es mostrarà un nou mètode gràfic presentat per Castillo *et al.* (2014a), el *CV-plot*. Aquest nou mètode gràfic és similar a l'*ME-plot*, però utilitza el coeficient de variació (CV) residual en comptes de l'esperança residual.

Sigui X una variable aleatòria contínua no negativa, la distribució residual de X sobre t es denota per $X_t = (X - t \mid X - t > 0)$ i el CV residual es defineix com

$$CV(t) = \sqrt{V(X_t)}/E(X_t) \tag{1.31}$$

on $V(X)$ i $E(X)$ corresponen a la variància i l'esperança, respectivament. Cal remarcar que $E(X_t) = ME(t)$ i que el CV residual és independent del paràmetre d'escala, de la mateixa manera que ho és el CV usual.

El CV residual d'una distribució GPD, donat $\xi < 1/2$, és constant donat per

$$CV\,(t) = cv_\xi = \sqrt{1/(1-2\xi)}, \tag{1.32}$$

i el seu valor és més gran que 1, més petit que 1 o igual a 1 d'acord amb $0 < \xi < 1/2$, $\xi < 0$ or $\xi = 0$.

Un CV residual constant caracteritza la distribució, és a dir, la distribució GPD és l'única que presenta aquesta propietat, veure Sullo i Rutherford (1977).

De la mateixa manera que Davison i Smith (1990) es van basar en la linealitat de l'esperança residual de la GPD per construir l'*ME-plot*, ara es pot definir de manera anàloga el *CV-plot*, basat en la propietat del CV residual de ser constant per a la mateixa família de distribucions.

El *CV-plot* és la representació de

$$\{(k,\ cv_n(x_{(k)})),\ 1 \leq k \leq n-1\} \tag{1.33}$$

on $\{x_{(i)}\}$ corresponen als estadístics d'ordre i $cv_n(t)$ és el coeficient de variació empíric de la distribució residual de X sobre t donat per

$$cv_n(t) = \frac{sd_t}{\bar{x}_t} \tag{1.34}$$

on sd_t i $\bar{x}_t$ són la desviació estàndar i la mitjana de X_t, respectivament.

Per a la representació del *CV-plot* s'ha escollit l'ordre dels llindars per a l'eix d'abcisses, però també es podrien fer servir els propis estadístics d'ordre. La corba representada és molt similar en els dos casos, simplement hi ha un canvi d'escala.

Castillo *et al.* (2014a) adjunten el codi R que permet realitzar aquesta gràfica amb els intervals de confiança incorporats per al cas exponencial. El seu interès es centra en aquest cas particular ($\xi = 0$), però els resultats trobats es poden generalitzar, tal i com es veurà en capítols posteriors. L'expressió que determina l'interval de confiança per a cada llindar t donat es desprèn del següent resultat.

Corol·lari 1.1. *Sigui X una variable aleatòria amb distribució exponencial, llavors $\sqrt{n(t)}(cv_n(t)-1)$ convergeix a un procés Gaussià amb mitjana 0 i funció de covariància donada per*

$$exp(-|s-t|/(2\mu)). \tag{1.35}$$

Aquesta és la funció de covariància d'un procés d'Ornstein-Uhlenbeck, la versió en temps continu d'un procés AR(1). És un procés estacionari de Markov Gaussià, en particular, per a cada t fixat

$$\sqrt{n(t)}(cv_n(t) - 1) \xrightarrow{d} N(0,\ 1). \tag{1.36}$$

A la Figura 1.5 es poden veure dos exemples de *CV-plot*. El primer, el de les dades daneses utilitzades anteriorment, mostra un exemple on el *CV-plot* no és útil perquè aquestes dades no tenen variància, de manera que no és possible obtenir cap informació, tot i que sembla que, a la part final de la gràfica, el *CV-plot* residual sí que acaba tendint cap a valors propers on hi ha existència de moments. Aquesta problemàtica es discutirà en els següents capítols i es presentarà una solució. El segon cas serveix per a il·lustrar la importància de considerar el CV residual enfront de només el CV, en aquest cas es presenta una mostra del valor absolut d'una distribució t de Student amb 4 graus de llibertat, que té CV igual a 1 però una cua pesada. Hi ha tests d'exponencialitat basats en el CV que no rebutjarien exponencialitat en mostres que provenen de distribucions com aquesta, ni en mostres que presenten cues lleugeres però amb CV igual a 1, com és el cas d'algunes mixtures de GPD; en capítols posteriors es presentarà una família de tests basats en el CV residual que serveixen per detectar aquests casos amb més precisió.

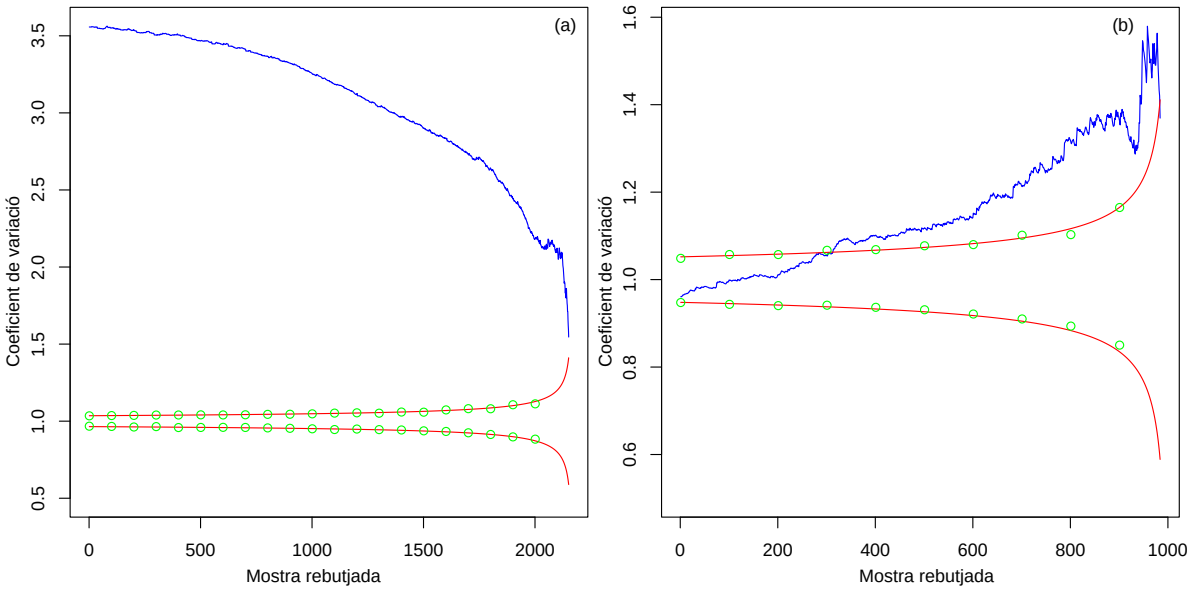

Figura 1.5: (a) *CV-plot* de les dades daneses sobre pèrdues en assegurances de focs. (b) *CV-plot* d'una mostra simulada de mida 1.000 del valor absolut d'una distribució t de Student amb 4 graus de llibertat. Les línies vermelles corresponen a l'interval de confiança teòric del 95% de la distribució exponencial i els punts verds al mateix interval però calculat mitjançant simulació.

1.4.3 Un nou enfoc per a cues pesades i exponencials

En el primer capítol s'ha comentat que un dels punts febles del mètode PoT és l'elecció del llindar a partir del qual la cua s'ha de distribuir com una GPD. Com ja s'ha dit, l'elecció, basada normalment en mètodes gràfics, sol ser subjectiva i sempre queda el problema de si s'ha agafat un llindar suficientment gran. Un llindar excessivament alt origina una mostra petita i dificulta l'estimació, però un llindar que no sigui prou elevat conté informació que no és de la cua i pertorba els resultats de l'estimació.

En algunes situacions, com per exemple en finances o en la variabilitat climatològica provocada per ciclons tropicals, les dades presenten un comportament semblant al d'una distribució de Pareto en un rang ampli però canvia per a valors molt grans cap a distribucions menys pesades. Si s'aplica el mètode PoT en aquests casos, queda clar que l'elecció del llindar serà crucial i els resultats poden distar bastant del que és raonable.

Castillo *et al* (2013) proposen un nou enfoc per a valors extrems, basat en l'ús de la distribució *full-tails gamma* (FTG), vàlid per a situacions com les que s'acaben de comentar.

Definició 1.8. Distribució FTG. La funció de densitat d'una distribució FTG ve donada per

$$f(x;\alpha,\theta,\rho) = \theta\ \ (\rho+\theta x)^{\alpha-1}\exp(-(\rho+\theta x))/\Gamma(\alpha,\rho) \tag{1.37}$$

amb suport a $(0, \infty)$, on $\alpha \in \mathbb{R}$, $\theta > 0$, $\rho > 0$ i $\Gamma(\alpha, \rho)$ és la funció gamma incompleta.

Aquesta família conté la distribució gamma, en particular la distribució exponencial, la distribució gamma truncada ($\alpha > 0$), la seva extensió per a $\alpha \leq 0$ i la distribució de Pareto. També presenta la propietat de ser tancada per truncament, com li passa a la GPD, i permet trobar distribucions tan a prop de la distribució de Pareto segons el que determinen les dades, però amb una major flexibilitat, ampliant les distribucions de cues pesades proporcionades per la GPD.

L'estimació de paràmetres per ML es pot realitzar mitjançant mètodes numèrics, veure Castillo *et al* (2013), i cal remarcar que, com passa amb el cas de l'estimador de Hill, aquest mètode només serà adequat en casos on clarament hi hagi cues pesades, perquè aplicat en cas de cues lleugeres pot donar resultats no coherents.

CAPÍTOL 2

Caracterització de distribucions mitjançant el CV residual

Donada una variable aleatòria X, es diu que aquesta queda totalment caracteritzada si es pot determinar la probabilitat de que X prengui valors en qualsevol interval de la recta real. La manera més habitual de caracteritzar variables aleatòries és mitjançant la funció de distribució o la funció de densitat, però existeixen altres maneres de fer-ho, per exemple, fent servir la funció generatriu de moments.

L'esperança residual caracteritza la funció de distribució mitjançant la relació

$$1 - F(t) = \frac{ME(0)}{ME(t)}\, exp\left[-\int_0^t \frac{ds}{ME(s)}\right]. \tag{2.1}$$

Gupta i Kirmani (2000) van demostrar que el CV residual caracteritza l'esperança residual mitjançant la següent relació

$$ME(t)\left(1 + CV^2(t)\right) = \int_0^t CV^2(s)ds + ME(1 + CV^2(0)) - t, \tag{2.2}$$

per tant, també caracteritzarà la distribució fent servir (2.1).

D'aquest resultat es desprèn que si el CV residual és constant, llavors la distribució X ha de ser GPD. Per tant aquestes distribucions queden unívocament determiandes pel seu CV residual.

En aquest capítol es presentarà la caracterització mitjançant el CV residual per a algunes variables aleatòries, es dibuixaran els *CV-plots* teòrics corresponents i es desenvoluparà la teoria asimptòtica que permetrà obtenir l'expressió dels intervals de confiança associats als *CV-plots* de la GPD.

2.1 Càlcul general del CV residual

El *CV-plot* teòric per a una GPD té la forma més simple que es pot trobar, una funció constant, i uns intervals de confiança que es poden calcular com es veurà a la Secció 2.3. El pas següent que es pot realitzar és buscar els *CV-plots* teòrics per a d'altres distribucions; ja que el CV residual les caracteritza i cada una tindrà, per tant, la seva gràfica corresponent. Això serveix a la pràctica per identificar els comportaments empírics que s'observen amb diferents conjunts de dades.

Per a aquest propòsit, tot i que el *CV-plot* s'ha definit anteriorment per a cada posició del llindar a la mostra o per a cada llindar, s'ha escollit ara la representació per quantils, és a dir, a cada valor $0 < p < 1$ se li assigna el CV residual per al corresponent quantil q_p; ja que d'aquesta manera es poden representar *CV-plots* de diferents distribucions en un mateix gràfic, independentment del suport que tingui cadascuna, i perquè aporta una visió directa del precentatge de dades que es va eliminant a cada pas.

En aquest apartat s'explicarà el càlcul general que es pot aplicar a qualsevol distribució, es presentaran els detalls particulars per a cada una de les distribucions escollides com a exemples i es representaran els *CV-plots* corresponents per a diferents valors dels paràmetres de forma.

Per poder calcular el CV residual, es necessita saber quina és al distribució dels excessos sobre un llindar, per tant s'ha de començar trobant la seva expressió.

Donada una variable aleatòria positiva contínua X amb funció de densitat $f(x)$, la distribució dels excessos de X per sobre d'un llindar t, donada per (1.13), es pot escriure com

$$F_t(x) = \int_t^{x+t} f(s)ds/(1 - F(t)). \tag{2.3}$$

D'aquesta manera, l'esperança i la variància condicionades es calculen com

$$ME(t) = \int_t^{x_F} (x - t)f(x)dx/(1 - F(t)) \tag{2.4}$$

$$\begin{aligned} V(t) = \int_t^{x_F} (x - t)^2 f(x)dx/(1 - F(t)) - ME(t)^2 = \\ = E((X - t)^2 |\, X > t) - ME(t)^2. \end{aligned} \tag{2.5}$$

Donat que $CV(t) = V(t)^{1/2}/ME(t)$, només cal calcular $ME(t)$ i $E((X-t)^2 |\, X > t)$, per aplicar, finalment, les fórmules (2.5) i (1.31).

No sempre es pot trobar una fórmula tancada per a aquestes expressions, però es poden programar fàcilment fent servir la funció `integrate` del paquet `stats` de l'R. Com a exemple d'un cas amb aquestes característiques, es representarà el *CV-plot* per

a un cas particular de la distribució Normal Inversa Gaussiana en valor absolut, fent servir les funcions `dnig` i `pnig` del paquet `fBasics` per obtenir les seves funcions de densitat i distribució, respectivament.

Cal remarcar, una altra vegada, que el *CV-plot* no depèn del paràmetre d'escala, tot i que, en algunes de les expressions que es presentaran a continuació, apareixerà degut a que aquestes no es simplificaran del tot per a la comoditat de la lectura.

2.2 Exemples

Exemple 2.1. Distribució gamma

Sigui $X \sim \Gamma(\alpha, \beta)$ una distribució gamma.

La funció de distribució de X és

$$F_X(x;\alpha,\beta) = \int_0^x \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s} ds, \tag{2.6}$$

on $\Gamma(x)$ és la funció gamma, $x > 0$, $\alpha > 0$ i $\beta > 0$.

L'esperança condicionada té la següent expressió:

$$\begin{aligned} ME(t) &= \int_t^\infty (x-t) \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx / (1 - F_X(t;\alpha,\beta)) = \\ &= (\alpha(1 - F_X(t;\alpha+1,\beta))/\beta - t(1 - F_X(t;\alpha,\beta))) / (1 - F_X(t;\alpha,\beta)). \end{aligned} \tag{2.7}$$

Es calcularà ara $E((X-t)^2 | \, X > t)$.

$$\begin{aligned} E((X-t)^2 | \, X > t) &= \int_t^\infty (x-t)^2 \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx / (1 - F_X(t;\alpha,\beta)) = \\ &= \left(\alpha(\alpha+1)(1-a)/\beta^2 - 2t\alpha(1-b)/\beta + t^2(1-c)\right) / (1-c). \end{aligned} \tag{2.8}$$

on $a = F_X(t;\alpha+2,\beta)$, $b = F_X(t;\alpha+1,\beta)$ i $c = F_X(t;\alpha,\beta)$.

Un cop realitzats els càlculs necessaris, com que interessarà agafar com a llindar el quantil k per a una distribució gamma amb paràmetres α i β, és a dir, $q_{k,\alpha,\beta}$, només cal canviar el valor del llindar t per aquest valor.

Anomenant $F_X(q_{k,\alpha,\beta};\alpha,\beta) = k$, $F_X(q_{k,\alpha,\beta};\alpha+1,\beta) = m$ i $F_X(q_{k,\alpha,\beta};\alpha+2,\beta) = n$, el CV residual queda de la següent manera

$$CV(q_{k,\alpha,\beta}) = \frac{(\alpha(\alpha+1)(1-n)(1-k) - \alpha^2(1-m)^2)^{1/2}}{\alpha(1-m) - q_{k,\alpha,\beta}\beta(1-k)}. \tag{2.9}$$

A la Figura 2.1 (b) es pot observar quin és el comportament del *CV-plot* de la distribució gamma en funció del valor del paràmetre α. Per a $\alpha < 1$ la gràfica és

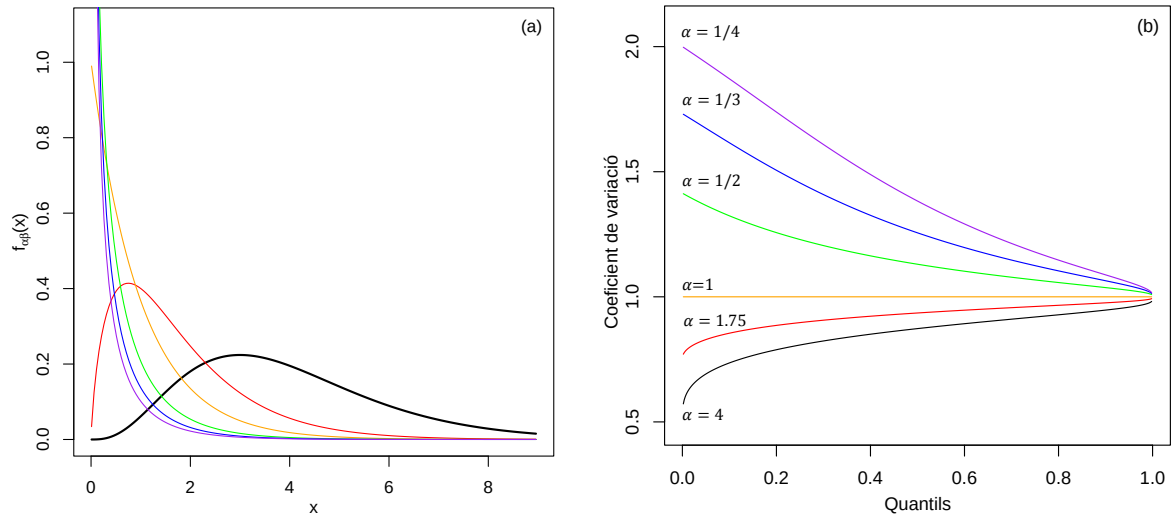

Figura 2.1: (a) Densitats de distribucions gamma per a diferents valors del paràmetre α, en tots els casos $\beta = 1$. (b) *CV-plots* teòrics per a les mateixes distribucions, identificades amb els mateixos colors.

decreixent tendint cap a 1, si $\alpha < 1$ és creixent tendint també cap a 1 i el cas $\alpha = 1$, que correspon a la distribució exponencial, té coeficient de variació residual constant igual a 1, com ja s'havia vist, ja que aquesta distribució és un cas particular de GPD.

A la Figura 2.1 (a) es poden veure les densitats amb les mateixos paràmetres escollits per a la gràfica (b), identificades amb el mateix color. Es pot observar com les densitats que es troben per sota de l'exponencial (color taronja) al principi del suport són les que després presenten un CV residual major que 1 (colors lila, blau i verd), mentre que les altres corresponen a cues lleugeres (colors vermell i negre). Aquesta distribució serveix com a model per comparar densitats i *CV-plots*, ja que en els casos posteriors ja no es mostraran les densitats.

Exemple 2.2. ***Distribució de Pareto Generalitzada***

Sigui X una $GPD(\xi, \psi)$ amb funció de distribució donada per (1.14).

L'esperança condicionada té la següent expressió:

$$ME(t) = \int_t^\infty (x-t)1/\psi(1+\xi x/\psi)^{-1-1/\xi}dx/(1-F_X(t;\alpha,\beta)) = \\ = (\psi+\xi t)/(1-\xi), \tag{2.10}$$

però només per a $\xi < 1$.

Es calcularà ara $E((X-t)^2|\, X > t)$.

$$E((X-t)^2|\, X > t) = \int_t^\infty (x-t)^2 1/\psi(1+\xi x/\psi)^{-1-1/\xi}dx/(1-F_X(t;\xi,\psi)) = \\ = (\psi+\xi t)^2/\left((1-\xi)^2(1-2\xi)\right), \tag{2.11}$$

però només per a $\xi < 1/2$.

D'on es dedueix fàcilment que el CV residual és igual a $\sqrt{1/(1-2\xi)}$, tal i com s'havia definit a (1.32).

A la Figura 2.2 es pot observar el comportament del *CV-plot* de ls GPD en funció de l'índex del valor extrem ξ. El CV residual serà constant major, igual o menor que 1 si $\xi > 0$, $\xi = 0$ i $\xi < 0$, respectivament.

Exemple 2.3. Distribució log-normal

Sigui $X \sim lnN(\mu, \sigma^2)$ una distribució log-normal amb paràmetres μ i σ.

La funció de distribució de X és

$$F_X(x;\mu,\sigma) = \Phi(\frac{ln(x)-\mu}{\sigma}), \tag{2.12}$$

on $\Phi(x)$ és la funció de distribució d'una variable normal estàndar, i la seva funció de densitat és

$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{ln(x)-\mu}{2\sigma^2}}. \tag{2.13}$$

L'esperança condicionada, després d'aplicar el canvi de variable $t = ln(x) - \mu$, té la següent expressió:

$$\begin{aligned} E(X-t \mid X>t) &= \int_t^\infty (x-t)\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{ln(x)-\mu}{2\sigma^2}} dx/(1-F_X(t;\mu,\sigma)) = \\ &= \left(\exp((\sigma^2+2\mu)/2)(1-\Phi(a-1)) - bt\right)/b \end{aligned} \tag{2.14}$$

on $a = (ln(t)-\mu)/\sigma$ i $b = 1 - F_X(t;\mu,\sigma)$.

Es calcularà ara $E((X-t)^2 | X>t)$, aplicant el canvi de variable $z = ln(x) - \mu$.

$$\begin{aligned} E((X-t)^2 | X>t) &= \int_t^\infty (x-t)^2 \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{ln(x)-\mu}{2\sigma^2}} dx/(1-F_X(t;\mu,\sigma)) = \\ &= \left(\exp(2\sigma^2+2\mu)(1-\Phi(a-2\sigma)) - 2\exp((\sigma^2+2\mu)/2)(1-\Phi\left(a-1\right))t + bt^2\right)/b. \end{aligned} \tag{2.15}$$

Un cop realitzats els càlculs necessaris, com que interessarà agafar com a llindar el quantil k per a una log-normal amb paràmetres μ i σ, és a dir, $q_{k,\mu,\sigma}$, només cal canviar el valor del llindar t per aquest valor.

Anomenant $F_X(q_{k,\mu,\sigma};\mu,\sigma) = \Phi\left(\frac{ln(q_{k,\mu,\sigma})-\mu}{\sigma}\right) = k$, $\Phi\left(\frac{ln(q_{k,\mu,\sigma})-\mu-\sigma}{\sigma}\right) = m$ i $\Phi\left(\frac{ln(q_{k,\mu,\sigma})-\mu-2\sigma^2}{\sigma}\right) = n$, el CV residual queda de la següent manera

$$CV(q_{k,\mu,\sigma}) = \frac{\left(e^{2\sigma^2+2\mu}(1-n)(1-k) - e^{\sigma^2+2\mu}(1-m)^2\right)^{1/2}}{e^{\frac{\sigma^2+2\mu}{2}}(1-m) - q_{k,\mu,\sigma}(1-k)}. \tag{2.16}$$

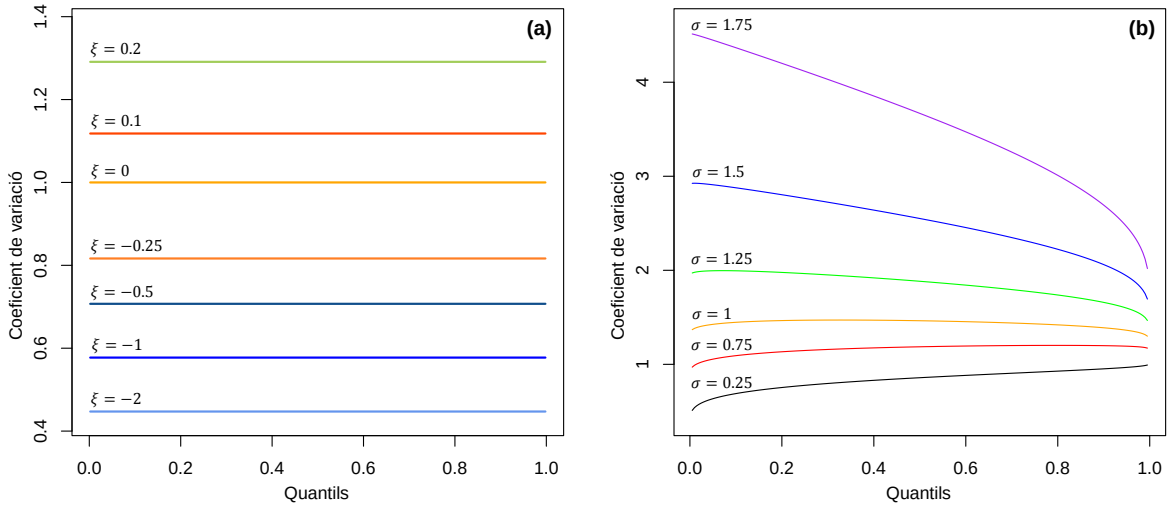

Figura 2.2: (a) *CV-plots* teòrics de la GPD per a diferents valors del paràmetre ξ. (b) *CV-plots* teòrics de distribucions log-normals per a diferents valors del paràmetre σ.

A la Figura 2.2 es pot observar quin és el comportament del *CV-plot* de la distribució log-normal en funció del valor del paràmetre σ. La gràfica és creixent per a alguns valors del paràmetre i decreixent per a d'altres, però en tots els casos tendeixen cap a 1 a mida que augmenten els quantils, amb una convergència cada cop més lenta com més gran és el valor de σ.

Exemple 2.4. Distribució *half normal*

Sigui $X \sim N(0, \sigma^2)$, el valor absolut d'aquesta distribució és una altra distribució, definida a l'interval $(0, \infty)$, que s'anomena *half normal* i es denota per $Y \sim HN(\sigma^2)$.

La funció de distribució de Y és

$$F_Y(x;\sigma) = \int_0^x \frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{-\frac{s^2}{2\sigma^2}} ds. \tag{2.17}$$

Sigui $F_X(x;0,\sigma)$ la funció de distribució de X, on $F_X(x;\mu,\sigma)$ és la distribució d'una variable normal amb paràmetres μ i σ, llavors es té que $F_Y(x;\sigma) = 2F_X(x;0,\sigma) - 1$, de manera que a partir dels valors tabulats de la distribució normal estàndard es pot trobar la distribució de la variable X i, a partir d'aquests, els de la variable Y.

L'esperança condicionada, aplicant el canvi de variable $z = \frac{x^2}{2\sigma^2}$, té la següent expressió:

$$\begin{aligned} E(X - t \mid X > t) &= \Big(\int_t^\infty (x-t)\frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{-\frac{x^2}{2\sigma^2}} dx\Big)/(1 - F_Y(t;\sigma)) = \\ &= \Big(\sigma\sqrt{2\pi^{-1}}\exp(-t^2/2\sigma^2) - t(1 - F_Y(t;\sigma))\Big)/1 - F_Y(t;\sigma). \end{aligned} \tag{2.18}$$

Es calcularà ara $E((X-t)^2 | X > t)$, aplicant elsl canvis de variables $z = \frac{x^2}{\sigma\sqrt{2}}$ i $s = \frac{x^2}{2\sigma^2}$ quan sigui convenient.

$$E((X-t)^2|\, X>t) = (\int_t^\infty (x-t)^2 \frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{-\frac{x^2}{2\sigma^2}} dx)/(1-F_Y(t;\sigma)) = \\ = \left((\sigma^2+t^2)(1-F_Y(t;\sigma)) - \sqrt{2\pi^{-1}}\sigma \exp(-t^2/2\sigma^2)t\right)/(1-F_Y(t;\sigma)). \tag{2.19}$$

Un cop realitzats els càlculs necessaris, com que interessarà agafar com a llindar el quantil k per a una *half normal* amb paràmetre σ, és a dir, $q_{k,\sigma}$, només cal canviar el valor del llindar t per aquest valor.

Es pot comprovar fàcilment que, per als quantils d'una *half normal*, es té la relació

$$q_{k,\sigma} = \sigma q_{k,1}, \tag{2.20}$$

i, considerant que $F_Y(x;\sigma) = 2F_X(x;0,\sigma) - 1$, es té $F_X(q_{k,1};0,1) = \frac{k+1}{2}$.

D'on es pot deduir que el quantil k d'una *half normal* amb parametre σ, $q_{k,\sigma}$, correspon al quantil $\frac{k+1}{2}$ d'una distribució normal estàndar multiplicat per σ.

Si es considera q_k^* el quantil $(k+1)/2$ d'una distribució normal estàndar, el CV residual queda de la següent manera

$$CV(q_{k,\sigma}) = \frac{\left((1-k)^2\pi + \sqrt{2\pi}q_k^*(1-k)e^{-\frac{q_k^{*2}}{2}} - 2e^{-\frac{q_k^{*2}}{2}}\right)^{1/2}}{\sqrt{2}e^{-\frac{q_k^{*2}}{2}} - q_k^*\sqrt{\pi}(1-k)}. \tag{2.21}$$

A la Figura 2.3 es pot veure el *CV-plot* de la distribució *half normal* com el límit del *CV-plot* del valor absolut d'una distribució t_ν (t de Student amb ν graus de llibertat) quan $\nu \to \infty$. El valor del coeficient de variació residual tendeix al valor 1 a mesura que augmenten els quantils, que és el valor del coeficient de variació residual de la distribució exponencial; fet que és d'esperar, ja que les cues d'una distribució *half normal* són exponencials, tot i que la convergència és lenta i aquest valor no s'assoleix fins al final de la gràfica.

Exemple 2.5. Distribució t de Student en valor absolut

Sigui $X \sim t_\nu$, una distribució t de Student amb ν graus de llibertat, considerem $Y = |X|$, una nova distribució definida a l'interval $(0,\infty)$.

La funció de distribució de Y és

$$F_Y(x;\nu) = \int_0^x \frac{2\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{s^2}{\nu})^{-\frac{\nu+1}{2}} ds = \int_0^x k(1+\frac{s^2}{\nu})^{-\frac{\nu+1}{2}} ds. \tag{2.22}$$

Sigui $F_X(x;\nu)$ la funció de distribució de X, llavors es té que $F_Y(x;\nu) = 2F_X(x;\nu) - 1$, de manera que a partir dels valors tabulats de la distribució de la variable X es pot trobar els de la distribució de la variable Y.

L'esperança condicionada, aplicant el canvi de variable $z = 1 + \frac{x^2}{\nu}$, té la següent expressió:

$$E(X-t \mid X>t) = \int_t^\infty k(x-t)(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}dx/(1-F_Y(t;\nu)) = \\ = \left(2\nu k/(\nu-1)\left(1+t^2/\nu\right)^{-(\nu-1)/2} - (1-F_Y(t;\nu))t\right)/(1-F_Y(t;\nu)), \tag{2.23}$$

però només per a $\nu > 1$, ja que es necessita $\frac{\nu-1}{2} > 0$.

Es calcularà ara $E((X-t)^2|\, X>t)$, aplicant el canvi de variable $z = 1+\frac{x^2}{\nu}$.

$$E((X-t)^2|\, X>t) = \int_t^\infty k(x-t)^2(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}dx/(1-F_Y(t;\nu)) = \\ = k\sqrt{\nu^3\pi}\Gamma((\nu-2)/2)/(2(\nu-2)\Gamma((\nu-1)/2))\left(1-F_Y(\frac{\sqrt{\nu-2}}{\sqrt{\nu}}t;\nu-2)\right)/(1-F_Y(t;\nu)) + \\ + \left(t^2(1-F_Y(t;\nu)) - k\nu/(\nu-1)\left(1+t^2/\nu\right)^{-\frac{\nu-1}{2}}\right)/(1-F_Y(t;\nu)), \tag{2.24}$$

però només per a $\nu > 2$, ja que es necessita $\nu - 2 > 0$.

A partir d'aquest punt es procediria com en els casos anterior, calculant l'expressió del CV residual per a un quantil k donat, obtenint un *CV-plot* diferent per a cada grau de llibertat donat,però donada la impossibilitat de reduir aquesta expressió de manera que sigui intel·ligible, s'ometrà la seva aparició.

A la Figura 2.3, es pot observar com, a mesura que augmenten els graus de llibertat, els *CV-plots* de les distribucions t de Student en valor absolut tendeixen al *CV-plot* de la distribució *half normal*, fet que és d'esperar, ja que el límit d'una distribució t quan els graus de llibertat tendeixen a infinit és una distribució normal estàndar.

Exemple 2.6. Distribució Normal Inversa Gaussiana en valor absolut

La distribució normal inversa gaussiana (NIG), proposada per Barndorff-Nielsen (1997), és una distribució contínua amb funció de densitat

$$f(x;\alpha,\beta,\mu,\delta) = a(\alpha,\beta,\mu,\delta)\, q\left(\frac{x-\mu}{\delta}\right)^{-1} K_1\left(\delta\alpha q\left(\frac{x-\mu}{\delta}\right)\right)\exp(\beta x) \tag{2.25}$$

on $a(\alpha,\beta,\mu,\delta) = \pi^{-1}\alpha\exp\left(\delta\sqrt{(\alpha^2-\beta^2)}-\beta\mu\right)$, $q(x) = \sqrt{1+x^2}$ i K_λ és la funció de Bessel modificada de tercer tipus amb índex λ donada per

$$K_\lambda(x) = \frac{1}{2}\int_0^\infty y^{\lambda-1}\exp(-x(y+y^{-1})/2)dy. \tag{2.26}$$

Els paràmetres de forma, α, escala, δ, asimetria, β, i localització, μ, satisfan $0 \le |\beta| \le \alpha$, $\mu \in \mathbb{R}$ i $\delta > 0$.

El fet de que la funció de densitat contingui la funció de Bessel dificulta poder trobar una expressió tancada per al CV residual, però, tal i com s'ha comentat abans, existeixen funcions de R que permeten programar-la numèricament.

Per il·lustrar el *CV-plot* en aquesta situació, es farà servir un cas concret de la distribució NIG. Considerant $\beta = \mu = 0$ i la reparametrització $\phi = \delta/\alpha$ i $\omega = \alpha\delta$, s'obté la distribució NIG simètrica amb mitjana 0, denotada per $X \sim NIG(\phi, \omega)$.

Sigui ara $Y = |X|$ una nova variable aleatòria, $\phi > 0$ és el seu paràmetre de forma i $\omega > 0$ el d'escala. Mitjançant les funcions `dnig` i `pnig` del paquet `fBasics` i la funció `integrate` del paquet `stats` de l'R, és possible crear un algoritme que repliqui els càlculs generals que s'han presentat al principi del capítol.

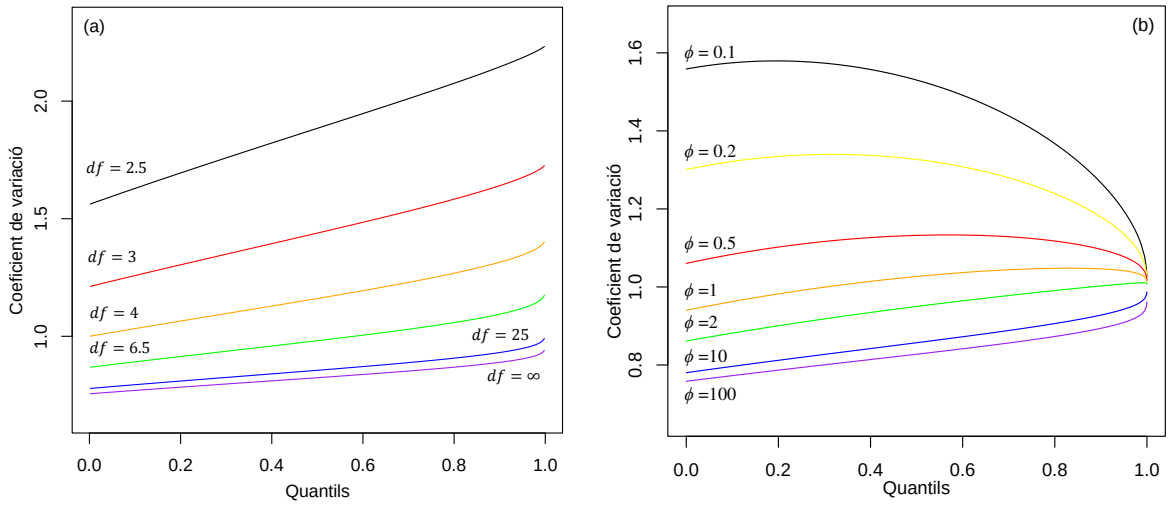

Figura 2.3: (a) *CV-plots* teòrics de distribucions t de Student en valor absolut per a diferents graus de llibertat (df) i per a la distribució *half normal* (df $=\infty$). (b) *CV-plots* teòrics de distribucions NIG simètriques amb mitjana 0 en valor absolut per a diferents valors del paràmetre ϕ.

A la Figura 2.3 es pot veure el coportament del *CV-plot* per a diferents valors del paràmetre ϕ. Per a valors petits, la gràfica acaba decreixent tendint cap al valor 1 i, a mida que creix el valor del paràmetre, el *CV-plot* és creixent però tendint igualment cap a 1.

2.3 Teoria asimptòtica

En aquesta secció, s'estudiarà la distribució asimptòtica del coeficient de variació residual per a la GPD com un procés aleatori indexat pel llindar. Aquesta distribució proporciona intervals de confiança puntuals per al *CV-plot* i un test de múltiples llindars que redueix el problema de considerar múltiples llindars per separat. Aquests resultats generalitzen els presentats a Castillo *et al.* (2014a).

Denotem $\mu_0(t) = P\{X > t\}$ i $\mu_k(t) = E[X^k 1_{(X>t)}]$, $k > 0$. S'assumirà $\mu_0(t) > 0$, per a tot t, a partir d'ara. Notem que

$$\mu_k(t) = \mu_0(t) E(X^k \mid X > t). \tag{2.27}$$

Donada una mostra $\{x_i\}$ de mida n, sigui $n(t) = \sum_{j=1}^{n} 1_{(X_j>t)}$ el nombre d'observacions que es troben sobre un cert llindar t. Per la llei dels grans nombres, $n(t)/n$ convergeix cap a $\mu_0(t)$.

El $cv_n(t)$, donat per la fórmula (1.34), és un estimador consistent de cv_ξ, assumint moment de segon ordre finit, ja que el límit en probabilitat de $cv_n(t)$ quan n tendeix a infinit és

$$m_{cv}(t) = \frac{\sqrt{\mu_2(t)\mu_0(t) - \mu_1(t)^2}}{\mu_1(t) - t\mu_0(t)} = cv_\xi. \tag{2.28}$$

Es defineix el $k-$èssim moment mostral estandaritzat de l'excés condicional com

$$W_{k,n}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left\{ X_j^k 1_{(X_j>t)} - \mu_k(t) \right\}, \tag{2.29}$$

per tant,

$$\sum_{j=1}^{n} X_j^k 1_{(X_j>t)} = \sqrt{n} W_{k,n}(t) + n\mu_k(t). \tag{2.30}$$

Cal observar que la constant normalitzadora $1/\sqrt{n}$ es fa servir per tenir $W_{k,n}(t) = O_p(1)$, amb ordres de convergència en notació de probabilitats. La covariància d'aquest procés aleatori ve donada per

$$cov(W_{i,n}(s), W_{j,n}(t)) = cov(X^i 1_{(X>s)}, X^j 1_{(X>t)}) = \mu_{i+j}(s \vee t) - \mu_i(s)\mu_j(t). \tag{2.31}$$

Encara que les quantitats cv i W_k depenen entre d'altres de n, la possible dependència d'aquest valor serà suprimida per simplicitat a partir d'ara. Fins i tot la dependència de t serà suprimida per a $W_k = W_k(t)$ i $\mu_k = \mu_k(t)$ en molts llocs.

Teorema 2.1. *Sigui X una variable aleatòria contínua no negativa amb moments finits de quart ordre. Llavors, es té la següent expansió*

$$\sqrt{n}(cv(t) - m_{cv}(t)) = \frac{\mu_0 W_2}{2(t\mu - t\mu_0)\sqrt{\mu_2\mu_0 - \mu_1^2}} + \frac{\mu_0(t\mu_1 - \mu_2)W_1}{(t\mu - t\mu_0)^2\sqrt{\mu_2\mu_0 - \mu_1^2}} +$$
$$+ \frac{(-2t\mu_1^2 + t\mu_0\mu_2 + \mu_1\mu_2)W_0}{2(t\mu - t\mu_0)^2\sqrt{\mu_2\mu_0 - \mu_1^2}} + O_p\left(\frac{1}{\sqrt{n}}\right). \tag{2.32}$$

Per als detalls de la demostració, veure Castillo *et al.* (2014a).

Teorema 2.2. *Sigui X una variable aleatòria GPD amb $\xi < 1/4$ i $\psi > 0$ paràmetres de forma i escala, respectivament; llavors $\sqrt{n}(cv(t) - cv_\xi)$ convergeix a un procés Gaussià amb mitjana 0 i funció de covariància donada per*

$$\rho_0(s,t) = exp((s \wedge t)/\psi) \tag{2.33}$$

per a $\xi = 0$ i

$$\rho_\xi(s,t) = \left(\frac{\psi + \xi(s \wedge t)}{\psi}\right)^{\frac{1}{\xi}} \cdot \frac{(1-\xi)^2}{(1-3\xi)(1-2\xi)^2(1-4\xi)(\psi + \xi(s \wedge t))^2} \cdot$$
$$\big((6\xi^4 - 9\xi^3 + 3\xi^2)(s \vee t)^2 + (12\psi\xi^3 - 10\psi\xi^2 + 4\psi\xi)(s \vee t)$$
$$+(8\xi^3 - 2\xi^2)(s \wedge t)(s \vee t) + (8\psi\xi^2 - 2\psi\xi)(s \wedge t) + 6\psi^2\xi^2 - \psi^2\xi + \psi^2\big) \tag{2.34}$$

per a $\xi \neq 0$ i $s \leq t$.

Demostració. A partir del Teorema 2.1 es té

$$\sqrt{n}(cv(t) - cv_\xi) = (W_0, W_1, W_2)a(t) + O_p(n^{-1/2})$$

on

$$a_1(t) = (1-\xi)(\xi t^2 + 2\psi^2 + 4\psi t + t^2)/2$$
$$a_2(t) = (-2\psi - t)(1-\xi)^2$$
$$a_3(t) = (1-\xi)^2(1-2\xi)/2$$

$$a'(t) = \left(\frac{\psi + \xi t}{\psi}\right)^{1/\xi} \cdot (a_1(t),\ a_2(t),\ a_3(t)) \,/((1-2\xi)^{1/2}\,(\psi - \xi t)^2).$$

Llavors, la matriu de covariàncies de $W = (W_0, W_1, W_2)'$, per (2.31), assumint $s \leq t$, és

$$cov(W_{i,n}(s), W_{j,n}(t)) = M(s,t) = (\mu_{i+j}(s \vee t) - \mu_i(s)\mu_j(t))_{i,j=0,1,2}.$$

Una mica d'àlgebra demostra

$$a'(s)M(s,t)a(t) = \rho_\xi(s,t).$$

□

Corol·lari 2.1. *En particular, sota les condicions del Teorema 2.2,*

$$\sqrt{n}(cv(0) - cv_\xi) \xrightarrow{d} N\left(0, \sigma_\xi^2\right) \tag{2.35}$$

on $\sigma_\xi^2 = (6\xi^2 - \xi + 1)(1-\xi)^2 / \left((1-2\xi)^2(1-4\xi)(1-3\xi)\right)$.

Demostració. Per a la prova del cas particular, només cal observar que

$$\rho_\xi(0,0) = \sigma_\xi^2.$$

□

Corol·lari 2.2. *Sigui X una variable aleatòria GPD amb $\xi < 1/4$ i $\psi > 0$* paràmetres de forma i d'escala*, respectivament, llavors $\sqrt{n(t)}(cv(t) - cv_\xi)$ convergeix a un procés Gaussià amb mitjana 0 i funció de covariància donada per*

$$\rho_\xi^*(s,t) = \left(\frac{\psi + \xi s}{\psi}\right)^{1/2\xi} \rho_\xi(s,t) \left(\frac{\psi + \xi t}{\psi}\right)^{1/2\xi} \tag{2.36}$$

En particular, per a cada t fixat

$$\sqrt{n(t)}(cv(t) - cv_\xi) \xrightarrow{d} N\left(0,\, \sigma_\xi^2\right). \tag{2.37}$$

Demostració. Recordem que $n(t)/n$ convergeix a $\mu_0(t) = Pr\{X > t\} > 0$. Per tant, si n tendeix cap a infinit $n(t)$ tendeix cap a infinit també. Es pot escriure

$$\sqrt{n(t)}\,(cv(t) - cv_\xi) = \sqrt{n(t)/n}\sqrt{n}\,(cv(t) - cv_\xi).$$

A partir de (2.30), s'obté
$\sum_{j=1}^{n} X_j^0 1_{(X_j > t)} = \sqrt{n} W_{0,n}(t) + n\mu_0(t) \Rightarrow n(t) = \sqrt{n} W_{0,n}(t) + n\left(\frac{\psi - \xi t}{\psi}\right)^{1/\xi}$

$$\frac{n(t)}{n} = \left(\frac{\psi - \xi t}{\psi}\right)^{1/\xi} + \frac{W_0}{\sqrt{n}}.$$

Llavors $\sqrt{n(t)} \approx \sqrt{n}\left(\frac{\psi - \xi t}{\psi}\right)^{1/2\xi}$ i es té, suposant $s \le t$,

$$\rho_\xi^*(s,t) = \left(\frac{\psi + \xi s}{\psi}\right)^{1/2\xi} \rho_\xi(s,t) \left(\frac{\psi + \xi t}{\psi}\right)^{1/2\xi}$$

□

Exemple 2.7. Dades sobre les onades a Bilbao.

Per il·lustrar com es pot realitzar un *CV-plot* amb un interval de confiança per a un paràmetre ξ donat, es faran servir dades sobre les onades a Bilbao, presentades a Castillo i Hadi (1997), les quals representen la mitjana per hora dels períodes de *zero-crossing* (temps que triga una ona en tornar a la seva posició incial) expressada en segons i corresponen a 179 observacions agafades a partir del llindar 7,5 segons.

Aquestes dades han estat molt estudiades i sembla adient ajustar-les per una GPD amb índex del valor extrem $\xi = -1$, és a dir, una distribució uniforme, veure Castillo i Serra (2015).

Segons l'equació (2.37), per a aquest cas es té

$$\sqrt{n(t)}(cv(t) - 1/\sqrt{3}) \xrightarrow{d} N\left(0, 8/45\right). \tag{2.38}$$

Per tant, l'interval de confiança quedarà centrat a $1/\sqrt{3} \approx 0{,}577$ i, a mida que s'augmenta el llindar, es va eixamplant perquè la variància cada vegada és més gran.

A la Figura 2.4 es poden veure dos *CV-plots* per a aquestes dades, un amb els intervals de confiança per les distribucions exponencial i uniforme i un altre només amb l'interval de confiança per al cas exponencial. Sembla que la suposició d'uniformitat de les dades és encertada.

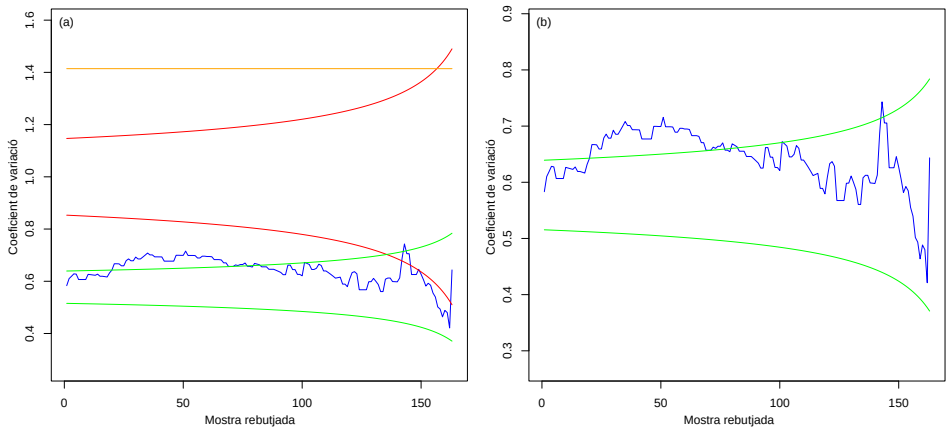

Figura 2.4: (a) *CV-plot* de les dades de Bilbao amb els corresponents intervals de confiança del 95% per a les distribucions uniforme (verd) i exponencial (vermell). La línia taronja correspon al CV residual per a $\xi = 1/4$. (b) *CV-plot* per a les mateixes dades però només amb l'interval per a la distribució uniforme.

2.4 Comentaris i propostes sobre el CV-plot per a la GPD

El procediment que s'ha presentat al prinicpi d'aquest capítol per calcular el CV residual es pot aplicar sempre i quan existeixin els moments de primer i segon ordre de la variable aleatòria X, és a dir, l'esperança i la variància.

De fet, també funciona per a variables aleatòries que no són estrictament positives, amb la condició que $E(X_+) < \infty$ i $V(X_+) < \infty$, on $X_+ = \{X \mid X > 0\}$. Però, per simplicitat, sempre es poden traslladar les dades a l'origen per tenir una mostra positiva i aplicar els resultats que s'han presentat fins al moment.

Sobre les restriccions del *CV-plot*, per al cas de la GPD, per exemple, només es pot calcular quan $\xi < 1/2$ i, per a una distribució t de Student, quan $\nu > 2$. Si el que es vol és calcular intervals de confiança, les condicions encara són més restrictives, ja que es necessiten moments finits fins a quart ordre, veure Teorema 2.1.

Per al cas de la GPD només es poden calcular intervals de confiança si $\xi < 1/4$. Per aquest motiu, en alguns *CV-plots*, apareix representada la funció constant corresponent a $cv_{1/4} = \sqrt{2} \approx 1{,}414$, per tal de marcar el límit fins al qual es pot utilitzar la teoria asimptòtica.

Malgrat aquestes limitacions, en el cas de la GPD es pot fer servir una propietat de dualitat existent entre les distribucions Fréchet i Weibull, que es pot traslladar a les distribucions GPD amb cua pesada i les de suport compacte, la qual permetrà realitzar una extensió del *CV-plot* per als casos on, en un principi, aquest no existeix. En el següent capítol es presentaran els resultats teòrics que permetran realitzar una transformació adient en els casos en els quals no hi ha existència dels moments necessaris.

CAPÍTOL 3

Dualitat entre cues lleugeres i pesades

Les distribucions de Fréchet i Weibull es troben estretament relacionades, ja que, sota la transformació -1/X, es pot passar d'una a l'altra, veure Embrechts *et al.* (1997) Secció 3.3.2. Per tant apareix una dualitat entre subfamílies de la GEV, més concretament, entre les que tenen índex del valor extrem positiu i les que el tenen negatiu, però això no passa amb la GPD.

Ara bé, ja que aquestes dues famílies es troben relacionades en el límit i mitjanant el domini d'atracció del màxim, es podria plantejar la qüestió de si hi ha alguna relació entre dominis d'atracció del màxim de variables aleatòries i com afectaria al cas concret de la GPD.

En aquest capítol es veurà que aquesta relació existeix i que servirà per trobar una relació de dualitat entre cues lleugeres i cues pesades. A més a més, sota una transformació prèvia, aquesta relació es farà servir per realitzar l'extensió del *CV-plot* per als casos on no hi ha moments finits suficients.

3.1 Relació entre els dominis d'atracció del màxim de H_ξ i $H_{-\xi}$

En el primer capítol d'aquesta tesi s'ha definit el concepte de domini d'atracció del màxim, Definició 1.2, i s'han presentat les condicions suficients de variació regular que ha de satisfer una distribució F per a pertànyer al domini d'atracció del màxim d'una distribució Fréchet o una distribució Weibull, veure Proposició 1.1 i Proposició 1.2. Totes dues condicions necessiten que una certa funció sigui $RV_{-\alpha}$ amb $\alpha > 0$, en el cas

Fréchet aquesta funció és $\bar{F}$ mentre que en el cas Weibull és $\bar{F}(x_F - x^{-1})$.

A partir de la inversió $-1/X$ les distribucions de Fréchet es poden transformar en distribucions de Weibull i a l'inrevés, per tant, és d'esperar que existeixi també una relació entre els dominis d'atracció del màxim de les distribucions Fréchet i Weibull. Els resultats que es presenten continuació defineixen aquesta relació.

Corol·lari 3.1. *Si $F \in D(H_\xi),\ \xi > 0$, on F és la funció de distribució de la v.a. X, llavors $F^* \in D(H_{-\xi})$, on F^* és la funció de distribució de la v.a. $X^* = -1/X$.*

Demostració. Sigui X una v.a. amb funció de distribució F tal que $F \in D(H_\xi),\ \xi > 0$, i sigui $X^* = -1/X$.

$F^*(x) = F(-1/x)$, on F^* és la funció de distribució de $X^* \Rightarrow \overline{F^*}(x) = \overline{F}(-1/x)$.

Per una altra banda, $x_F = +\infty$, perquè $F \in D(H_\xi),\ \xi > 0$, llavors

$$x_{F^*} = sup\{x \in \mathbb{R} :\ F^*(x) < 1\} = sup\{x \in \mathbb{R} :\ F(-1/x) < 1\} = 0$$

i $\overline{F^*}(x_{F^*} - x^{-1}) = \overline{F^*}(-x^{-1}) = \overline{F}(x)$.

Per la Proposició 1.1,

$$\overline{F^*}(x_{F^*} - x^{-1}) = \overline{F}(x) = x^{-1/\xi}L(x),\ \xi > 0\ \Rightarrow \overline{F^*}(x_{F^*} - x^{-1}) \in RV_{-1/\xi},$$

i per la Proposició 1.2,

$$\overline{F^*}(x_{F^*} - x^{-1}) \in RV_{-1/\xi} \Longleftrightarrow F^* \in D(H_{-\xi}).$$

□

Corol·lari 3.2. *Si $F \in D(H_{-\xi}),\ \xi > 0$, on F és la funció de distribució de la v.a. X, llavors $F^* \in D(H_\xi)$, on F^* és la funció de distribució de $X^* = x_F - 1/X$.*

Demostració. Sigui X una v.a. amb funció de distribució F tal que $F \in D(H_{-\xi}),\ \xi > 0$, i sigui $X^* = x_F - 1/X$.

$F^*(x) = F(x_F - 1/x)$, on F^* és la funció de distribució de $X^* \Rightarrow \overline{F^*}(x) = \overline{F}(x_F - 1/x)$.

Per la Proposició 1.2,

$$\overline{F^*}(x) = \overline{F}(x_F - 1/x) = x^{-1/\xi}L(x),\ \xi > 0 \Rightarrow \overline{F^*}(x) \in RV_{-1/\xi}.$$

Llavors, per la Proposició 1.1,

$$\overline{F^*}(x) \in RV_{-1/\xi} \Longleftrightarrow F^* \in D(H_\xi)$$

□

Aquests resultats, permeten realitzar transformacions de manera que distribucions sota el domini d'atracció del màxim Fréchet es transformen en distribucions sota el domini d'atracció del màxim Weibull i viceversa. A més a més, com que es té que si la distribució límit del màxim és GEV amb paràmetre ξ (GEV_ξ), llavors la distribució límit dels excessos sobre un llindar serà GPD amb el mateix paràmetre (GPD_ξ), això permet ampliar el resultat per a aquestes darreres.

Corol·lari 3.3. *Sigui X una v.a. tal que el límit de la seva distribució residual sobre un llindar es comporta com una GPD amb índex del valor extrem ξ, llavors el límit de la distribució residual sobre un llindar de $X^* = -1/X$ és una GPD amb índex del valor extrem $-\xi$.*

D'aquesta manera, el problema de no tenir moments d'ordre 1 i 2 per a que l'*ME-plot* i el *CV-plot* siguin consistents (GPD amb $\xi > 1$ i $\xi > 1/2$, respectivament) queda pal·liat. En cas de trobar-se en una situació problemàtica, només cal aplicar la transformació $X^* = -1/X$, fer servir les eines gràfiques o estadístiques que es considerin adients per trobar l'índex del valor extrem ξ de la distribució límit GPD i llavors assignar a X una GPD amb el mateix paràmetre però canviat de signe.

Exemple 3.1. Considerem una GPD amb paràmetre $\xi = 1$. La distribució de Cauchy, per exemple, té com a distribució límit per sobre d'un llindar aquesta distribució.

Sigui X una variable aleatòria estàndar de Cauchy amb la següent funció de densitat

$$f(x) = \frac{1}{\pi(1 + x^2)},\ x \in \mathbb{R}. \tag{3.1}$$

S'ha generat una mostra de mida 200 d'aquesta variable aleatòria, fent servir la funció `rcauchy` del paquet `stats` de l'R, i s'han representat els *CV-plots* de la mostra original i de la transformació $-1/X$ traslladada a l'origen.

A la Figura 3.1 es pot observar com el *CV-plot* de la mostra original no aporta cap informació, la gràfica no presenta cap regió més o menys constant a partir de cap llindar i es troba molt per sobre del límit a partir del qual ja no funciona la teoria asimptòtica ($\xi = 1/4$), en canvi, el *CV-plot* de la mostra transformada sí que presenta una regió més o menys constant dins de l'interval de confinaça de la distribució uniforme, que correspon a una GPD amb $\xi = -1$.

La cua de la Cauchy és GPD_1 i la de la seva transformació $-1/X$ és GPD_{-1}, es satisfà, per tant, el resultat demostrat en el Corol·lari 3.1.

3.2 Transformació $Y = -1/(X + c) + 1/c$

La GPD(ξ,ψ) es troba estandaritzada ja que totes les seves observacions prenen valors positius, el suport de la distribució és (0,σ) on $\sigma = \infty$ per a $\xi \geq 0$ i $\sigma = \psi/|\xi|$

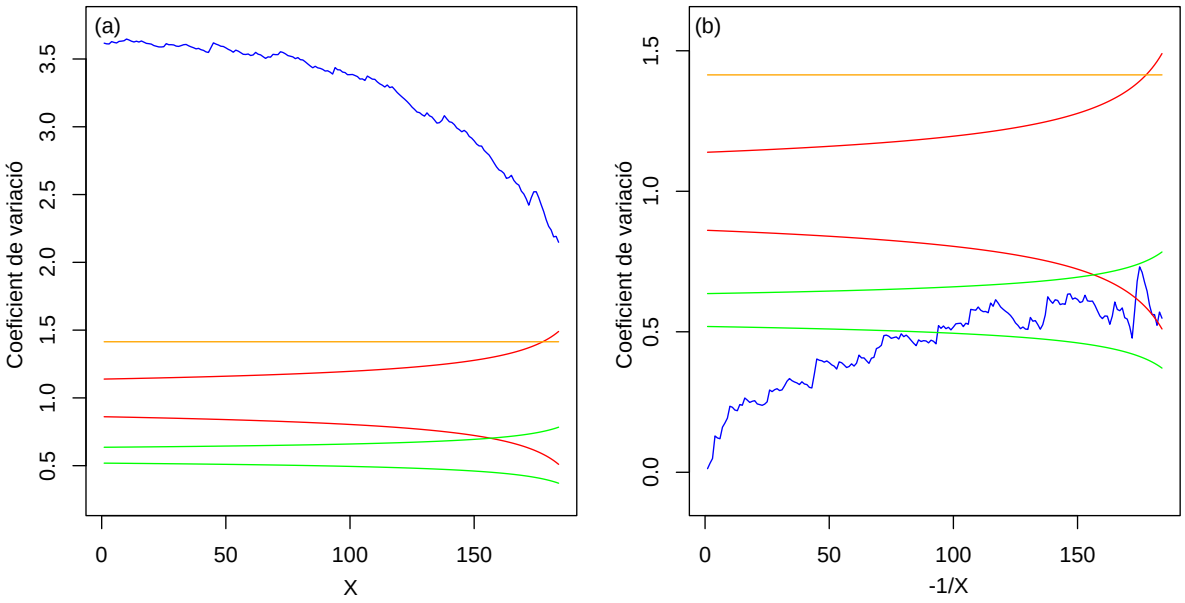

Figura 3.1: (a) *CV-plot* d'una mostra de mida 200 d'una variable aleatòria Cauchy estàndar. (b) *CV-plot* de les mateixes dades un cop aplicada la transformació $-1/X$. En ambdós casos les línies verdes i vermelles corresponen als intervals de confiança del 95% per a les distribucions uniforme i exponencial, respectivament, i la línia taronja al CV residual per a una $GPD_{1/4}$.

per a $\xi < 0$. La GPD es pot expandir incloent un paràmetre de localització mitjançant $Y = X + \mu$, de manera que el comportament de X a prop de σ serà el mateix que el de Y a prop de $\sigma + \mu$. De la mateixa manera, es pot expandir la transformació $X^* = -1/X$ a $Y = -1/(X + c)$, donat $c \geq 0$ o $c \leq -\sigma$, per a que la transformació romangui monòtona creixent a l'interval $(0, \sigma)$.

Teorema 3.1. *Sigui X una v.a. GPD(ξ,ψ) amb suport (0,σ) i $c \geq 0$ o $c \leq -\sigma$, llavors $Y = -1/(X + c)$ té distribució GPD amb paràmetre de localitzció si i només si $c = \psi/\xi$. A més a més, $Z = Y + 1/c$ té distribució GPD(-ξ,ξ^2/ψ).*

Demostració. Si X és una v.a. GPD estàndar, la transformació $Y = -1/(X + c)$, on $c \in \mathbb{R}$, té densitat

$$f_Y(y) = \psi^{1/\xi} \left(y/((\psi - \xi c)y - \xi)\right)^{1+1/\xi} y^{-2}, \ \psi > 0$$

amb suport a $-1/c \leq y < 0$, si $\xi \geq 0$, i a $-1/c \leq y < -1/(c + \sigma)$ si $\xi < 0$.

Si es considera $c = \psi/\xi$, la densitat de la variable Y,amb el mateix suport, pren la forma

$$f_Y(y) = \psi^{1/\xi} \left(y/ - \xi\right)^{1+1/\xi} y^{-2}, \ \psi > 0.$$

Donat el canvi $Z = Y + \xi/\psi = -1/(X + \psi/\xi) + \xi/\psi$, es pot veure fàcilment que la seva funció de densitat és

$$f_Z(z) = 1/(\xi^2/\psi)\left(1 + (-\xi)x/(\xi^2/\psi)\right)^{-1-1/(-\xi)}$$

amb suport a $0 \leq z \leq \xi/\psi$, si $\xi > 0$, i a $0 \leq z < \infty$ si $\xi < 0$.

□

Corol·lari 3.4. *Sigui $\xi > 0$, $\psi > 0$ i $c = \psi/\xi$, llavors una v.a. x té distribució GPD(ξ,ψ) si i només si $Z = X/(c(X+c))$ té distribució GPD(ξ_z,ψ_z) amb $\xi_z = -\xi$, $\psi_z = \xi^2/\psi$ i suport (0, ξ/ψ).*

Demostració. En el sentit cap a la dreta, es pot veure directament a partir del Teorema 3.1, perquè $c > 0$ i $Z = X/(c(X+c)) = -1/(X+c) + 1/c$.

El sentit cap a l'esquerra també és conseqüència del Teorema 3.1, perquè la inversa de la transformació anterior és

$$X = c^2 Z/(1-cZ) = Z/(c_2(Z+c_2)) = -1/(Z+c_2) + 1/c_2$$

on $c_2 = -1/c = -\xi/\psi$. El suport de Z és $(0,\ \psi/|\xi|)$=$(0,\ \xi/\psi)$ i $Z + c_2 > 0$ (equivalentment $c_2 \leq -\xi/\psi$), llavors X és una funció monòtona creixent de Z i el Teorema 3.1 prova el resultat. □

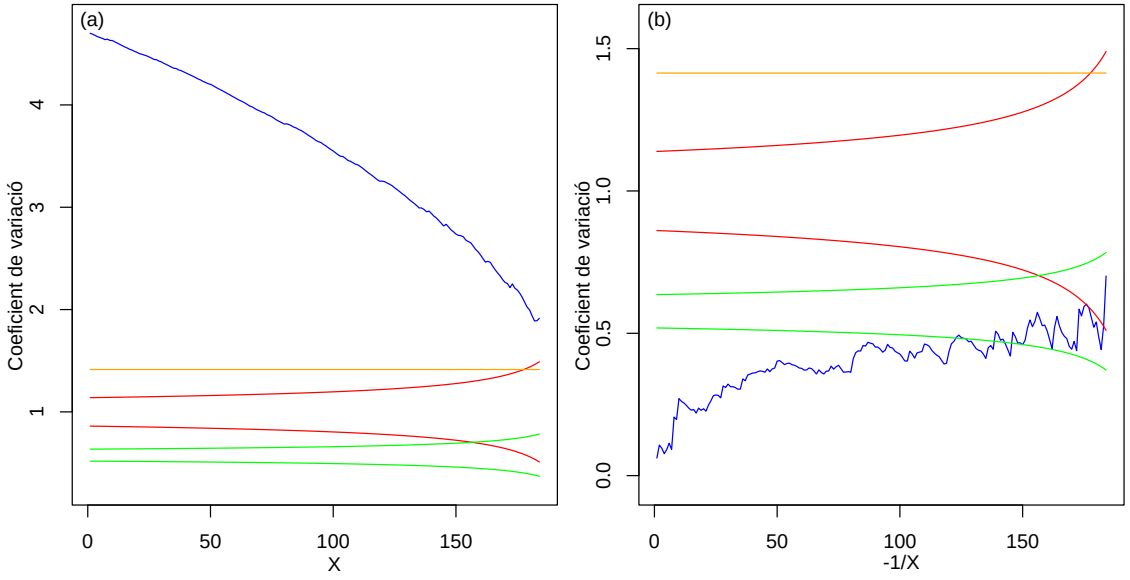

Figura 3.2: (a) *CV-plot* d'una mostra de mida 200 d'una variable aleatòria GPD_1. (b) *CV-plot* de les mateixes dades un cop aplicada la transformació $-1/X$. En ambdós casos les línies verdes i vermelles corresponen als intervals de confiança del 95% per a les distribucions uniforme i exponencial, respectivament, i la línia taronja al CV residual per a una $GPD_{1/4}$.

A continuació s'il·llustrarà, fent servir diferents *CV-plots*, la importància d'aplicar la transformació $Y = -1/(X+c) + 1/c$ amb el valor de c adequat. S'ha generat

una mostra de mida 200 d'una GPD_1, s'ha aplicat la trasnformació $Y = -1/X$ i, a continuació, s'han dibuixat els *CV-plots* corresponents. Per simular una mostra d'una variable aleatòria GPD es pot fer servir la funció `rgpd` del paquet `evir` de l'R.

A la Figura 3.2 es pot veure com, tot i que la simulació inicial és GPD_1, la seva transformació no presenta un *CV-plot* constant des de l'inici del gràfic, però sí que acaba entrant dins de l'interval de confiança d'una GPD_{-1}, que correspon a la distribució uniforme.

Si es vol aplicar la transformació $Y = -1/(X + c) + 1/c$ amb $c = \psi/\xi$, primer s'han d'estimar els paràmetres, a la pràctica es pot fer mitjançant ML, per exemple. A la Figura 3.3 es pot veure com, un cop estimats els paràmetres ξ i ψ per ML i aplicada la transformació $Y = -1/(X + \psi/\xi) + \xi/\psi$ a les dades utilitzades a la Figura 3.2, el *CV-plot* sí que es correspon amb el d'una GPD_{-1} des del seu inici.

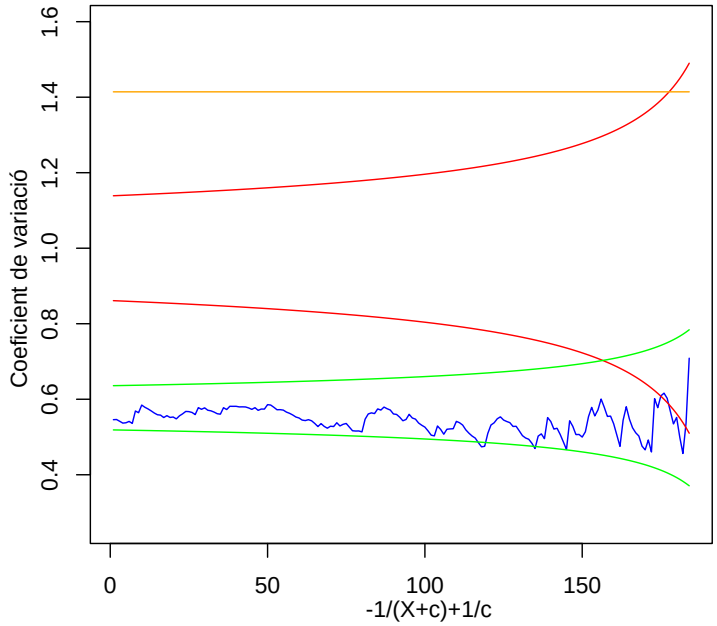

Figura 3.3: *CV-plot* de la mostra GPD_1 de la Figura 3.2 després d'aplicar-li la transformació $Y = -1/(X + \psi/\xi) + \xi/\psi$.

3.3 Exemple amb les dades daneses

A la Figura 1.5 es podia veure com el *CV-plot* de les dades daneses no era capaç d'aportar informació degut a la falta de moments. La gràfica començava amb un valor molt elevat i anava decreixent, però quedant bastant per sobre de l'interval de confiança de la distribució exponencial.

A continuació, s'aplicarà la darrera transformació presentada a l'apartat anterior a aquestes dades per poder extreure'n més informació. Es consideraran les pèrdues per sobre de 10 milions de corones daneses, 109 dades exactament, per realitzar els càlculs, ja que és el llindar que es proposa a McNeil (1997) a partir d'analitzar l'*ME-plot* de les dades, veure Figura 1.3.

Abans de tot, cal traslladar les dades a l'origen per poder realitzar les estimacions dels paràmetres mitjançant ML, les quals són $\widehat{\xi} = 0{,}5$ i $\widehat{\psi} = 7$. Seguidament, amb aquesta informació, ja es pot realitzar la transformació $Y = -1/(X + \psi/\xi) + \xi/\psi$.

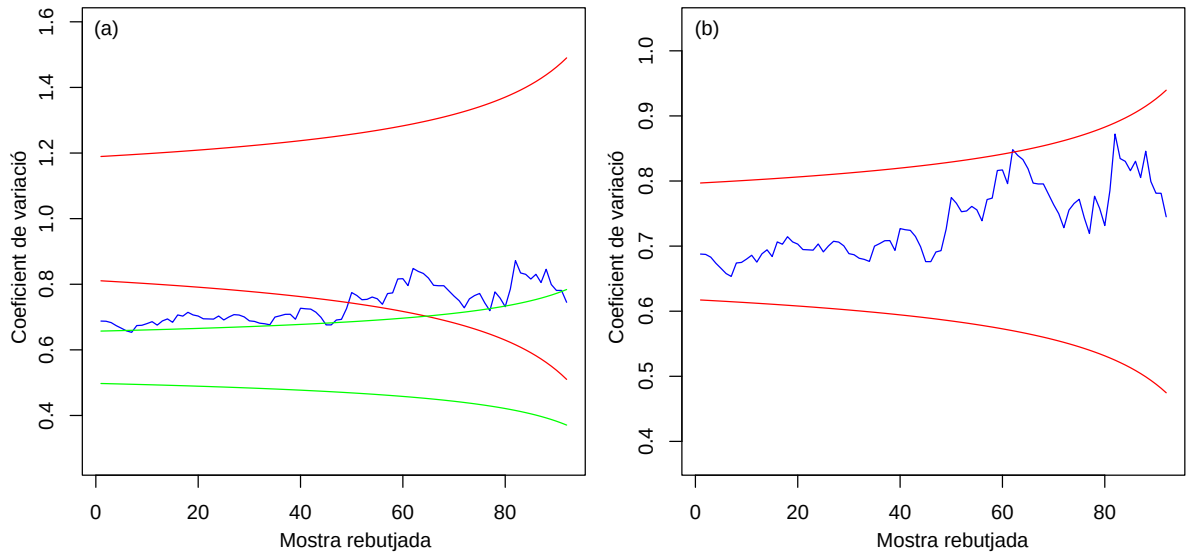

Figura 3.4: (a) *CV-plot* de les dades daneses per sobre de 10 milions de pèrdues després d'aplicar-li la transformació $Y = -1/(X + \psi/\xi) + \xi/\psi$. Les línies verdes i vermelles corresponen als intervals de confiança del 95% de les distribucions uniforme i exponencial, respectivament. (b) *CV-plot* per a la mateixa transformació amb l'interval de confiança del 95% per a una GPD amb $\xi = -\,0{,}5$.

A la Figura 3.4, a l'esquerra, es pot veure el *CV-plot* de la transformació amb els intervals de confiança de les distribucions uniforme i exponencial, es pot observar com la gràfica queda totalment acotada per aquestes dues distribucions, de manera que la cua ha de correspondre a una distribució GPD amb $-1 \leq \xi \leq 0$, que és una distribució de suport compacte, tal i com era d'esperar. A la dreta es pot veure el *CV-plot* de la mateixa transformació però amb l'interval de confiança del 95% per a una GPD amb $\xi = -\,0{,}5$, el qual conté la majoria dels valors dels CV residuals.

La conclusió a la qual es pot arribar després d'aplicar aquests canvis és que una GPD amb paràmetre ξ al voltant de 0,5 és una estimació adient per a la cua de la distribució, resultat que és coherent amb els resultats proposats per McNeil (1997). Cal recordar que els mètodes gràfics són eines que permeten realitzar una visió descriptiva de les dades i la subjectivitat de l'investigador afecta en gran mesura a les conclusions, es necessiten

eines d'inferència per poder realitzar estudis més acurats de les dades i les estimacions dels paràmetres associats.

En el següent capítol es presentaran els tests T_m que permetran realitzar la inferència necessària sobre les dades. Aquests tests es podran fer servir tant per contrastar la bondat de l'ajust com per estimar el llindar a partir del qual es pot considerar que la distribució és GPD mitjançant un algoritme automàtic de selecció de llindar, el qual s'aplicarà a les dades daneses per completar l'anàlisi que s'està realitzant.

CAPÍTOL 4

Contrastos amb múltiples llindars

Castillo *et al.* (2014a) van presentar un nou test per a exponencialitat, anomenat T_m, que utilitza múltiples llindars escollits mitjançant les successives potències de 1/2. Aquest test múltiple es pot extendre a qualsevol GPD que tingui moments d'ordre 2, ja que es necessita existència de variància, i es pot ampliar la manera d'escollir els llindars utilitzant potències de p amb $0 < p < 1$.

És important poder disposar d'un test per recolzar els mètodes gràfics. A més a més, un test de múltiples llindars avalua la hipòtesi nul·la al llarg de tota la cua i no només en un punt. Com ja s'ha comentat, existeixen tests basats en el CV que poden portar a conclusions errònies. Per exemple, la distribució exponencial és l'única que té CV residual a 1, però no és l'única que té CV igual a 1, es poden trobar exemples tant de cues pesades com de suport compacte que compleixen aquesta propietat. Un test basat en el CV no sempre pot detectar les possibles diferències, mentre que un test basat en múltiples llindars té menys probabilitat d'error.

En aquest capítol es presentarà l'extensió de l'estadístic T_m, anomenat $T_{m,p}$. Considerar la distribució com a procés en funció del llindar permet trobar les distribucions asimptòtica i aproximada de $T_{m,p}$. Per finalitzar, es presentaran algunes simulacions realitzades per comparar l'aproximació als valors asimptòtics segons la mida mostral.

4.1 Tests T_m i estimació

Donada una mostra $\{x_j\}$ d'una GPD_ξ, per a qualsevol conjunt possible de llindars $t_0 < t_1 < ... < t_m$, siguin $n(t_k)$ el nombre d'observacions del conjunt $\{x_j : x_j > t_k\}$,

$cv(t_k)$ el CV empíric del conjunt traslladat a l'origen donat per (1.30), on $0 \leq k \leq m$, i cv_ξ el CV teòric donat per (1.32).

Es defineix l'estadístic T basat en aquesta selecció de múltiples llindars com

$$T = \sum_{k=0}^{m} n(t_k) \left(cv(t_k) - cv_\xi\right)^2 . \tag{4.1}$$

La distribució de T no depèn del parametre d'escala i és senzill simular-la. És important notar que, sota la hipòtesi nul·la de GPD_ξ, s'esperen valors petits per a T i valor proper a $\sqrt{1/(1-2\xi)}$ per a cada $cv(t_k)$. Per tant, valors alts per a T mostrarien que una GPD amb índex del valor extrem ξ no seria un model adient per a les dades.

Els llindars $\{t_k\}$ poden ser arbitraris, però es pot obtenir una simplificació pràctica agafant llindars de manera que la probabilitat de la cua sigui p^k (potència de p), sota la hipòtesi nul·la de GPD. Donada una mostra $\{x_j\}$ de mida n, independentment de la distribució de la qual provingui, $x_{(0)} = 0$, $x_{((1-p)n)}$, $x_{((1-p^2)n)}$, $x_{((1-p^3)n)}$, ... són els llindars que aproximadament deixen per sobre d'ells les probabilitats 1, p, p^2, p^3..., per tant, aquests són els valors que es necessiten per treballar amb el conjunt de llindars escollit. Per a una mostra general, el quantils q_k corresponents als darrers np^k elements són considerats (si $p = 1/2$, per exemple, q_1 és la mediana, q_2 és el tercer quartil, ...).

Per a una variable aleatòria GPD amb índex del valor extrem ξ, $Pr\left\{X > \frac{\psi}{\xi}\left(p^{-k\xi} - 1\right)\right\} = p^k$, per tant, $q_k \approx \frac{\psi}{\xi}\left(p^{-k\xi} - 1\right) \approx x_{(n-np^k)}$ (per al cas $\xi = 0$, distribució exponencial amb paràmetre μ, es té que aquests valors seran aproximadament equiespaiats ja que $Pr\{X > (-\mu log(p))k\} = p^k$).

Agafant el conjunt de llindars corresponent als quantils de la mostra que s'acaben de descriure, (4.1) esdevé

$$T_{m,p} = n \sum_{k=0}^{m} p^k \left(cv(q_k) - cv_\xi\right)^2 \tag{4.2}$$

Es pot observar que hi ha una gran quantitat de possibilitats a l'hora d'escollir un estadístic $T_{m,p}$. Per tal que s'hagi d'escollir el mínim número de paràmetres i l'estudi sigui el més objectiu possible, es proposa una partició concreta a partir de l'únic valor de m escollit per l'investigador, de manera que la p es fixarà a partir d'aquest valor.

Un cop escollit el valor de m, el valor de p es calcularia de la següent manera

$$p = \exp(\log(ns/n)/m) \tag{4.3}$$

on n és la mida de la mostra i ns la mida de la mostra a partir del llindar més gran, és a dir, la mínima mida de mostra que es faria servir en el càlcul dels CV residuals. Es recomana agafar $ns = 8$, però aquest és un valor que es pot modificar.

A partir d'ara, per simplicitat i perquè l'únic valor que cal escollir quan s'utilitza un test $T_{m,p}$ és m, a aquests tests se'ls anomenarà T_m en molts llocs.

La finalitat d'aquests tests T_m és fer-los servir en contrasts de l'estil:

$$\begin{cases} H_0 : \text{La distribució de la qual prové la mostra és } GPD_\xi \\ H_1 : \text{La distribució de la qual prové la mostra no és } GPD_\xi \end{cases}$$

També es poden fer servir en casos en els que l'índex del valor extrem és desconegut, agafant com a ξ el valor corresponent al cv_ξ que minimitza l'estadístic T_m. Aquest valor es pot trobar fàcilment derivant i igualant a 0, la seva expressió és

$$\hat{\xi} = (c\hat{v}_\xi{}^2 - 1)/(2c\hat{v}_\xi{}^2) \tag{4.4}$$

on $c\hat{v}_\xi = (1-p)\sum_{k=0}^{m} p^k cv(q_k)/(1-p^{m+1})$.

En aquesta situació, amb ξ desconegut, el contrast que es realitza és el següent:

$$\begin{cases} H_0 : \text{La distribució de la qual prové la mostra és GPD} \\ H_1 : \text{La distribució de la qual prové la mostra no és GPD} \end{cases}$$

Per tant, donada un mostra a partir d'un cert llindar fixat, amb un d'aquests tests es pot contrastar la hipòtesi de GPD per a aquesta.

Exemple 4.1. En acabar el capítol anterior, es recomanava una estimació per al paràmetre ξ de 0,5 en el cas de les dades daneses per sobre de 10 milions, basant-se en el *CV-plot*. Aquesta gràfica ha de servir de suport i per realitzar l'anàlisi descriptiva, però com a eina inferencial és millor utilitzar els tests que s'acaben de presentar.

A la Taula 4.1 es poden veure els resultats de realitzar el contrast per a ξ = -0,5 conegut amb els tests T_m, per a diferents valors de m, i el p-valor obtingut mitjançant simulació de 10^4 mostres, per a les 109 dades transformades. En tots els casos el p-valor és prou gran per concloure que no hi ha evidències significatives per poder rebutjar la hipòtesi nul·la de $GPD_{-0,5}$, en el cas de les dades re-transformades seria $\xi = 0,5$, valor que concorda amb el valor obtingut si s'estima el paràmetre per ML, veure McNeil (1997).

De la mateixa manera, es podria realitzar el contrast per a ξ desconegut. A la Taula 4.2 es poden veure els valors de l'estadístic T_m, els p-valors calculats per simulació de 10^4 mostres i l'estimació de ξ donada per (4.4). Com en el cas anterior, tots els p-valors són prou grans per concloure que no hi ha evidències significatives per poder rebutjar la hipòtesi nul·la de GPD, tot i que les estimacions de ξ es troben al voltant de -0,4 en comptes del valor -0,5.

En qualsevol dels dos casos, amb índex del valor extrem conegut o desconegut, no es pot rebutjar la hipòtesi de GPD per a les dades daneses per sobre de 10 milions de corones, fet que dóna més suport encara a les conclusions de McNeil (1997) sobre aquestes dades.

Taula 4.1: Estadístic T_m i p-valor per al contrast amb ξ conegut igual a -0,5 aplicat a la transformació de les dades daneses per sobre de 10 milions de corones.

m	$\mathbf{T_m}$	**p-valor**
5	1.05	0.61
10	1.87	0.64
15	2.71	0.64
20	4.38	0.50

Taula 4.2: Estadístic T_m i p-valor per al contrast amb ξ desconegut aplicat a la transformació de les dades daneses per sobre de 10 milions de corones, també es mostra el valor de ξ que minimitza l'estadístic i que serveix com a estimador.

m	$\mathbf{T_m}$	**p-valor**	$\hat{\boldsymbol{\xi}}$
5	0.86	0.58	-0.43
10	1.24	0.75	-0.40
15	1.86	0.75	-0.40
20	2.98	0.63	-0.39

4.1.1 Estimació del paràmetre ξ

Existeixen diversos mètodes per a l'estimació del paràmetre ξ d'una GPD; el més conegut és l'estimació ML però es poden trobar altres mètodes estudiats per Hosking i Wallis (1987), Castillo i Hadi (1997), Zhang i Stephens (2009) o Song i Song (2012), entre d'altres. Malgrat disposar de totes aquestes opcions, s'ha vist que aquestes tècniques no sempre són factibles.

Al Capítol 1 ja s'ha comentat que el més habitual és fer servir ML per a l'estimació del paràmetre ξ un cop s'ha escollit un llindar, veure Coles (2001), i, per al cas de la bondat de l'ajust, Choulakian i Stephens (2001) recomanen també aquest mètode d'estimació en el cas de paràmetres desconeguts abans d'aplicar els tests de Cramér-von Mises o d'Anderson-Darling.

Els tests T_m es poden utilitzar per estimar el valor del paràmetre ξ a partir de la seva minimització, obtenint una fórmula tancada, veure (4.4), i donat que l'estimador ML per a la GPD és un estimador eficient però que pot tenir problemes quan les mostres són de mida petita, veure Castillo i Daoudi (2009), en aquest apartat, es compararan l'estimació ML amb l'estimació fent servir diferents tests T_m per a mostres petites.

L'estimació utilitzant els tests T_m es realitzarà fent servir els valors $m = 0$, $m = 2$ i $m = 10$, amb una significació del 20%. El cas particular T_0 correspon a calcular l'estimació del paràmetre ξ a partir del coeficient de variació empíric de tota la mostra

sota la hipòtesi de GPD, donant com a resultat l'estimació

$$\hat{\xi} = (1 - 1/cv^2)/2 \tag{4.5}$$

on cv ve donat per l'equació (1.30).

A més a més, aquesta estimació coincideix amb la que sorgeix fent servir el mètode dels moments (MOM), presentat per Hosking i Wallis (1987). Degut a aquesta coincidència i que es vol deixar clara la diferència entre fer servir el coeficient de variació simplement o utilitzar múltiples llindars, a aquest mètode d'estimació se l'anomenarà CV en comptes de T_0.

Es compararan, per tant, les estimacions del paràmetre ξ mitjançant ML, CV, T_2 i T_{10}. A la Taula 4.3 es poden veure els errors quadràtics mitjans sobre 1.000 simulacions per a aquestes estimacions, per a diverses mides de mostra i diferents valors del paràmetre ξ.

Es pot concloure que, per a mostres de mida petita, els mètodes que millor funcionen són T_2 i T_{10} quan $\xi < 0$, el primer quan la mida és més petita o igual que 50 i el segon quan es va augmentant la mida de la mostra, i, quan $\xi \leq 0$, T_2 continua essent el millor mètode mentre que CV és el que té un error més petit quan es va augmentant la mida de la mostra. Es pot observar també com, per a mostres de mida 200 i $\xi = 0{,}3$ l'MSE més petit correspon al mètode MLE, quan la mostra es va fent gran el millor mètode és MLE, però en mostres no tan grans, quan el test T_m pot fallar degut a la falta de moments, és normal que aquest mètode comenci a donar millors resultats.

4.1.2 Estimació del paràmetre ψ

Els mètodes d'estimació existents que s'han mencionat fins ara, MLE o MOM per exemple, proporcionen una estimació tant per a l'índex del valor extrem ξ com per al de forma ψ d'una GPD, cosa que no succeeix amb els mètodes T_m, ja que el coeficient de variació no depèn del paràmetre d'escala.

La qüestió que sorgeix llavors és quina és la millor manera d'estimar ψ un cop s'han fet servir els mètodes T_m. Cal recordar que aquest valor és necessari si es vol aplicar el mètode PoT per trobar l'estimació completa de la distribució de la cua, equació (1.18).

Per intentar respondre aquesta pregunta, s'han comparat sis mètodes diferents d'estimació del paràmetre ψ donat que ξ sigui conegut.

- MLE. Maximitzant la funció de log-versemblança, $l(\xi, \psi)$, respecte ψ, equació (1.27). Per realitzar aquesta tasca es fa servir la funció `optimize` del paquet `stats` de l'R.

- LE. Igualant l'equació de versemblança, $\partial l/\partial\psi$, a 0 i aïllant ψ. Per realitzar aquesta tasca es fa servir la funció `nleqslv` del paquet `nleqslv` de l'R.

Taula 4.3: Comparació de l'error quadràtic mitjà d'estimacions de ξ per a mostres de diverses mides petites i diferents valors de l'índex del valor extrem d'una GPD. Els mètodes fets servir són: el coeficient de variació (CV), màxima versemblança (MLE) i T_m amb $m = 2$ i $m = 10$. En negreta està marcat el valor més petit corresponent a cada una de les diferents combinacions.

Mètode	**Mida**	$\boldsymbol{\xi}$								
		-0.5	**-0.4**	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**
CV	20	13.10	10.39	9.03	7.79	6.80	6.66	6.96	7.26	8.65
MLE	20	34.27	29.97	26.15	23.62	20.46	18.14	18.16	17.24	17.62
T_2	20	**5.79**	**5.18**	**4.77**	**4.58**	**4.47**	**4.85**	**5.68**	**6.74**	**8.57**
T_{10}	20	6.02	5.42	5.16	4.93	4.80	5.24	6.10	7.17	9.16
CV	30	7.68	6.39	5.09	4.26	3.93	3.83	4.17	4.65	**5.60**
MLE	30	12.68	10.12	8.43	7.52	7.26	7.28	7.79	8.19	8.88
T_2	30	3.81	**3.21**	**2.78**	**2.51**	**2.54**	**2.79**	**3.36**	**4.34**	5.82
T_{10}	30	**3.75**	3.33	3.04	2.88	2.96	3.30	3.93	5.02	6.73
CV	50	4.17	3.34	2.72	2.39	2.15	2.16	2.31	2.61	**3.31**
MLE	50	3.45	3.25	2.84	3.00	3.15	3.36	3.59	3.89	4.44
T_2	50	**2.15**	**1.89**	**1.62**	**1.64**	**1.68**	**1.92**	**2.30**	**2.87**	3.94
T_{10}	50	2.16	1.86	1.65	1.67	1.76	2.07	2.45	3.13	4.36
CV	100	1.92	1.58	1.30	1.12	1.00	1.02	**1.16**	**1.33**	**1.76**
MLE	100	1.09	1.05	1.05	1.11	1.22	1.34	1.55	1.70	1.94
T_2	100	1.17	0.99	0.85	**0.83**	**0.87**	**0.97**	1.26	1.57	2.16
T_{10}	100	**0.90**	**0.83**	**0.80**	0.85	1.00	1.17	1.54	1.94	2.67
CV	200	0.95	0.74	0.62	0.52	0.48	**0.51**	**0.58**	**0.74**	1.01
MLE	200	0.41	0.42	0.42	0.46	0.52	0.60	0.67	0.78	**0.90**
T_2	200	0.65	0.52	0.44	0.41	**0.44**	0.52	0.66	0.89	1.26
T_{10}	200	**0.41**	**0.38**	**0.37**	**0.40**	0.51	0.64	0.85	1.14	1.64

- f0. Resolent l'equació $f(\psi) = 0$ respecte ψ donada

$$f(\psi) = 1/(1+\psi) - 1/n \sum_{i=0}^{n} 1/(1+\xi x_i/\psi) \tag{4.6}$$

  on aquesta expressió es dedueix de simplificarl'equació de versemblança i sumar 1 a cada terme. Per realitzar aquesta tasca es fa servir també la funció `nleqslv` de l'R.

- Minf2. Minimitzant el quadrat de $f(\psi)$. Per realitzar aquesta tasca es fa servir també la funció `optimize` de l'R.

- Mediana. Aïllant ψ de l'expressió de la mediana d'una GPD s'obté l'estimació

$$\hat{\psi} = \xi Me/(2^{\xi} - 1). \tag{4.7}$$

  on Me és la mediana empírica de la mostra.

- Mitjana. Aïllant ψ de l'expressió de la mitjana d'una GPD s'obté l'estimació

$$\hat{\psi} = \bar{X}(1-\xi). \tag{4.8}$$

  on $\bar{X}$ és la mitjana empírica de la mostra.

Els MSE s'han calculat sobre 10.000 mostres simulades de cada una de les diferents condicions plantejades. S'han considerat mostres de mides 20, 30, 50, 100 i 200, també s'han utilitzat diferents valors del paràmetre ξ i diferents valors del parà metre ψ. Ara bé, com que els resutats són anàlegs per als diferents valors del paràmetre ψ, per simplicitat, només es mostraran els resultats obtinguts per a les simulacions quan $\psi = 1$.

A la Taula 4.4 es pot veure com l'estimació per ML és la que té un MSE menor en tots els casos. Per a $\xi < 0$, clarament és la millor, per a $\xi = 0$, tots els mètodes són equivalents excepte el de la mediana i, per a $\xi > 0$, es pot veure que hi ha més d'un mètode que funciona millor.

Per tant, un cop fets servir els mètodes T_m i quan sigui necessari estimar el valor del paràmetre ψ, el mètode que es farà servir serà MLE sobre ψ suposant ξ conegut.

4.2 Distribució asimptòtica

És possible escriure l'estadístic $T_{m,p}$ de la forma $T_{m,p} = V'V$, on

$$V' = \sqrt{n}\left[(cv(q_0) - cv_\xi)\,,\, p^{1/2}\,(cv(q_1) - cv_\xi)\,,\, \ldots\,,\, p^{m/2}\,(cv(q_m) - cv_\xi)\right] \tag{4.9}$$

i $q_k = Q(1-p^k)$, on $Q(p)$ ve donada per (1.7).

Taula 4.4: MSE per als diferents mètodes d'estimació del paràmetre ψ utilitzats, fent servir diferents mides de mostra i diferents valors del paràmetre ξ per a mostres simulades, totes elles amb $\psi = 1$.

Mètode	**Mida**	**ξ**								
		-0.5	**-0.4**	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**
MLE	20	**1.44**	**1.82**	**2.39**	**3.08**	**3.89**	**4.98**	**6.14**	**7.49**	**8.81**
LE	20	13.21	8.47	4.68	3.31	**3.89**	**4.98**	**6.14**	> 100	> 100
f0	20	13.19	8.47	4.68	3.31	**3.89**	**4.98**	**6.14**	8.37	21.93
Minf2	20	3.27	2.95	2.73	3.13	3.90	**4.98**	**6.14**	**7.49**	**8.81**
Mediana	20	6.65	7.27	7.84	8.76	9.28	10.24	11.35	12.71	13.64
Mitjana	20	2.51	2.78	3.12	3.60	4.09	**4.98**	6.26	8.43	12.34
MLE	30	**0.89**	**1.16**	**1.53**	**2.05**	**2.61**	**3.41**	**4.24**	**4.76**	**5.50**
LE	30	12.47	8.02	3.84	2.23	**2.61**	**3.41**	**4.24**	> 100	> 100
f0	30	12.46	8.01	3.84	2.23	**2.61**	**3.41**	**4.24**	**4.76**	5.68
Minf2	30	2.75	2.26	2.32	2.49	2.63	**3.41**	**4.24**	**4.76**	**5.50**
Mediana	30	4.56	4.79	5.35	5.74	6.27	7.02	7.62	8.09	8.78
Mitjana	30	1.67	1.82	2.06	2.38	2.73	3.41	**4.31**	5.38	7.95
MLE	50	**0.47**	**0.64**	**0.90**	**1.23**	**1.60**	**2.02**	**2.38**	**2.89**	**3.37**
LE	50	11.89	7.92	3.38	**1.41**	**1.60**	**2.02**	**2.38**	> 100	> 100
f0	50	11.88	7.92	3.38	**1.41**	**1.60**	**2.02**	**2.38**	**2.89**	3.41
Minf2	50	2.32	1.80	1.57	1.39	1.62	**2.02**	**2.38**	**2.89**	**3.37**
Mediana	50	2.76	2.96	3.27	3.68	3.88	4.17	4.42	4.95	5.33
Mitjana	50	0.99	1.09	1.25	1.45	1.67	**2.02**	2.46	3.33	5.17
MLE	100	**0.20**	**0.31**	**0.43**	**0.62**	**0.82**	**1.00**	**1.20**	**1.43**	**1.60**
LE	100	11.43	8.15	3.41	0.77	**0.82**	**1.00**	**1.20**	**1.43**	> 100
f0	100	11.43	8.15	3.41	0.77	**0.82**	**1.00**	**1.20**	**1.43**	1.69
Minf2	100	1.98	1.61	1.07	0.92	0.83	**1.00**	**1.20**	**1.43**	**1.60**
Mediana	100	1.42	1.55	1.70	1.80	1.93	2.07	2.22	2.44	2.55
Mitjana	100	0.51	0.56	0.63	0.73	0.85	**1.00**	1.26	1.68	2.49
MLE	200	**0.09**	**0.14**	**0.21**	**0.30**	**0.39**	**0.50**	**0.60**	**0.71**	**0.80**
LE	200	11.32	8.14	3.94	**0.45**	**0.39**	**0.50**	**0.60**	**0.71**	> 100
f0	200	11.32	8.14	3.94	**0.45**	**0.39**	**0.50**	**0.60**	**0.71**	**0.80**
Minf2	200	1.87	1.46	0.83	0.64	0.41	**0.50**	**0.60**	**0.71**	**0.80**
Mediana	200	0.70	0.78	0.84	0.92	0.96	1.05	1.11	1.20	1.28
Mitjana	200	0.25	0.28	0.31	0.36	0.41	**0.50**	0.62	0.84	1.23

La distribució asimptòtica de $T_{m,p}$ es pot trobar a partir del Teorema ari 2.2 de la següent manera. Es té $q_k \approx \frac{\psi}{\xi}\left(p^{-k\xi}-1\right)$, per tant, asimptòticament, la matriu de covariàncies per a V és

$$\Sigma_{m,p} = \left(p^{i/2}\rho_\xi(q_i,q_j)\,p^{j/2}\right)_{i,j=0,\dots,m} =$$

$$\left(\frac{p^{1/2((1-\xi)(i\vee j)+(2\xi-1)(i\wedge j))}\left(p^{-(i\vee j)\xi}\left(6\xi^2-9\xi+3\right)+p^{-(i\wedge j)\xi}\left(8\xi-2\right)\right)\left(\xi-1\right)^2}{\left(2\xi-1\right)^2\left(3\xi-1\right)\left(4\xi-1\right)}\right)_{i,j=0,\dots,m} \tag{4.10}$$

Teorema 4.1. *La distribució asimptòtica de $T_{m,p}$ és $\sum_{i=0}^m \lambda_i Z_i^2$ amb Z_i distribuits com a $N(0,1)$ independents i λ_i els valors propis de $\Sigma_{m,p}$.*

Demostració. Pel teorema central del límit V és asimptòticament normal multivariant $N(0,\Sigma_m)$. Llavors, amb un argument clàssic, $\Sigma_{m,p} = A\Lambda A'$ amb A matriu ortogonal i Λ la matriu diagonal dels valors propis. Es dóna que $V = A\Lambda^{1/2}Z$ amb Z asimptòticament normal multivariant amb la identitat com a matriu de covariància, $N(0,I)$. Llavors $T_{m,p} = V'V = Z'\Lambda Z = \sum_{i=0}^m \lambda_i Z_i^2$, perquè A és una matriu ortogonal. □

Exemple 4.2. Per a $m=2$, $p=1/2$ i $\xi=0$,

$$\Sigma_{2,1/2} = \begin{pmatrix} 1 & 1/\sqrt{2} & 1/2 \\ 1/\sqrt{2} & 1 & 1/\sqrt{2} \\ 1/2 & 1/\sqrt{2} & 1 \end{pmatrix}$$

i els valors propis vénen donats per

$$\lambda_0 = \left(5+\sqrt{17}\right)/4,\ \lambda_1 = 1/2,\ \lambda_2 = \left(5-\sqrt{17}\right)/4.$$

Cal notar també que, per a $m=0$, la distribució asimptòtica de T_0 és simplement una distribució χ_1^2.

Amb tantes combinacions possibles, és difícil comparar aquests estadístics, perquè cadascun d'ells té un valor esperat en funció del valor de m, però mitjançant una petita modificació es pot aconseguir que siguin comparables.

A partir de (2.37) es pot deduir fàcilment que

$$n(t)/\sigma_\xi^2\left(cv(t)-cv_\xi\right)^2 \xrightarrow{d} \chi_1^2. \tag{4.11}$$

Tot i que cada un dels sumands de $T_{m,p}$ no són independents, el que sí que es pot veure és que l'esperança de cada $T_{m,p}/\sigma_\xi^2$ és $m+1$, llavors es té que l'esperança de

$T_{m,p}/\left(\sigma_\xi^2(m+1)\right)$ és 1, independentment del valor de m. Per tant, dividint els estadístics $T_{m,p}$ per $\sigma_\xi^2(m+1)$, sí que es poden comparar per a diferents parelles de m i p.

Aquesta modificació de l'estadístic no suposa cap inconvenient a l'hora de trobar la seva distribució asimptòtica perquè és possible utilitzar $cT_{m,p}$, on $c \in \mathbb{R}$, i no cal tornar a fer tots els càlculs, ja que es poden relacionar els valors propis de $T_{m,p}$ amb els de $cT_{m,p}$.

Teorema 4.2. *Siguin λ_0, λ_1, ..., λ_m els valors propis de la matriu de covariàncies $\Sigma_{m,p}$ associada a $T_{m,p}$, llavors els valors propis de la matriu de covariàncies associada a $cT_{m,p}$ són $c\lambda_0$, $c\lambda_1$, ..., $c\lambda_m$.*

Demostració. Siguin els valors λ_0, λ_1, ..., λ_m les solucions de l'equació característica $|\Sigma_{m,p} - \lambda Id| = 0$, $\Sigma^*_{m,p} = \left((cp)^{i/2}\rho(q_i,q_j)\,(cp)^{j/2}\right)_{i,j=0,\dots,m} = c\Sigma_{m,p}$ la matriu de covariàncies associada a $cT_{m,p}$ i β_0, β_1, ..., β_m els valors propis de $\Sigma^*_{m,p}$.

Per trobar els valors propis de $\Sigma^*_{m,p}$ fem

$$|\Sigma^*_{m,p} - \beta Id| = |c\Sigma_{m,p} - \beta Id| = c^{m+1}|\Sigma_{m,p} - (\beta/c)Id| = 0$$

d'on es pot observar que $\lambda = \beta/c$ i per tant es pot escriure $\beta_i = c\lambda_i$ $i = 0, ..., m$, tal i com es volia veure. □

Exemple 4.3. Per a $m = 2$, $p = 1/2$ i $\xi = 0$, els valors propis de la matriu de covariàncies per a $T_{2,1/2}/\left(3\sigma_0^2\right)$ vénen donats per

$$\lambda_0 = \left(5 + \sqrt{17}\right)/12,\ \lambda_1 = 1/6,\ \lambda_2 = \left(5 - \sqrt{17}\right)/12.$$

4.3 Distribució aproximada

La distribució asimptòtica de $T_{m,p}$, donada pel Teorema 4.1, proporciona una manera de calcular els p-valors per a mostres grans sense simulació pesada. Per exemple, si la mida de la mostra és $n = 2.000$ i $m = 3$, el mètode directe necessita mostres de 2.000 nombres aleatoris GPD i la distribució asimptòtica només necessita mostres de 4 nombres aleatoris normals.

Per altra banda, la distribució asimptòtica de $T_{m,p}$, es pot aproximar per $a + b\chi^2_\nu$, on χ^2_ν té distribució gamma amb paràmetres $(2/\nu, 2)$ i les constants a, b i ν s'ajusten pels tres primers moments de $\Sigma_{i=0}^m \lambda_i Z_i^2$. Això porta a resoldre:

$$a + b\nu = \sum_{i=0}^{m} \lambda_i,\ b^2\nu = \sum_{i=0}^{m} \lambda_i^2,\ b^3\nu = \sum_{i=0}^{m} \lambda_i^3. \tag{4.12}$$

D'aquesta manera, es poden calcular p-valors amb menys esforç encara, ja que només cal simular mostres d'un valor de la distribució gamma.

Taula 4.5: Valors propis de la matriu de covariàncies i paràmetres per a la distribució aproximada de $T_{m,p}$, per a diferents valors de m i de l'índex del valor extrem, ξ. En tots els casos $p = 1/2$.

p = 1/2		**Valors propis**					**Paràmetres**		
$\boldsymbol{\xi}$	**Test**	$\boldsymbol{\lambda_0}$	$\boldsymbol{\lambda_1}$	$\boldsymbol{\lambda_2}$	$\boldsymbol{\lambda_3}$	$\boldsymbol{\lambda_4}$	**a**	**b**	$\boldsymbol{\nu}$
-0.1	**T0**	0.5356					0.0000	0.5356	1.0000
	T1	0.8081	0.2630				0.1157	0.7559	1.2639
	T2	0.2097	0.4055	0.9915			0.2555	0.8818	1.5323
	T3	0.1925	0.2824	0.5508	1.1165		0.3934	0.9531	1.8349
	T4	0.1844	0.2372	0.3671	0.6862	1.2029	0.5248	0.9952	2.1633
0	**T0**	1.0000					0.0000	1.0000	1.0000
	T1	1.7071	0.2929				0.2000	1.6667	1.0800
	T2	0.5000	0.2192	2.2808			0.4792	2.1818	1.1554
	T3	2.7503	0.3103	0.7420	0.1974		0.7971	2.5758	1.2435
	T4	1.0000	0.2500	3.1380	0.1878	0.4241	1.1323	2.8764	1.3446
0.1	**T0**	2.8929					0.0000	2.8929	1.0000
	T1	5.4393	0.3464				0.3036	5.4187	1.0117
	T2	0.2336	0.6755	7.7696			0.7824	7.7096	1.0242
	T3	0.2063	0.3488	1.1189	9.8974		1.4010	9.7710	1.0409
	T4	0.1947	0.2689	0.5036	1.6591	11.8380	2.1290	11.6134	1.0622

A la Taula 4.5 es mostren els valors propis i els paràmetres associats a diferents tests (m =0, 1, 2, 3 i 4) per a diferents valors del paràmetre ξ (-0.1, 0, 0.1) i un valor de $p = 1/2$. De la mateixa manera que es poden relacionar els valors propis de l'estadístic $cT_{m,p}$ amb els de $T_{m,p}$, també es poden relacionar els seus paràmetres associats a la distribució aproximada.

Siguin a, b i ν els paràmetres de l'estadístic $T_{m,p}$ i a^*, b^* i ν^*els de cT_m, on

$$a + b\nu = \sum_0^m \lambda_i,\ b^2\nu = \sum_0^m \lambda_i^2,\ b^3\nu = \sum_0^m \lambda_i^3 \tag{4.13}$$

$$a^* + b^*\nu^* = \sum_0^m c\lambda_i,\ b^2\nu = \sum_0^m c\lambda_i^2,\ b^3\nu = \sum_0^m c\lambda_i^3 \tag{4.14}$$

Resolent els dos sistemes es pot comprovar que

$$a^* = ca,\ b^* = cb,\ v^* = v \tag{4.15}$$

4.4 Simulacions

Un cop presentades les distribucions asimptòtica i aproximada, es pot estudiar si realment el que diu la teoria es compleix a la pràctica, és a dir, si les distribucions asimptòtica i aproximada ajusten bé la distribució real de l'estadístic $T_{m,p}$ per a mostres de mida gran, per veure si en alguns casos funcionen millor que en altres.

La Taula 4.6 mostra els punts crítics corresponents als percentils 90, 95 i 99, obtinguts per simulació, per als estadístics $T_m/((m+1)\sigma_\xi^2)$ amb m =20 i 40 i ξ = -0,5, -0,4, -0,2, -0,1, 0, 0,1, i 0,2, per a mostres de mida 500, 1000 i 2000, així com els valors obtinguts per simulació de les distribucions asimptòtica i aproximada. Les simulacions han estat executades 10.000 vegades en cada cas. Es pot veure que els mètodes asimptòtic i aproximat són útils per trobar p-valors aproximats quan $\xi < 0$, ja que en aquesta situació ajusten molt bé, fins i tot per a mostres de mida 500, però es necessita mostra de mida més gran quan el valor del paràmetre creix per obtenir un ajust millor, i per a valors de ξ més grans o iguals que 0, a mesura que s'apropa a la no existència dels quatre primers moments, $\xi \geq 1/4$, aquests dos mètodes deixen de ser útils.

Taula 4.6: Valors crítics per simulació directa de $T_{m,p}/((m+1)\sigma_\xi^2)$, per a mostres de mida 500,1.000 i 2.000, i de les distribucions asimpòtica i aproximada (amb 10.000 simulacions).

	$\mathbf{T_{20,1/2}/(21\sigma_\xi^2)}$			$\mathbf{T_{40,1/2}/(41\sigma_\xi^2)}$			$\mathbf{T_{20,1/2}/(21\sigma_\xi^2)}$			$\mathbf{T_{40,1/2}/(41\sigma_\xi^2)}$		
	90	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
$\boldsymbol{\xi}$	**-0.5**						**-0.1**					
500	1.65	2.03	2.9	1.63	1.97	2.85	1.14	1.48	2.62	1.16	1.5	2.69
1000	1.55	1.85	2.56	1.55	1.82	2.5	1.24	1.57	2.59	1.21	1.54	2.54
2000	1.52	1.76	2.36	1.52	1.77	2.36	1.26	1.55	2.59	1.25	1.58	2.57
Asimp.	1.5	1.71	2.17	1.47	1.68	2.13	1.74	2.08	2.95	1.73	2.08	2.9
Aprox.	1.54	1.75	2.18	1.5	1.7	2.13	1.79	2.15	2.95	1.75	2.12	3.03
$\boldsymbol{\xi}$	**-0.4**						**0**					
500	1.62	2	2.97	1.62	2	2.99	0.78	1.03	2.1	0.77	1.03	2.08
1000	1.56	1.84	2.62	1.52	1.84	2.56	0.87	1.15	2.17	0.86	1.13	2.08
2000	1.51	1.79	2.42	1.52	1.79	2.5	0.93	1.21	2.34	0.93	1.21	2.22
Asimp.	1.54	1.76	2.22	1.53	1.77	2.25	1.96	2.5	3.8	1.94	2.47	3.75
Aprox.	1.54	1.76	2.27	1.52	1.74	2.28	2	2.54	3.88	2	2.51	3.74
$\boldsymbol{\xi}$	**-0.3**						**0.1**					
500	1.55	1.94	3.02	1.58	1.96	2.98	0.36	0.47	1.03	0.36	0.47	0.98
1000	1.5	1.84	2.72	1.47	1.78	2.56	0.44	0.61	1.37	0.44	0.58	1.31
2000	1.49	1.78	2.47	1.47	1.77	2.52	0.51	0.68	1.73	0.5	0.69	1.54
Asimp.	1.58	1.81	2.32	1.57	1.82	2.37	2.28	3.09	5.15	2.28	3.08	5.14
Aprox.	1.57	1.83	2.32	1.57	1.82	2.33	2.17	2.92	4.82	2.26	3.04	4.9
$\boldsymbol{\xi}$	**-0.2**						**0.2**					
500	1.4	1.77	3.01	1.38	1.78	3.09	0.07	0.08	0.15	0.07	0.08	0.16
1000	1.41	1.74	2.69	1.4	1.69	2.62	0.09	0.12	0.26	0.09	0.11	0.23
2000	1.41	1.71	2.56	1.41	1.74	2.52	0.11	0.14	0.39	0.11	0.14	0.39
Asimp.	1.62	1.89	2.52	1.62	1.88	2.53	2.56	3.53	6.15	2.67	3.75	6.36
Aprox.	1.64	1.92	2.56	1.62	1.91	2.54	2.5	3.52	6.03	2.6	3.61	6.19

4.5 Selecció de llindars

La teoria asimptòtica diu que la cua d'una distribució es comporta com una GPD a partir d'un llindar prou gran, seria intuitiu que, en cas de rebutjar la hipòtesi nul·la fent servir un test T_m, es triés un llindar més gran i es tornés a realitzar un contrast. També seria raonable que, si aquest procés es realitza iterativament mentre hi hagi evidències de rebuig, arribarà un moment en el qual s'acceptarà que la mostra prové d'una GPD.

En aquest apartat, es presenta un algoritme que determina quin és el llindar a partir del qual la cua es pot considerar GPD utilitzant tests T_m. Es pot fer servir tant quan l'índex del valor extrem és conegut com desconegut i, en aquest segon cas, proporciona també una estimació per a ξ.

Un cop fixat el valor m es procedeix de la manera següent:

- Calcular el valor de p donat per (4.3).
- Es construeix la partició $t_0 \leq t_1 \leq \cdots \leq t_m$ de manera que cada t_k correspongui al quantil $(1 - p^k)$.
- Es realitza el contrast T_m per a tota la mostra. Si no es rebutja H_0, t_0 es retorna com a estimador del llindar a partir del qual hi ha GPD, en cas contrari es passa al pas següent.
- Es realitza el contrast T_{m-1} per a la mostra $\{x_j : \ x_j > t_1\}$. Es pot comprovar que la manera d'escollir la partició inicial fa que el valor de p no variï a mesura que es disminueix el valor de m i la mida de la mostra a partir dels llindars escollits. Si no es rebutja H_0, t_1 es retorna com a estimador del llindar a partir del qual hi ha GPD, en cas contrari es passa al pas següent.
- Es disminueix el valor de m i la mida de la partició i es realitza el contrast T_{m-q} per a la mostra $\{x_j : \ x_j > t_q\}$ tantes vegades com calgui fins que no es rebutja H_0.

A l'hora d'aplicar l'algoritme, s'ha d'escollir el valor de m però també el d'una significació, la qual perd una mica el seu significat tal i com s'aplica en un contrast ja que es poden arribar a a aplicar m contrastos en el pitjor dels casos i cada un d'aquests no són independents, de manera que si es desitja quantificar la significació global de l'algoritme no és un fet inmediat. Els valors proposats són 0,05, 0,1, 0,15 i 0,2.

Tot i que algun dels valors proposats per a la significació pot ser relativament gran, cal pensar que, en el tests T_m, el que es vol contrastar és H_0 i no H_1 com passa normalment. En el cas dels tests T_m fets servir iterativament, s'ha de fer servir una significació major per donar sentit a aquesta nova visió que es planteja sobre les hipòtesis si es vol ser més restrictiu a l'hora de trobar un llindar raonable. Un tipus de plantejament com aquest es pot trobar en àrees com la de la bioequivalència en medicaments, veure Ocaña *et al.* (2008). En aquestes situacions, tot i recomanar una significació d'un 5% per als contrasts, a més a més, es realitzen comprovacions addicionals per donar la conclusió final del contrast i veure si dos medicaments són o no equivalents.

Exemple 4.4. Per il·lustrar l'aplicació d'aquest algoritme, es considerarà la transformació de les dades daneses $Y = -1/X$ amb un llindar de dos milions de corones, en

el Capítol 6 es tornarà a parlar sobre aquestes dades i s'explicarà el motiu de l'elecció d'aquest llindar.

A la Figura 4.1 es poden veure els *CV-plots* corresponents a aquesta transformació amb els intervals de confiança del 95% de les distribucions exponencial i uniforme a l'esquerra i amb l'interval amb la mateixa confiança per a una GPD amb $\xi = -0{,}5$.

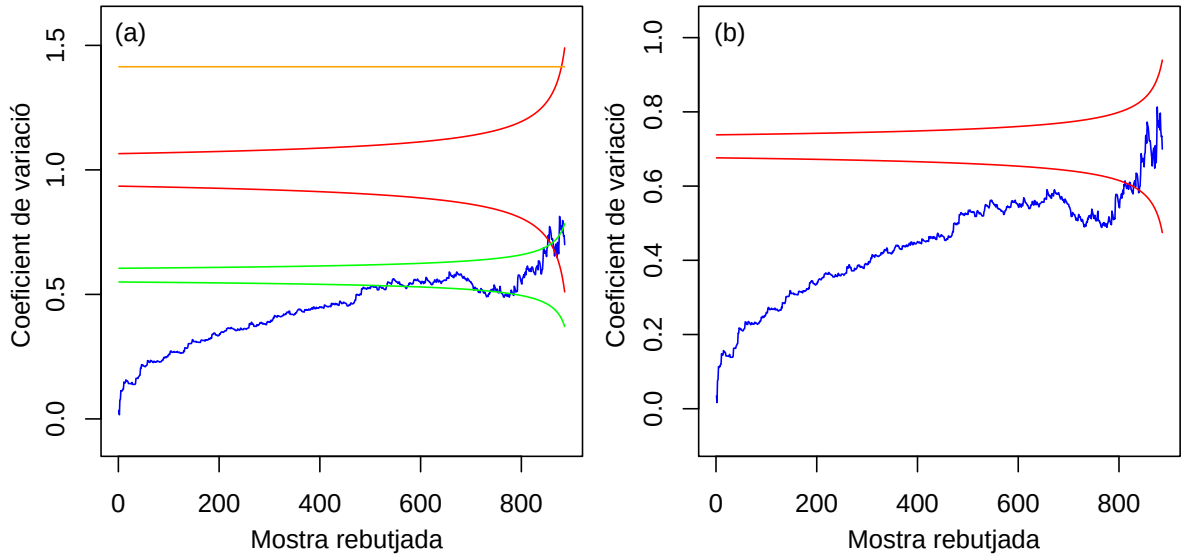

Figura 4.1: (a) *CV-plot* de les dades daneses per sobre de 2 milions de pèrdues després d'aplicar-li la transformació $Y = -1/X$. Les línies verdes i vermelles corresponen als intervals de confiança del 95% de les distribucions uniforme i exponencial, respectivament. (b) *CV-plot* per a la mateixa transformació amb l'interval de confiança per a una GPD amb $\xi = -0{,}5$.

Per a la mostra de mida 903, un cop transformada, es fa servir un valor de m igual a 20 i càlcul del p-valor mitjançant 10.000 simulacions, sobre la significació cal dir que el resultat és el mateix utilitzant qualsevol dels valors proposats. Després d'aplicar l'algoritme, s'obtenen els següents resultats d'estimacions per a Y

$$\hat{\xi} = -0,62$$
$$\text{Valor estimat del llindar } t = -0,125$$
$$\text{Mida de la cua a partir del llindar} = 110$$

Ara només cal desfer la transformació. Per a ξ, simplement cal canviar el signe de l'estimador i, per al llindar, cal desfer la transformació i sumar-li el valor mínim de la mostra original, ja que prèviament s'havien traslladat les dades a l'origen. La mida de la cua a partir del llindar es manté igual.

Les estimacions de ξ i del llindar t per a la variable X original són finalment

$$\hat{\xi} = 0,62$$
$$\text{Valor estimat del llindar } t = 9,997$$

Un valor de $\xi = 0{,}62$ continua sent proper a 0,5, però la novetat d'aquest algoritme és que també proporciona una estimació per al llindar t a partir del qual comença la cua de la distribució, no és necessari mirar cap gràfic, com es fa habitualment amb l'*ME-plot*, i desapareix, per tant, la subjectivitat de cada investigador a l'hora de seleccionar el llindar. A més a més, el llindar estimat és coherent amb els resultats que s'havien obtingut anteriorment, on s'escollia un llindar de 10 milions amb una cua de 109 dades i un valor estimat $\xi = 0{,}5$.

4.5.1 Variacions en els paràmetres de T_m

Un cop l'ús dels estadístics $T_{m,p}$ queda limitat a l'elecció del valor de m, perquè el valor de p queda fixat a partir d'aquest, sorgeix la qüestió de si hi ha algun valor de m més eficaç, també com afecta l'elecció d'una significació o una altra als resultats finals. En aquest apartat es realitzaran diferents combinacions de mida de mostra, valor de l'índex del valor extrem ξ, significació i valor de m per veure com es comporta el mètode de selecció de llindars en l'estimació del llindar i del paràmetre ξ.

Per realitzar aquesta tasca, s'han simulat mostres GPD per a cada combinació de les mides 50, 100 i 200, amb els valors -0,25, 0 i 0,1 per al paràmetre ξ, fent servir en tots els casos un valor de $\psi = 1$, ja que els mètodes T_m no depenen del paràmetre d'escala. En concret, s'han simulat 1.000 casos per a cada una de les 9 combinacions escollides i a cada un d'aquests casos se li ha aplicat el mètode T_m amb un valor de m de 5, 8, 10, 15, 20 i 30, cada un amb les significacions del 5, 10, 15 i 20%, és a dir, a cada mostra se li han aplicat 24 vegades el mètode T_m amb diferents paràmetres. El càlcul dels p-valors necessaris s'ha realitzat mitjançant 1.000 simulacions en cada cas.

Com que el mètode T_m dóna com a resultat tant l'estimació de l'índex del valor extrem ξ com del llindar a partir del qual es pot considerar que la cua és GPD, es poden estudiar aquests dos valors i la mesura de l'error que es farà servir és l'error quadràtic mitjà (MSE) de l'estimació del llindar i de l'estimació del paràmetre ξ. Com que totes les mostres simulades són GPD, es considerarà que el llindar correcte en tots els casos és el 0 i el valor de ξ correcte el que s'hagi fet servir per a fer la simulació, és a dir, -0,25, 0 i 0,1.

A la Taula 4.7 es poden veure els MSE per a les estimacions del llindar. Es pot observar com, fixada la mida de mostra i el valor del paràmetre ξ, els MSE són molt semblants, independentment del valor de m i de la significació escollida. Per una altra

banda, es pot veure com, fixat ξ, el MSE disminueix a mesura que augmenta la mida de la mostra quan aquest és 0,25 i 0, en canvi augmenta quan $\xi = 0,1$. Finalment, es pot observar com el MSE va augmentant a mesura que augmenta el valor de ξ, fet que té sentit si es considera que, com més a prop s'estigui de la falta de moments.

Taula 4.7: MSE per a l'estimació del llindar.

ξ		-0.25			0			0.1		
Mida		50	100	200	50	100	200	50	100	200
m	sig	MSE llindar			MSE llindar			MSE llindar		
5	0.05	0.023	0.000	0.000	1.006	1.597	0.110	1.959	3.944	7.162
5	0.10	0.023	0.000	0.000	1.006	1.597	0.194	1.959	3.944	7.162
5	0.15	0.023	0.000	0.000	1.006	1.597	0.321	1.959	3.944	7.162
5	0.20	0.023	0.000	0.000	1.006	1.597	0.447	1.959	3.944	7.162
8	0.05	0.029	0.005	0.000	0.855	1.517	0.098	1.642	3.680	6.892
8	0.10	0.029	0.005	0.000	0.855	1.517	0.207	1.642	3.680	6.892
8	0.15	0.029	0.005	0.000	0.855	1.517	0.301	1.642	3.680	6.892
8	0.20	0.029	0.005	0.000	0.855	1.517	0.413	1.642	3.680	6.892
10	0.05	0.029	0.004	0.000	0.926	1.531	0.103	1.823	3.618	7.063
10	0.10	0.029	0.004	0.000	0.926	1.531	0.178	1.823	3.618	7.063
10	0.15	0.029	0.004	0.000	0.926	1.531	0.296	1.823	3.618	7.063
10	0.20	0.029	0.004	0.000	0.926	1.531	0.400	1.823	3.618	7.063
15	0.05	0.029	0.004	0.000	0.904	1.409	0.069	1.812	3.276	6.255
15	0.10	0.029	0.004	0.000	0.904	1.409	0.147	1.812	3.276	6.255
15	0.15	0.029	0.004	0.000	0.904	1.409	0.228	1.812	3.276	6.255
15	0.20	0.029	0.004	0.000	0.904	1.409	0.316	1.812	3.276	6.255
20	0.05	0.028	0.005	0.000	0.856	1.473	0.079	1.714	3.495	6.412
20	0.10	0.028	0.005	0.000	0.856	1.473	0.141	1.714	3.495	6.412
20	0.15	0.028	0.005	0.000	0.856	1.473	0.232	1.714	3.495	6.412
20	0.20	0.028	0.005	0.000	0.856	1.473	0.346	1.714	3.495	6.412
30	0.05	0.028	0.004	0.000	0.827	1.452	0.075	1.641	3.405	6.069
30	0.10	0.028	0.004	0.000	0.827	1.452	0.115	1.641	3.405	6.069
30	0.15	0.028	0.004	0.000	0.827	1.452	0.194	1.641	3.405	6.069
30	0.20	0.028	0.004	0.000	0.827	1.452	0.301	1.641	3.405	6.069

A la Taula 4.8 es poden veure els MSE per a les estimacions del paràmetre ξ. Es pot observar com, fixada la mida de mostra i el valor del paràmetre ξ, els MSE són molt semblants, independentment del valor de m i de la significació escollida. Per una altra banda, es pot veure com, fixat ξ, el MSE disminueix a mesura que augmenta la mida de la mostra en tots els casos. Finalment, es pot observar com el MSE va augmentant

a mesura que augmenta el valor de ξ, fet que té sentit si es considera que, com més a prop s'estigui de la falta de moments, tot i que ho fa amb menys diferència que en el cas del MSE per als llindars.

Taula 4.8: MSE per a l'estimació de ξ.

ξ		**-0.25**			**0**			**0.1**		
Mida		**50**	**100**	**200**	**50**	**100**	**200**	**50**	**100**	**200**
m	**sig**	**MSE ξ**			**MSE ξ**			**MSE ξ**		
5	0.05	0.017	0.008	0.004	0.027	0.019	0.008	0.030	0.023	0.020
5	0.10	0.017	0.008	0.004	0.027	0.019	0.009	0.030	0.023	0.020
5	0.15	0.017	0.008	0.004	0.027	0.019	0.010	0.030	0.023	0.020
5	0.20	0.017	0.008	0.004	0.027	0.019	0.012	0.030	0.023	0.020
8	0.05	0.019	0.008	0.004	0.028	0.018	0.007	0.031	0.022	0.020
8	0.10	0.019	0.008	0.004	0.028	0.018	0.008	0.031	0.022	0.020
8	0.15	0.019	0.008	0.004	0.028	0.018	0.009	0.031	0.022	0.020
8	0.20	0.019	0.008	0.004	0.028	0.018	0.010	0.031	0.022	0.020
10	0.05	0.019	0.008	0.004	0.027	0.02	0.008	0.030	0.023	0.018
10	0.10	0.019	0.008	0.004	0.027	0.02	0.009	0.030	0.023	0.018
10	0.15	0.019	0.008	0.004	0.027	0.02	0.010	0.030	0.023	0.018
10	0.20	0.019	0.008	0.004	0.027	0.02	0.012	0.030	0.023	0.018
15	0.05	0.020	0.009	0.004	0.030	0.019	0.008	0.033	0.022	0.018
15	0.10	0.020	0.009	0.004	0.030	0.019	0.009	0.033	0.022	0.018
15	0.15	0.020	0.009	0.004	0.030	0.019	0.010	0.033	0.022	0.018
15	0.20	0.020	0.009	0.004	0.030	0.019	0.012	0.033	0.022	0.018
20	0.05	0.021	0.009	0.004	0.031	0.019	0.008	0.034	0.023	0.017
20	0.10	0.021	0.009	0.004	0.031	0.019	0.008	0.034	0.023	0.017
20	0.15	0.021	0.009	0.004	0.031	0.019	0.010	0.034	0.023	0.017
20	0.20	0.021	0.009	0.004	0.031	0.019	0.011	0.034	0.023	0.017
30	0.05	0.021	0.009	0.004	0.029	0.021	0.008	0.032	0.024	0.017
30	0.10	0.021	0.009	0.004	0.029	0.021	0.009	0.032	0.024	0.017
30	0.15	0.021	0.009	0.004	0.029	0.021	0.010	0.032	0.024	0.017
30	0.20	0.021	0.009	0.004	0.029	0.021	0.011	0.032	0.024	0.017

CAPÍTOL 5

Aplicacions en ciències de la computació: valors extrems en sistemes informàtics encastats

Les funcionalitats de nombres en els dominis de l'automoció, l'espai, la indústria aeroespacial i de ferrocarril que impliquen un sistema informàtic està augmentant ràpidament. Per exemple, mentre que en el passat el volant d'un cotxe tenia una connexió mecànica a les rodes, ara aquesta connexió és electromecànica; un sensor detecta l'angle de direcció i el transmet a un ordinador, el qual el processa i acciona uns motors elèctrics sobre les rodes. Tot aquest procés està subjecte a estrictes limitacions de temps, ja que afecta a la seguretat de tot el cotxe.

Per als sistemes informàtics implicats en funcions crítiques relacionades amb la seguretat, cal demostrar que la computació requerida per processar noves entrades s'executa correctament i a temps. En el cas particular de la sincronització, el pitjor temps d'execució (WCET) dels programes que duen a terme els càlculs necessita ser estimat. Existeixen mètodes per estimar el WCET dels programes, cadascun amb el seu propi conjunt de restriccions, Wilhelm *et al.* (2008). Recentment, dins de l'anàlisi probabilístic de temps (PTA) s'ha generat una variant anomenada *Measurement-Based Probabilistic Timing Analysis* (MBPTA), la qual ha sorgit com un poderós mètode per obtenir estimacions del WCET en sistemes de computació internament complexos, Cucu-Grosjean *et al.* (2012). L'MBPTA es basa en plataformes; el temps d'execució de les quals pot ser modelitzat amb una variable aleatòria mitjançant una recopilació de mesures de temps d'execució convenient, Cazorla *et al.* (2013), una correcta aplicació de l'EVT permet estimar la distribució dels temps d'execució més elevats. Aquesta distribució pot ser llavors utilitzada per obtenir el temps d'execució que podria ser excedit amb

una probabilitat donada (arbitràriament petita), la qual és anomenada probabilitat del WCET. La investigació actual sobre MBPTA s'ha centrat en una aplicació particular de l'EVT. No obstant això, una sèrie de supòsits utilitzats en aquestes aplicacions no s'han estudiat. En aquest treball s'analitzen alguns d'aquests supòsits i les seves implicacions fent servir alguns dels últims descobriments en el camp de l'EVT, Castillo *et al.* (2014a) i Castillo i Serra (2015). En particular, l'estudi es basa en l'anàlisi d'aplicacions d'automoció mitjançant l'ús del conjunt de referència EEMBC Autobench, Poovey (2007). Aquest conjunt EEMBC reflecteix l'actual demanda del món real d'alguns sistemes encastats d'automoció en temps real crucials. Es considera una arquitectura de maquinari que implementa algunes característiques d'alt rendiment, com ara multicores i jerarquies de memória cau anàlogues a les que hi ha en Slijepcevic *et al.* (2014), que ha estat modelitzada per mitjà d'un simulador de cicle de precisió basat en SoCLib, veure SoClib.

L'article s'organitza de la manera següent; a la Secció 2 es realitza un anàlisi exploratori de les dades, a la Secció 3 s'estudia el supòsit d'independència i s'escull l'enfoc adient per a l'estudi adient dels conjunts, a la Secció 4 es presenten els estadístics que es faran servir per contrastar la hipòtesi bàsica d'exponencialitat i la hipòtesi de GPD, a la Secció 5 es procedeix a estimar l'índex del valor extrem, a la Secció 6 es fan servir els resultats obtinguts per aplicar el mètode *Peak over Threshold* (PoT) i realitzar el càlcul del *Value-at-risk* (VaR) i, finalment, a la Secció 7 s'exposen les principals conclusions de l'estudi.

5.1 Anàlisi exploratori de les dades

Per realitzar l'anàlisi es disposa de mostres de mida 1.000 per a cada un dels 16 conjunts que formen part de l'EEMBC AutoBench, Poovey (2007), el qual està format per una coneguda sèrie de programes de temps reals que són utilitzats en alguns sistemes encastats d'automoció.

Aquests conjunts de dades, els quals representen temps d'execució en cicles, són, en ordre alfabètic: `a2time`, `aifftr`, `aifirf` , `aiifft`, `basefp`, `bitmnp`, `cacheb`, `canrdr`, `idctrn`, `iirflt`, `matrix`, `pntrch`, `puwmod`, `rspeed`, `tblook` i `ttsprk`. A partir d'ara, aquests conjunts s'anomenaran amb una A (d'automoció) i el nombre que ocupin en la llista anterior, per exemple, per fer referència a `basefp` es farà servir la notació A5.

En primer lloc s'han realitzat els histogrames de les dades, en tots els casos s'observen distribucions unimodals amb valors extrems a la cua dreta. A la Figura 5.1 es poden veure els histogrames dels temps d'execució per als conjunts A2 i A11, s'observa que per al conjunt A2 apareixen alguns valors extrems allunyats de la mediana, en canvi, per al conjunt A11, es pot veure com apareix una petita nova agrupació de valors a la dreta

molt allunyada.

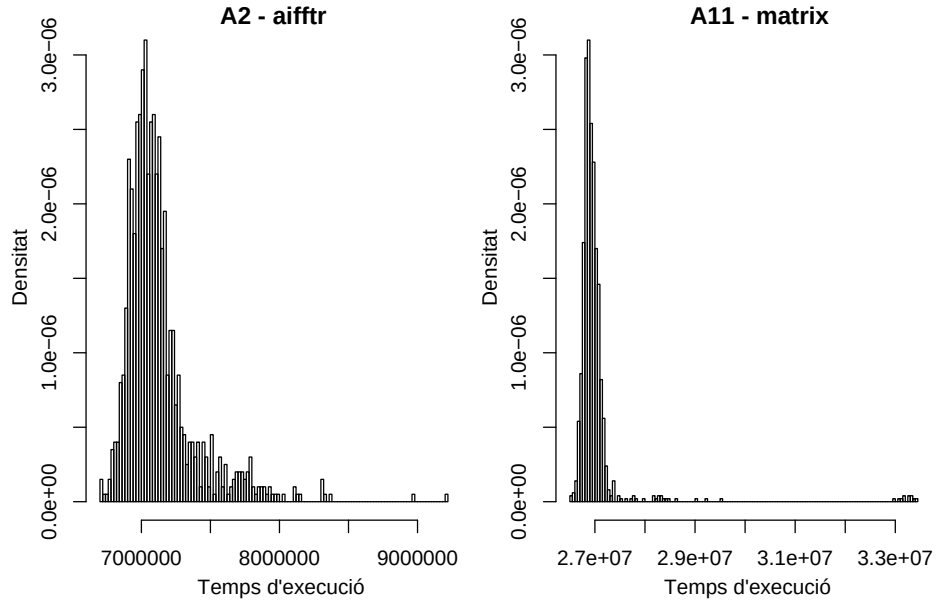

Figura 5.1: Histogrames dels temps d'execució dels conjunts A2 i A11.

Donada la presència de valors extrems, s'han realitzat els *box-plots* per observar-los visualment. A la Figura 5.2 es mostren els *box-plots* de sis conjunts representatius dels comportaments diferenciats que presenten les cues de tots els casos estudiats, incloent els conjunts A2 i A11 mostrats anteriorment. En el cas A5 les dades són bastant homogènies i s'observen els valors atípics molt agrupats, en els casos A2 i A9 els valors atípics es troben una mica més dispersos que en el cas anterior, però encara bastant agrupats amb només un parell de valors molt extrems, en A8 els valors atípics es van dispersant d'una manera més homogènia i, finalment, en els casos A11 i A12 s'observen clarament dues agrupacions de valors extrems. Aquesta diferència en el tipus de comportament de la cua pot venir donada per la naturalesa del programa i el muntatge de la màquina.

Si es consideren el rang interquartílic, $IQR = (Q_3 - Q_1)$ el qual es correspon amb l'amplada de la caixa, W_1 la dada més petita encara dins de $Q_1 - 1.5IQR$ i W_2 la dada més gran encara dins de $Q_3 + 1.5IQR$, es pot observar que la majoria de valors que queden fora de l'interval (W_1, W_2) són majors que W_2, amb proporcions que varien entre un 0,8% i un 8,7% sobre la mostra total. Cal fer notar que, sota normalitat, la proporció de valors que queda dins d'aquest interval és un 99%. Un primer indicador numèric que es pot fer servir per a la classificació de les cues és la distància entre el valor màxim i W_2 en termes de l'IQR. A la Taula 5.1, a la columna "Dist.", es pot veure com aquests valors varien entre 0,5 i 32,2 en termes de l'IQR, una distància petita es correspondria amb cua lleugera, una distància mitjana amb cua exponencial i per a valors elevats es podria tenir cua pesada.

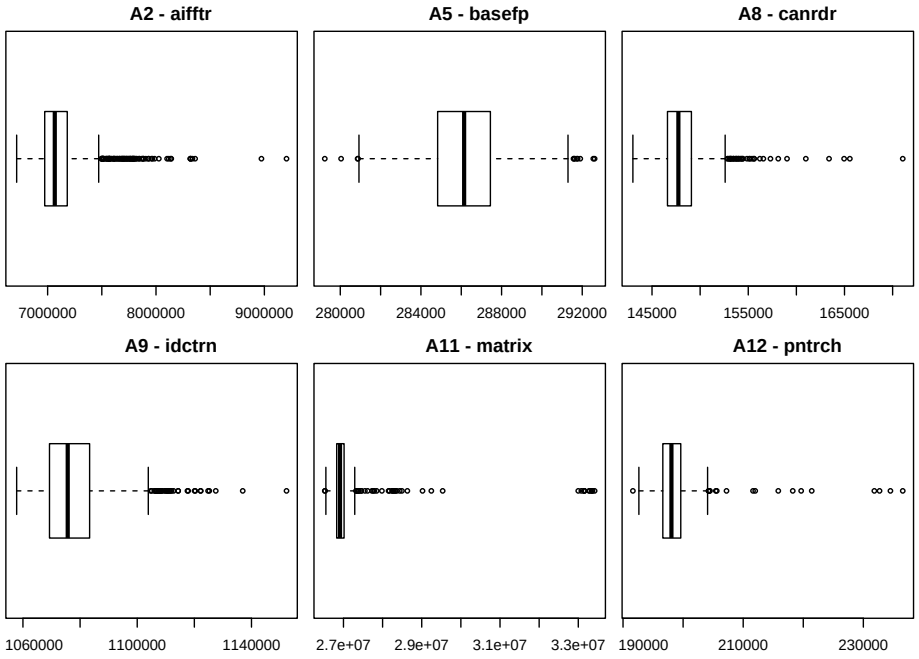

Figura 5.2: *Box-plots* dels temps d'execució de sis conjunts representatius.

Tant amb la visualització dels *box-plots* com amb la dels histogrames, es pot veure com els valors de la cua es troben per sobre de la mediana dels conjunts en tots els casos, per tant es recomana realitzar els estudis amb les dades a partir d'aquesta, ja que valors anteriors no aportaran cap informació doncs els models que s'utilitzaran per a les cues són distribucions decreixents.

5.2 Test d'independència

Els dos enfocs presentats en el Capítol 1 per a l'EVT, el model de blocs de màxims (GEV) i el model d'excessos per sobre d'un llindar (GPD), es basen en la hipòtesi d'independéncia de les dades, encara que pot ser aplicada amb un petit canvi en el cas de processos estacionaris (Coles, 2001, Ch 5). Per aplicar GEV, només es necessita independència dels màxims, la qual cosa pot ocòrrer en casos estacionaris on les observacions poden ser dependents en funció del temps, però el malgastament de dades fent servir aquest enfonc fa que sigui recomanable utilitzar els excessos sobre un llindar en el cas que la independència pugui ser assumida.

Un simple i popular test d'independència fa servir el primer coeficient d'autocorrelació, $\hat{\rho}_1$, rebutjant amb un nivell d'un 5% de significació si $\sqrt{n}\,|\hat{\rho}_1| > 1,96$. El test aplica el resultat $\sqrt{n}\varrho_k \sim N(0,1)$ per al coeficient d'autocorrelació en el *lag* k, aproximadament, quan H_0 és veritat. Quan es consideren molts coeficients d'altres ordres, alguns

poden ser significatius inclús si H_0 és certa. Una manera més natural de combinar h coeficients en un únic estadístic és amb el test de Ljung-Box, donat per

$$LB = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{5.1}$$

on n és la mida de la mostra i $\hat{\rho}_k$ és l'autocorrelació de la mostra en el *lag* k. Sota la hipòtesi d'independència, i la suposició que les dades tenen com a mínim moments finits d'ordre quatre, l'estadístic segueix una distribució χ_h^2. L'estadístic LB es fa servir en investigació de sèries temporals per contrastar la no autocorrelació de les dades. Altres tests, com el de canvi de signe, essencialment només consideren l'autocorrelació d'ordre 1, el desavantatge dels quals és una baixa potència causada per la pèrdua d'informació.

Aquesta metodologia permetrà decidir entre els dos enfocs proposats, GEV o GPD. Si les dades són independents, s'escollirà el model d'excessos sobre un llindar per no perdre tantes, en canvi, si les dades són dependents però els màxims són independents, s'escollirà el model per blocs de màxims. A més a més, permetrà comprovar que es compleix una de les hipòtesis del model, fet que no sempre és considerat pels investigadors, tal com queda patent a Resnick (1997), quan diu "*It is customary in many insurance studies involving heavy tailed phenomena to assume independence without actually statistically checking this important fact*".

Un factor a tenir en compte a l'hora d'aplicar el test és la necessitat d'existència de moments, fet que no sempre es satisfà quan es tenen cues pesades, "*It is not difficult to find examples of heavy tailed data which require infinite mean models for adequate fits*", veure Resnick (1997). Existeixen tècniques que poden ajudar a solucionar aquest problema, donada una variable aleatòria X amb cua pesada, amb la transformació $log(X)$ s'obtenen cues exponencials i amb la transformació $-1/X$ cues lleugeres, veure Castillo i Padilla (2015).

Es realitzarà el test de Ljung-Box per a X, $log(X)$ i $-1/X$, ja que sota un canvi d'escala la independència es conserva. A la Taula 5.1, a la columna "p-valor", es troben els p-valors més significatius per a cada conjunt en aplicar el test amb 20 *lags*, cal remarcar que les tres funcions donen p-valors molt semblants. Es pot veure que cap d'aquests casos és significatiu, per tant no hi ha evidències per rebutjar la independència de les dades i l'enfoc d'excessos sobre un llindar és l'escollit per a l'estudi de les dades.

Cal recordar que el valor del paràmetre ξ de la GPD que s'esculli per modelitzar els excessos per sobre d'un llindar determinarà el tipus de cua que s'obtindrà. Per a $\xi < 0$ la cua és lleugera, per a $\xi = 0$ la cua és exponencial, per a $0 < \xi < 1/4$ la cua és pesada, amb existència de moments fins a ordre quatre, i per a $\xi > 1/4$ considerarem la cua molt pesada.

5.3 Identificant el comportament de la cua

Per realitzar l'estudi del comportament de la cua es començarà realitzant els *CV-plots*. A la Figura 5.3 es poden veure els *CV-plots* corresponents a alguns conjunts representatius, les línies vermelles i verdes corresponen als intervals de confiança del 95% per a les distribucions exponencial i uniforme, respectivament, i la línia taronja constant correspon aproximadament a 1.4, el CV residual per a $\xi = 1/4$, el límit per al qual hi ha existència de moments de quart ordre. Aquestes gràfiques han estat realitzades a partir de la mediana de cada conjunt com a primer llindar, per tant, sobre conjunts de mida 500. Es pot observar com es traslladen al *CV-plot* els comportaments detectats amb els *box-plots*. Per a A2 i A9, s'observen cues exponencials, per a A5 una cua lleugera, per a A8 una cua pesada, però amb existència de moments d'ordre 4 i, finalment, per a A11 i A12, cues molt pesades que acaben amb un descens pronunciat cap a l'interval de confiança de la distribució exponencial. El comportament que s'observa per a A11 i A12 es té també per a A3 i A7, els quals són sospitosos de tenir falta d'existència de moments. Al final del capítol, en la Figura 5.6 es poden veure els *CV-plots* dels 16 conjunts de dades estudiats.

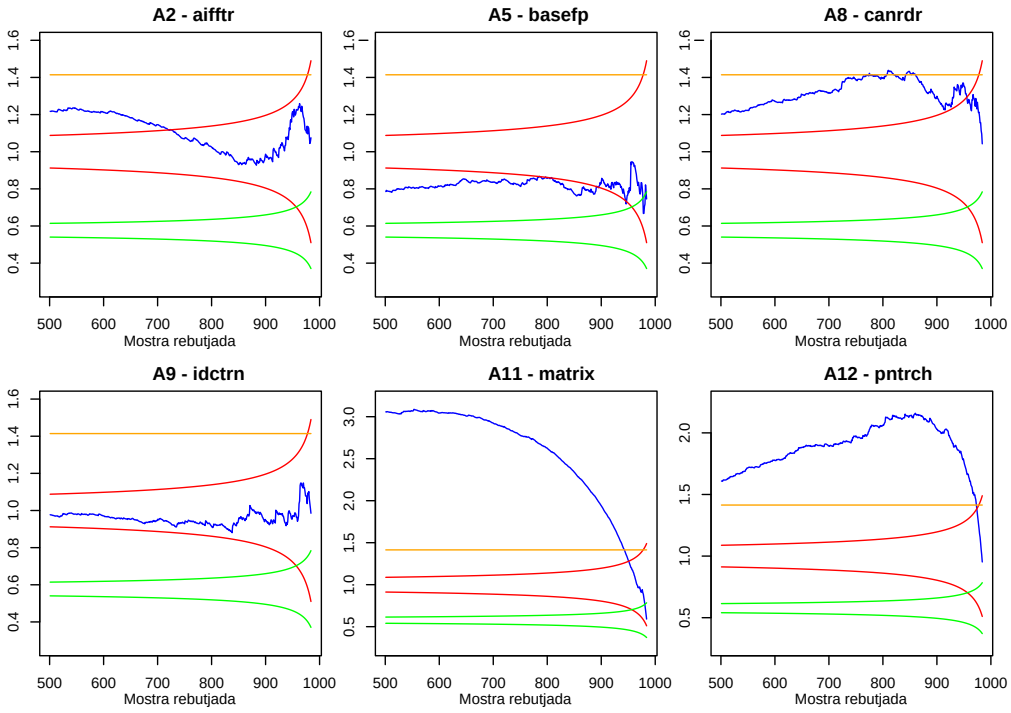

Figura 5.3: *CV-plots* dels temps d'execució de sis conjunts representatius. Les línies vermelles i verdes representen els intervals de confiança del 95%, respectivament, i la línia taronja representa la frontera a partir de la qual no hi ha existència de moments d'ordre quatre.

Els sistemes en temps real només utilitzen una forma restringida de programació, que garanteix que els programes sempre han d'acabar; la recursivitat no es permesa o es limita explícitament com han de ser els comptes d'iteració dels bucles. Considerant aquesta restricció, els models adequats seran els de suport compacte, índex del valor extrem negatiu per a la GPD, o la distribució exponencial, ja que és recomanable afegir els casos frontera en aquestes situacions, veure Castillo i Serra (2015). Si només es considerés el model bàsic, aquesta posició podria ser bastant conservadora en alguns casos amb estimacions massa restrictives quan realment no és necessari.

Per poder abordar aquesta situació, es faran servir els estadístics de contrast donats per (4.2). Aplicant el contrast sota exponencialitat, $\xi = 0$, amb $m = 20$ a les darreres 500 dades de cada conjunt, es rebutja exponencialitat en vuit dels casos: A1, A2, A3, A5, A7, A8, A11 i A12. Aplicant aquest contrast amb $m = 20$ sota la hipòtesi de GPD sense conèixer el valor de ξ, no es podria rebutjar GPD amb cua lleugera per a A1, A5, A6, A10, A13 i A15. Per al casos A6, A10, A13 i A15, en els quals no es rebutjava exponencialitat, ens trobaríem en la situació que seria massa restrictiu considerar-la, i cal considerar la cua lleugera. Per a A4, A9, A14 i A16 no es rebujaria GPD amb valors de ξ molt propers a 0, per tant es considera adient mantenir l'exponencialitat. Finalment, per als casos A2 i A8, no es rebutja GPD amb cua pesada, situació que contradiu la restricció de temps finit per als programes, i per als casos A3, A7, A11 i A12, es rebutja també GPD, de manera que considerar com a llindar la mediana de les dades no és adient.

A la Taula 5.1, es poden veure els casos per als quals s'escull exponencialitat o cua lleugera amb 500 dades, veure les columnes "Tipus" i "Mida de la cua".

5.4 Estimació de l'índex del valor extrem

Per estimar l'índex del valor extrem dels sis conjunts que no han quedat definits amb els contrastos per a GPD i exponencialitat, es farà servir l'algoritme de selecció de llindars presentat a la Secció 4.5, sota la hipòtesi d'exponencialitat i sota GPD sense conèixer l'índex del valor extrem de la cua. Aquest algoritme, a més d'estimar el llindar a partir del qual cal considerar que els excessos es comporten com una GPD, serveix també per estimar l'índex del valor extrem, és a dir, el valor de ξ.

A la Taula 5.1, a les columnes"ξ"i"Mida de la cua", es poden veure l'estimació per a l'índex del valor extrem de la GPD i el número de valors que quedarien per sobre del llindar escollit, per a sis dels casos es considera adient una cua lleugera (L), en tots ells a partir de la mediana, i per a la resta exponencialitat (E), en cinc casos a partir de la mediana o un valor proper i la resta deixant una cua petita però amb prou valors per a explicar els extrems. Aquests darrers casos coincideixen amb els conjunts sospitosos de tenir falta d'existència de moments, A3, A7, A11 i A12, i un que es troba en el límit,

Taula 5.1: Taula resum dels resultats per als 16 conjunts de dades. Es mostra, per a cada variable, el nom, el número d'extrems (valors majors que W_2), la distància en termes d l'IQR, p-valor més significatiu de realitzar el test de Ljung-Box amb 20 *lags* per a X, $log(X)$ i $-1/X$, valor de ξ estimat proposat, mida de la cua a partir del llindar escollit, tipus de cua, L per a lleugera i E per a exponencial, i VaR al 99,9% aplicant el mètode PoT.

Var.	**Nom**	**Extr.**	**Dist.**	**p-valor**	$\boldsymbol{\xi}$	**Mida cua**	**Tipus**	$\mathbf{VaR_{99.9\%}}$
A1	a2time	26	1.3	0.311	-0.2	500	L	124520
A2	aifftr	87	8.2	0.814	0.06	328	E	8499954
A3	aifirf	46	17.8	0.828	0.07	61	E	221294
A4	aiifft	60	5.0	0.506	0.04	500	E	7804243
A5	basefp	8	0.5	0.707	-0.25	500	L	291706
A6	bitmnp	38	1.7	0.399	-0.13	500	L	702430
A7	cacheb	47	18.5	0.156	-0.85	50	E	3068246
A8	canrdr	35	7.3	0.142	0.16	75	E	163376
A9	idctrn	39	3.4	0.459	-0.04	500	E	1144535
A10	iirflt	21	1.1	0.938	-0.15	500	L	179419
A11	matrix	43	32.2	0.742	-0.04	50	E	34231921
A12	pntrch	18	10.8	0.851	0.04	33	E	226807
A13	puwmod	16	1.1	0.412	-0.17	500	L	228167
A14	rspeed	31	2.8	0.713	0.04	500	E	72632
A15	tblook	34	1.5	0.741	-0.11	500	L	88498
A16	ttsprk	18	2.3	0.959	0.03	500	E	192864

A8.

A continuació, s'aplicarà la transformació, $Y = -1/(X + c) + 1/c$ on $c = \psi/\xi$, presentada en el Capítol 3, per solucionar el problema. Aplicant aquesta transformació als cinc casos A3, A7, A8, A11 i A12, no es pot rebutjar exponencialitat per a les 266, 14, 500, 12 i 329 darreres dades respectivament.

A la Figura 5.4 es poden veure els *CV-plots* per als casos A11 i A12, els quals presenten no existència de moments de tots els ordre necessaris, i per al cas A8, que pot semblar una mica dubtós fins que no s'han eliminat 800 dades de la mostra original. Es pot comparar el canvi dels *CV-plots* per a aquests casos observant els de la Figura 5.3, que corresponen a les dades originals; es pot veure com es passa de gràfiques amb valors elevats a gràfiques que queden entre els intervals de confiança de les distribucions uniforme i exponencial.

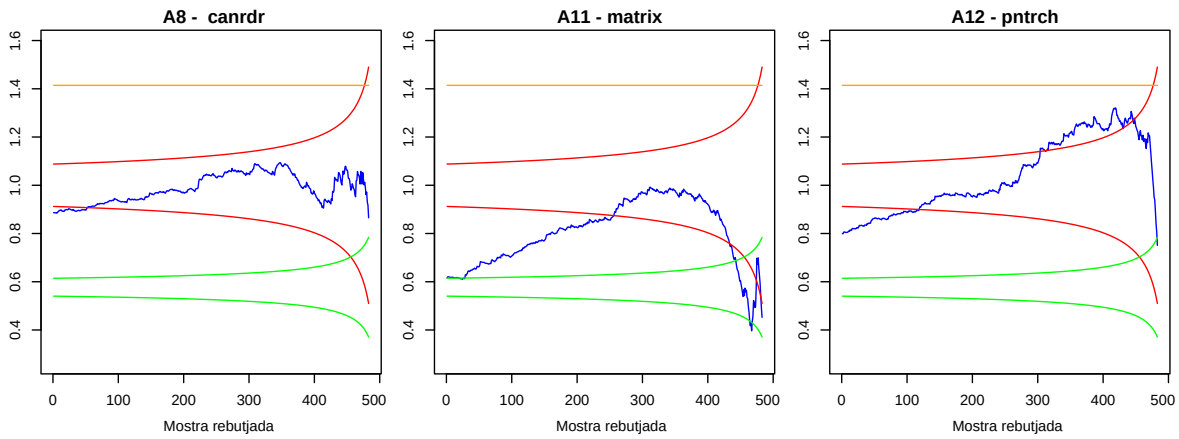

Figura 5.4: *CV-plots* per a les transformacions $Y = -1/(X + c) + 1/c$, $c = \psi/\xi$, sobre el temps d'execució de tres conjunts representatius. Les línies vermelles i verdes representen els intervals de confiança del 95%, respectivament, i la línia taronja representa la frontera a partir de la qual no hi ha existència de moments d'ordre quatre.

Aplicant l'algoritme de selecció de llindars sota hipòtesi de GPD sense conèixer ξ a A3, A7, A8, A11 i A12, els índexs del valor extrem estimats són 0, 0,84, 0,02, 0,95 i -0,13, respectivament. De manera que l'exponencialitat continuaria sent l'estimació adient.

Una alternativa per a l'estimació de l'índex del valor extrem és el mètode Hill, el qual sempre donarà estimacions positives per a l'índex de la cua, ja que es basa en el supòsit de tenir una cua pesada. Com que aquest estimador presenta alguns problemes de biaix, diverses alternatives han sorgit per solucionar aquest problema, entre elles un nou estimador basat en paràmetres de segon ordre que es presenta a Gomes i Pestana (2007). En l'àmbit de les dades financeres, s'han comparat estimacions realitzades mitjançant

el CV residual i el mètode Hill no esbiaixat, veure Castillo *et al.* (2014b).

Si s'aplica aquest mètode a A3, A7, A8, A11 i A12, s'obtenen unes estimacions de 0,61, 0,83, 0,45, 0,88 i 0,40 per a l'índex del valor extrem, amb mides de la cua de 390, 391, 392, 187 i 392, respectivament. Els resultats obtinguts no són consistents amb els que s'obtenen mitjançant la transformació i s'observa discrepància en l'elecció de llindars. Potser aquests mètodes no són els més adients per a aquests tipus de casos i cal pensar un altre, una mixtura de distribucions sembla un model que es podria aplicar com alternativa a l'exponencialitat que s'ha decidit en tots els casos conflictius.

5.5 PoT i VaR

Hi ha dos punts importants a l'hora de considerar el PoT: la selecció del llindar i l'estimació dels paràmetres, veure Coles (2001). El llindar s'escull normalment fent servir un mètode gràfic, en aquest cas es proposa l'ús del *CV-plot*, però gràcies a l'algoritme de selecció de llindars presentat a la Secció 4.5, tant l'estimació del llindar com la del paràmetre ξ es trobaran a partir d'aquest i el mètode gràfic servirà només de suport.

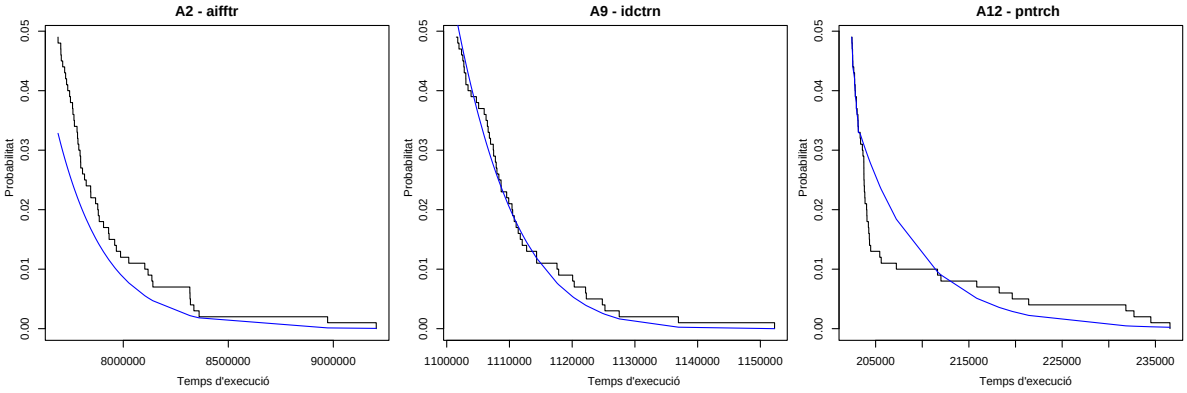

Figura 5.5: *Tail function* empírica a partir del quantil 95 per a A2, A9 i A12. La línia blava correspon a la mateixa funció realitzada a partir dels paràmetres estimats.

Finalment, s'han calculat alguns quantils elevats per a cada conjunt. En termes de risc, aquest concepte s'expressa com a *value at risk (VaR)*. Per a un valor petit p, $VaR_p = v$ si i només si $1 - F(v) = p$. La Taula 5.1 mostra els diferents valors per a cada llindar i índex del valor extrem proposats del $VaR_{99\%}$. A la Figura 5.5 es pot veure l'ajust per a tres conjunts a partir del quantil 95% després de l'estimació del paràmetre ψ per MLE. Sembla que els ajustos són bons per utilitzar-los per fer prediccions.

5.6 Conclusions

S'ha aplicat una nova metodologia en EVT a 16 conjunts de dades sobre automoció descrites en aquest capítol, per predir el temps màxim d'execució dels programes, amb una certa seguretat. La metodologia es basa en els punts següents:

1. Decidir entre GEV i GPD en base al test d'independència de Ljung-Box realitzat entre escales diferents que eviten el possible problema de no existència de moments.

2. Decidir entre el tipus de cua (lleugera, exponencial, pesada i molt pesada) en base als gràfics anomenats *CV-plots* i al test d'exponencialitat $T_{m,p}$, per a $\xi = 0$, que utilitza múltiples llindars.

3. Estimació de l'índex del valor extrem i del llindar a partir del qual es realitza l'estimació amb un nou algoritme de selecció de llindars basat en contrastos de GPD per a llindars múltiples. En el cas de cues molt pesades, l'algoritme s'aplica a una transformació de les dades que garanteix l'existència de moments.

4. Aplicació del mètode PoT que divideix la mostra entre el cos i la cua, a partir del llindar seleccionat anteriorment i amb l'estimació feta de l'índex del valor extrem.

La metodologia ha funcionat correctament en 14 dels 16 casos i presenta dubtes per als conjunts A11 i A12. En aquests casos s'han fet servir també altres metodologies que no han resultat concloents. La decisió que es proposa és acceptar exponencialitat per a un llindar molt elevat que ens fa sospitar el problema de contaminació per una distribució normal en una proporció molt baixa i per un llindar molt i molt elevat.

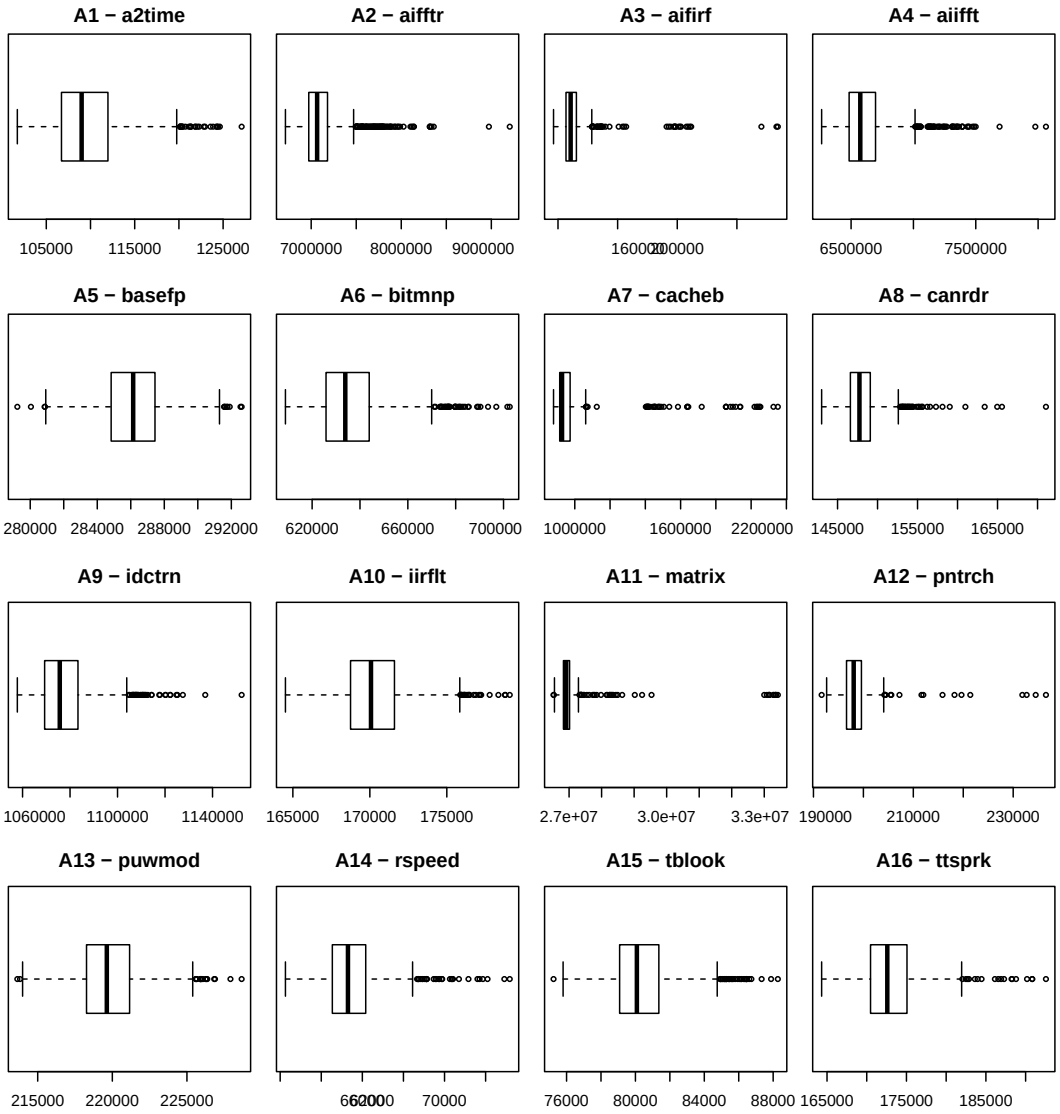

Figura 5.6: *Box-plots* dels temps d'execució per als 16 conjunts de dades.

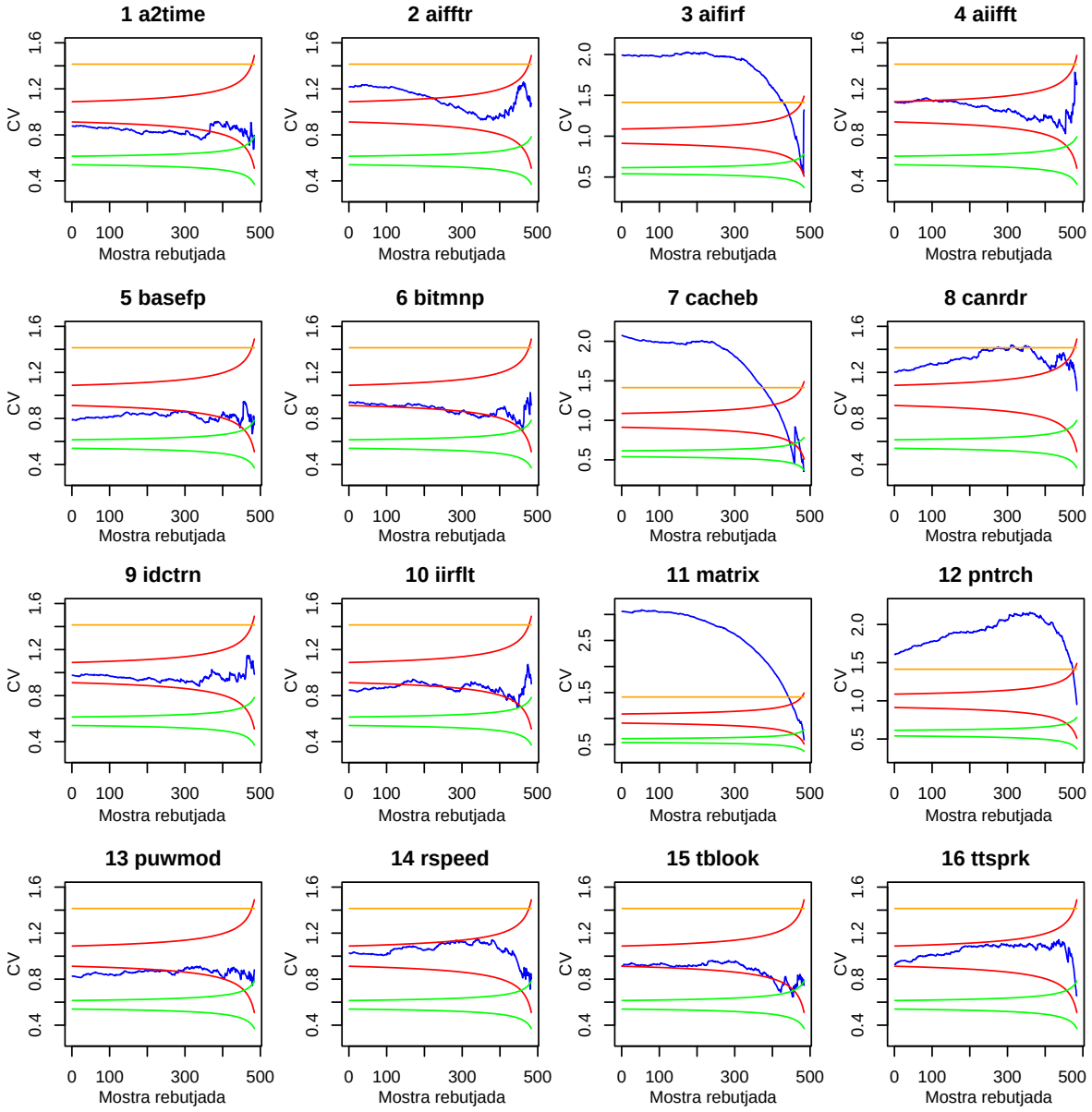

Figura 5.7: *CV-plots* dels temps d'execució per als 16 conjunts de dades. Les línies vermelles i verdes representen els intervals de confiança del 95%, respectivament, i la línia taronja representa la frontera a partir de la qual no hi ha existència de moments d'ordre quatre.

CAPÍTOL 6

Altres exemples d'aplicació a dades reals

En el capítol anterior es mostra un exemple d'aplicació d'EVT en un camp on el seu ús és relativament recent, però en camps com les finances o la hidrologia fa dècades que s'està aplicant aquesta teoria amb èxit. La predicció de grans quantils per controlar pèrdues o riscos naturals utilitzant EVT és una pràctica habitual.

En aquest capítol s'estudiaran alguns del casos més habituals en els quals s'aplica EVT i es comentaran alguns dels supòsits que, de vegades, els investigadors deixen una mica de banda, quan són una part fonamental de l'estudi, com la independència de les dades o l'adequació del model.

6.1 Dades daneses

El primer conjunt de dades que es tractarà en aquest capítol és el de les dades daneses que s'han fet servir com a exemple per il·lustrar els diferents gràfics i mètodes que s'han anat presentant. Com que els resultats principals ja s'han mostrat, es faran alguns comentaris sobre les dades i un resum de les conclusions. Aquestes dades es poden trobar al paquet `evir` de l'R amb el nom de `danish`.

Les dades sobre pèrdues en assegurances per focs van ser estudiades per primera vegada a McNeil (1997) i han estat molt comentades ja que són un bon exemple de l'aplicació del mètode PoT. Cal destacar Resnick (1997), en el qual els principals comentaris van dirigits al diagnòstic per avaluar l'adequació del model de cua pesada i al diagnòstic de la independència.

Per al diagnòstic d'un model de cua pesada es proposen mètodes suplementaris a l'*ME-plot* i al *QQ-plot*, un dels quals és el *Hill-plot*. Tots ells avalen la hipòtesi de model

de cua pesada. Per al diagnòstic de la independència, es presenten diferents tests, ja que a McNeil (1997) es dóna per bona a partir del gràfic d'autocorrelacions, que ni tan sols es mostra a l'article. L'únic que rebutja independència de les dades és el test de rangs, per tant la hipòtesi de la independència de les dades sembla també adient.

Degut a que la independència és una de les hipòtesis clau del model, el primer que es farà serà tractar el seu estudi en aquestes dades, tal com s'ha fet amb les dades de temps d'execució, amb el test de Ljung-Box. Per començar, el test de Ljung-Box no es pot aplicar directament a les dades perquè és necessària l'existència de moments finits fins a ordre quatre per poder aplicar-lo. Malgrat tot, aquesta falta de moments no és un problema greu, ja que es troba resolt amb les transformacions presentades en el capítol 3. Un cop aplicades en el cas de les dades daneses, es pot suposar que la nova mostra poseeix moments de tots els ordres i es pot aplicar el test perquè es satisfan les hipòtesis.

Si es considera la mostra de 2.156 dades i s'apliquen tant la transformació $Y = -1/X$ com $Y = -1/(X + c) + c$, els p-valors del test de Ljung-Box són significatius, per tant la hipòtesi d'independència per a aquestes dades hauria de ser qüestionada, ja que sota canvis d'escala la independència de les dades es manté. Ara bé, si es considera un llindar de dos milions de corones i les mateixes transformacions, els p-valors ja no són significatius. És per aquest motiu que a l'Exemple 4.4 el llindar considerat era de dos milions de corones, per garantir la hipòtesi d'independència de les dades. Per al cas d'un llindar de 10 milions, que s'ha utilitzat per realitzar contrastos amb els tests T_m, evidentment, tampoc s'obtenen p-valors significatius en realitzar el test a les dades transformades.

És important verificar bé les hipòtesis per no arribar a conclusions errònies. Per exemple, si s'aplica el test de Ljung-Box a les 2.156 dades daneses sense transformar, no es rebutja la independència de les dades, però cal tenir en compte que una hipòtesi que han de satisfer les dades per poder aplicar aquest test és l'existència de moments de quart ordre, si això no es té en compte i es dóna per vàlid el resultat del test, qualsevol càlcul posterior s'estaria realitzant sota condicions incorrectes.

Un cop estudiada la independència, i assumint que es satisfà aquesta hipòtesi, només cal resumir els resultats que s'han obtingut a través dels diferents estudis que han fet servir mètodes gràfics i d'estimació, contrastos i algoritmes de selecció de llindars.

Fent servir el mètode PoT, es pot considerar com a llindar raonable un valor de 10 milions de corones a través de l'observació de l'*ME-plot* i una estimació per a ξ de 0,5 mitjançant ML. Del *Hill-plot* i l'estimador de Hill modificat s'obtenen unes estimacions de ξ properes a 0,7, proposant un model amb cua més pesada que l'anterior. L'ús dels tests T_m ha permès contrastar la hipòtesi de que aquestes 109 darreres dades provenen d'una GPD, tant si es considera $\xi = 0{,}5$ com ξ desconegut, obtenint una estimació propera a 0,5 en el darrer cas.

Per finalitzar, considerant les dades a partir d'un llindar de dos milions de corones,

s'ha aplicat l'algoritme de selecció de llindars que fa servir els tests T_m obtenint un llindar molt proper a 10 milions però amb una estimació una mica major que 0,5.

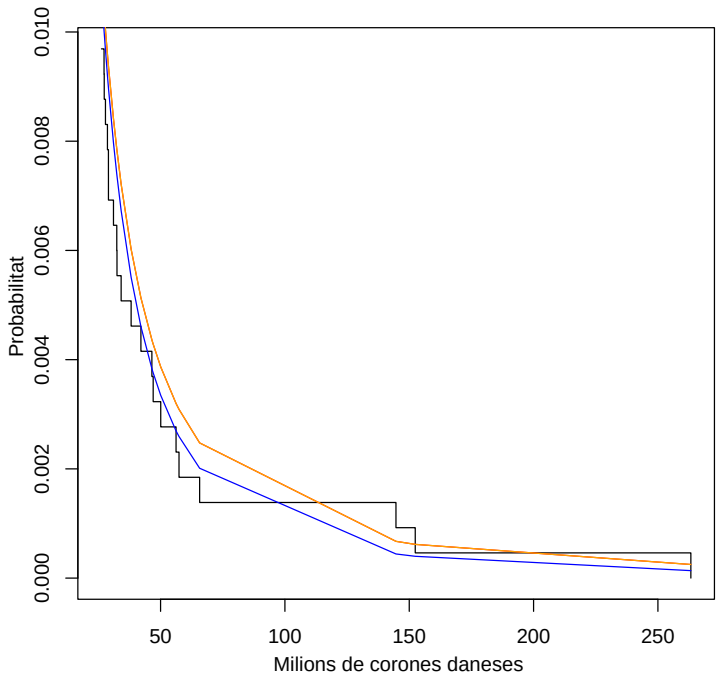

Figura 6.1: En negre, CCDF empírica de les dades daneses, en blau CCDF fent servir un llindar de 10 milions i $\xi = 0{,}5$ i en taronja CCDF fent servir un llindar de 9,997 milions de corones daneses i $\xi = 0{,}62$. En tots els casos es mostren les funcions a partir del quantil 99.

A la Figura 6.1 es pot veure la funció complementària de la distribució acumulada (CCDF) a partir del quantil 99 en diferents casos. En negre hi ha respresentada la CCDF empírica de les dades daneses, en blau la CCDF amb llindar i paràmetre donats pel mètode PoT i, finalment, en taronja la CCDF amb llindar i paràmetre donats per l'algoritme de selecció de llindars. Aquestes dues últimes funcions són molt properes, es pot observar que la primera ajusta millor entre 40 i 100 milions de corones daneses i la segona cap a la part final fins al 260 milions de corones daneses.

6.2 Dades financeres

Les finances són un dels camps on més s'aplica l'EVT, sobretot per calcular el VaR. No podia faltar, per tant, un exemple d'aplicació en aquesta àrea. En aquest apartat es realitzarà un estudi de les cotitzacions de les empreses IBM (IBM) i Microsoft (MSFT), l'índex industrial Down Jones (DJI) i el tipus de canvi Euro/Dòlar (EUSD) durant el mateix període de temps, des de l'1 de gener de 1999 al 17 de novembre de 2005.

Les dades que s'estudiaran exactament són els rendiments logarítmics negatius en valor absolut de les dades, és a dir, la cua de les pèrdues. Es dibuixaran els *CV-plots* corresponents, s'aplicarà el mètode T_m a aquests rendiments, es compararan els resultats amb altres mètodes, entre ells el presentat a Gomes i Pestana (2007), i també es realitzaran dues reflexions referents a l'adequació del model de cua pesada per a dades financeres i la diferenciació entre *outlier* i valor extrem.

Abans de començar l'estudi convé realitzar una descripció inicial de les dades. En aquest cas, una representació de les sèries temporals corresponents sembla el més adient. A la Figura 6.2 es poden veure les sèries corresponents als rendiments logarítmics dels quatre conjunts de dades; cal fixar-se especialment en l'apartat (a), ja que aquest gràfic permetrà il·lustrar una de les reflexions esmentades en el paràgraf anterior, la que fa referència als *outliers*.

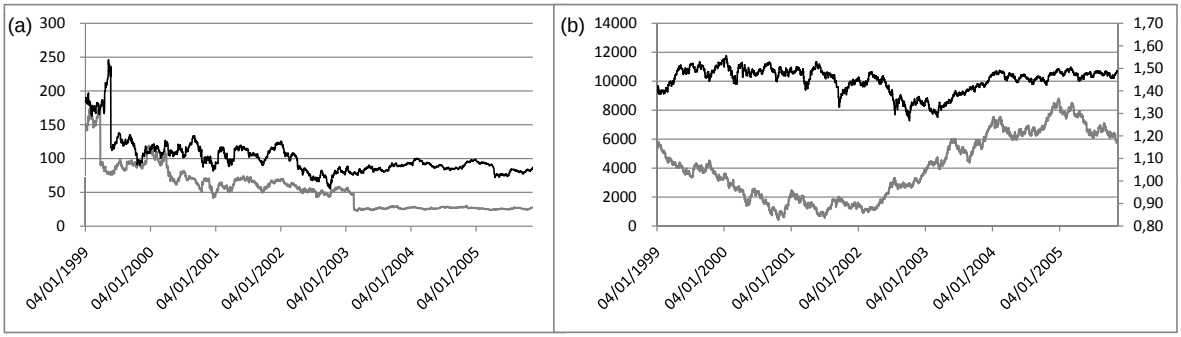

Figura 6.2: (a) Valors de tancament diaris de les dades IBM (línia negra) i MSFT (línia grisa). (b) Valors de tancament diaris de les dades DJI (línia negra i eix esquerre) i tipus de canvi diaris EUSD (línia grisa i eix dret).

Si s'observen amb deteniment les sèries IBM i MSFT, es pot veure com, abans d'arribar a mitjans de 1.999, hi ha una caiguda important en cada un dels dos casos i una altra, a principis de 2003, només per a MSFT. Aquests valors corresponen a dies en els quals es va realitzar un *split* de les accions corresponents. Un *split* és una acció corporativa en la qual una empresa divideix les seves accions existents en múltiples accions. Tot i que el nombre d'accions en circulació s'incrementa en un múltiple específic, el valor total monetari de les accions segueix sent el mateix en comparació amb les quantitats *pre-split*, perquè la divisió no afegeix cap valor real a l'empresa.

Aquests tres punts concrets, que provoquen un descens elevat dels rendiments, no han de ser considerats com a valors extrems sino com a *outliers*, perquè aquest descens que es produeix no és un descens real del valor monetari i només és degut a l'increment del número d'accions. Cal anar molt amb compte amb la distinció entre aquests dos conceptes, perquè un valor extrem sí que s'ha d'incloure a l'estudi, ja que és una da-

da que s'ha observat sota condicions normals, mentre que un *outlier* és preferible que desaparegui d'aquest, ja que la seva aparició va lligada a algun fet no normal.

A continuació, s'iniciarà l'estudi de les dades amb la representació dels *CV-plots*. En els casos de IBM i de MSFT tant amb *splits* com sense per poder veure les diferències. A més d'extreure informació sobre les cues, un dels casos servirà per introduir la segona de les reflexions que es comentaven a l'inici d'aquest apartat, la que fa referència als models de cues pesades per a dades financeres.

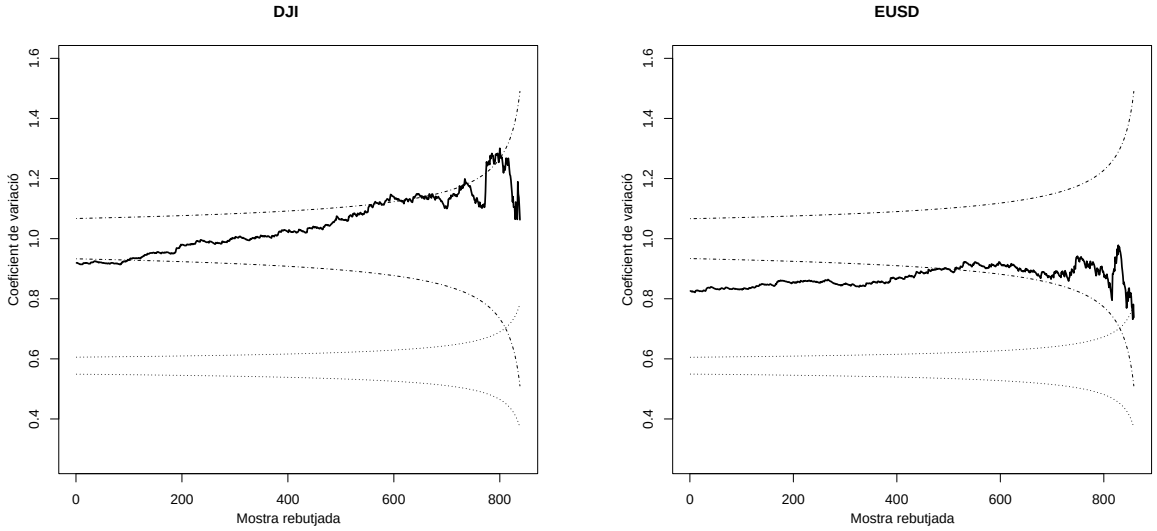

Figura 6.3: CV-plots dels rendiments logarítmics negatius en valor absolut per a DJI i EUSD. Els intervals de confiança per a les distribucions exponencial i uniforme es troben representats amb línies puntejades.

A la Figura 6.3 es poden veure els *CV-plots* per a DJI i EUSD. En el primer cas, la gràfica comença per sota de l'interval de confiança del 95% de la distribució exponencial i va ascendint de manera que algun valor se'n surt, però la majoria de valors queden a dins de l'interval, de manera que un model exponencial o de cua pesada molt propera a l'exponencial sembla adequat. En el segon cas, es pot observar clarament com la gràfica queda per sota de l'interval de confiança del 95% per a la distribució exponencial al llarg dels primers 500 llindars per acabar-hi entrant, en aquest cas el model a escollir sembla un de cua lleugera o exponencial.

Aquests dos casos són un bon exemple per reflexionar sobre el model que cal aplicar a dades financeres. Molts investigadors donen per fet que les dades financeres han de tenir cua pesada sense verificar d'alguna manera aquesta suposició i fan servir, entre d'altres, l'estimador de Hill, el qual retorna sempre una estimació de ξ positiva. Com passa amb el cas de la independència, si es consideren resultats quan les suposicions incials no es satisfan, les conclusions poden no ser correctes.

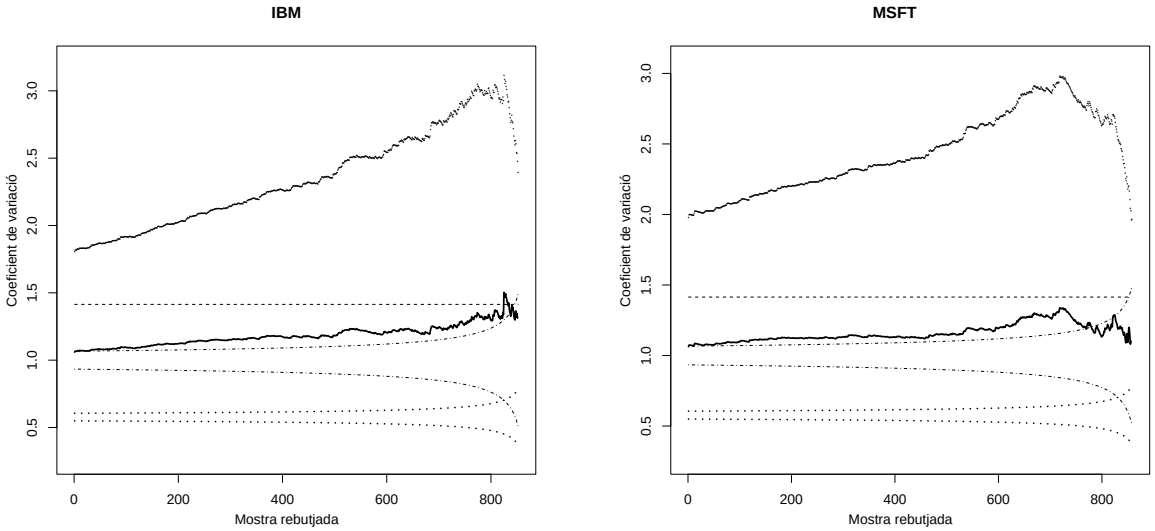

Figura 6.4: CV-plots dels rendiments logarítmics negatius en valor absolut per a IBM i MSFT, sense *splits* i considerant-los (gràfics per sobre de l'existència de moments de quart ordre, línia puntejada constant) . Els intervals de confiança del 95% per a les distribucions exponencial i uniforme es troben representats amb línies puntejades.

A la Figura 6.4 es poden veure els *CV-plots* per a IBM i MSFT, amb i sense *splits*. Com es pot observar, la diferència és evident, quan es consideren els *splits* les cues són molt pesades, mentre que si es treuen, les gràfiques es troben més a prop de la distribució exponencial. En tots dos casos sembla que el model adequat sí que seria el de cua pesada propera a l'exponencial, ja que els CV residuals queden molt a prop del límit superior de l'interval de confiança.

Per continuar l'estudi, es faran servir diferents mètodes per modelitzar les dades i es compararan els resultats. Per una banda, s'agafaran els màxims mensuals i s'estimaran els paràmetres per ML d'una GEV, per una altra, es farà ML de tota la msotra sota exponencialitat, GPD i FTG i, per acabar, s'aplicaran el mètode T_{20} i l'algoritme de Gomes i Pestana (2007). Aquests mètodes es denotaran per GEV, EXP, GPD, FTG, T_{20} i GP, respectivament. Cal fer notar que, excepte només en un cas, l'aplicació del mètode T_{20} ha donat els mateixos resultats fent servir les diferents significacions proposades i per tant es considera la significació 0,05.

Un cop estimats els paràmetres s'han calculat els VaRs al 99,9%. A la Taula 6.1 es poden veure els resultats obtinguts. Per a DJI i EUSD, tots els mètodes són coherents amb el que s'observba en els *CV-plots*, un model exponencial o de cua poc pesada per a DJI i un model exponencial o de cua lleugera per a EUSD, excepte GP, que en els dos casos estima cues molt pesades. Cal recordar que aquest mètode es basa en la hipòtesi de tenir cua pesada, si s'aplica sense aquesta condició, els resultats són inadequats.

Taula 6.1: VaR al 99, 9% i estimació del paràmetre ξ per als conjunts de dades MSFT, IBM, DJI i EUSD i els diferents mètodes aplicats. Per a IBM i MSFT també apareixen els casos sense *splits*, IMBs i MSFTs. L'asterisc per a FTG en MSFT, DJI i EUSD indica que el model seleccionat és la distribució gamma que acaba tenint cua exponencial, per això líndex de la cua és 0.

	mètode	**99,9%**	$\boldsymbol{\xi}$		**mètode**	**99,9%**	$\boldsymbol{\xi}$
MSFT				MSFTs			
	GEV	0.660	0.42		GEV	0.276	0.19
	EXP	0.124	0		EXP	0.114	0
	GPD	0.179	0.15		GPD	0.124	0.04
	FTG	0.160	0*		FTG	0.178	0.13
	T_{20}	0.431	0.42		T_{20}	0.159	0.12
	GP	0.174	0.35		GP	0.149	0.32
IBM				IBMs			
	GEV	0.660	0.42		GEV	0.333	0.28
	EXP	0.110	0		EXP	0.104	0
	GPD	0.144	0.11		GPD	0.112	0.03
	FTG	0.147	0.11		FTG	0.166	0.15
	T_{20}	0.335	2.62		T_{20}	0.153	0.14
	GP	0.181	0.39		GP	0.161	0.36
DJI				EUSD			
	GEV	0.104	0.17		GEV	0.026	-0.23
	EXP	0.058	0		EXP	0.035	0
	GPD	0.051	-0.06		GPD	0.022	-0.21
	FTG	0.05	0*		FTG	0.026	0*
	T_{20}	0.064	0,06		T_{20}	0.024	-0.16
	GP	0.068	0.31		GP	0.027	0.26

Per a IBM i IBMs (dades sense els *splits*), es pot veure com els mètodes FTG i GP no es veuen quasi afectats, mentre que per a la resta varia considerablement l'estimació passant d'una cua molt pesada a una propera a l'exponencial. Per a MSFT i MSFTs la situació és anàloga al cas anterior. En aquesta ocasió també s'observa que, excepte GP, la majoria de mètodes són coherents amb el que es veia en els *CV-plots*, que un model de cua pesada però propera a l'exponencial sembla adient.

Per finalitzar l'estudi, es compararan els VaRs calculats per a EUSD fent servir les dades corresponents des del 17 de novembrede 2.005 fins al 5 de gener de 2.014. A la Figura 6.5 es poden veure la sèrie temporal dels rendiments logarítmics i les línies de VaRs corresponents: lila per a GPD, taronja per a T_{20}, vermella per a FTG i GEV, blava per a GP i verda per a EXP.

En aquest període de temps hi ha al voltant d'unes 2.000 dades, per tant el que s'esperaria és que només dues vegades es traspasés la línia del VaR. Es pot veure com els casos més realistes són GPD i T_{20}, tot i que hi ha un període a finals de 2008, corresponent a la crisi financera que van patir els Estats Units, on diverses vegades es traspassen tots els VaRs excepte el corresponent a EXP. Una altra vegada queda clara la importància de comprovar l'adequació del model abans d'estimar paràmetres, els mètodes que proposen cues exponencials o cues pesades ofereixen estimacions massa restricitves per al VaR.

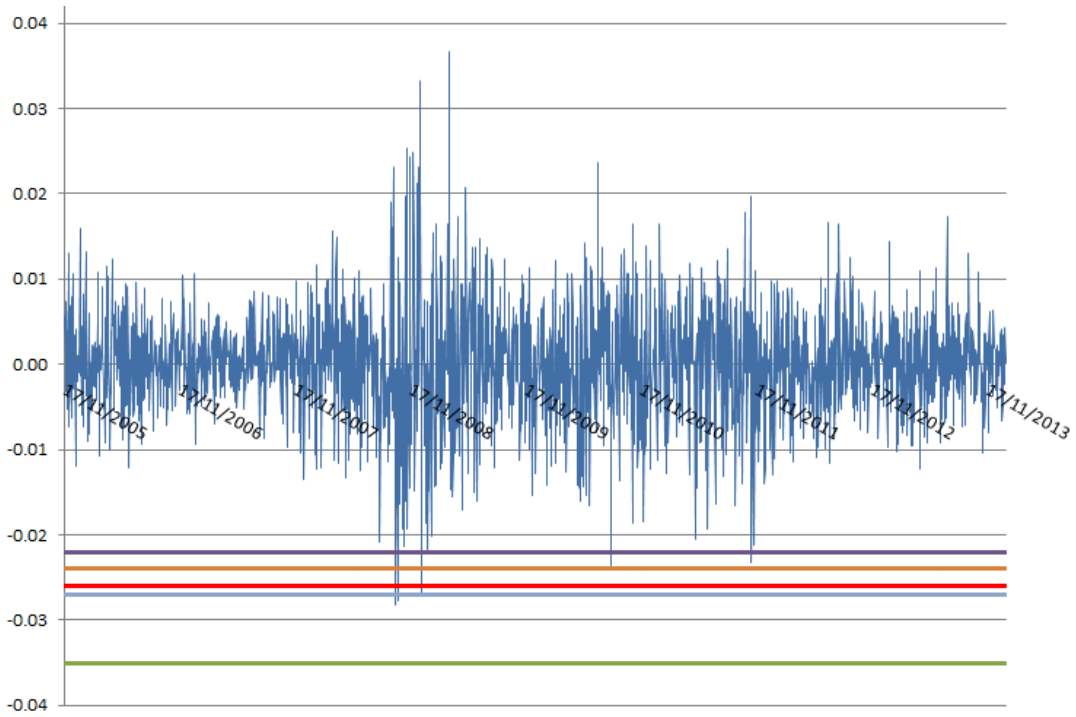

Figura 6.5: Sèrie temporal dels rendiments logarítmics per a EUSD des del 17 de novembre de 2.005 fins al 5 de gener de 2.014. En lila, taronja, vermell, blau i verd els VaRs al 99% obtinguts mitjançant els mètodes GPD, T_{20}, FTG/GEV, GP i EXP, respectivament.

CAPÍTOL 7

Conclusions del treball i futures línies de recerca

En aquest treball s'ha presentat un estudi exhaustiu de les noves eines basades en el coeficient de variació residual que es poden aplicar en EVT. Concretament, s'ha ampliat l'estudi del *CV-plot* a distribucions diferents de l'exponencial i s'han presentat nous estadístics de contrast per a la GPD.

Aquestes eines són aplicables quan l'enfoc escollit per modelitzar les cues és el dels excessos sobre un llindar i quan es fa servir el mètode PoT. Els estudis realitzats, aplicant les noves eines a les diferents dades analitzades, porten a remarcar el següents punts:

- Tot i que la teoria probabilista de l'EVT està desenvolupada, la seva aplicació pràctica comporta una sèrie de problemes a nivell d'estimació de paràmetres i selecció de llindars.

- L'utilització del test de Ljung-Box amb les transformacions presentades en el Capítol 3 serveix tant per contrastar la hipòtesi d'independència de les dades com per escollir entre GEV i GPD.

- Existeix una dualitat entre cues lleugeres i pesades per a la GPD, la qual es pot associar a una transformació, que permet solucionar el problema de l'existència de moments, de manera que les noves eines presentades es poden fer servir en qualsevol situació. La transformació que permet passar de cues pesades a lleugeres, i a l'inversa, ha estat desenvolupada com a part de les novetats presentades en aquesta tesi.

- L'ús del mètode gràfic *CV-plot* aporta avantatges en l'estudi de valors extrems respecte l'*ME-plot* i el *Hill-plot*, en el primer cas perquè el CV residual no depèn del

paràmetre d'escala i és més fàcil detectar una recta constant que una funció lineal, de manera que la incertesa es redueix essencialment de tres a un sol paràmetre, i en el segon perquè es pot aplicar tant a cues lleugeres com pesades, mentre que el *Hill-plot* només s'aplica a cues pesades. A més a més, el *CV-plot* es pot aplicar com a eina de bondat d'ajust per a distribucions diferents de la GPD.

- S'han presentat uns nous estadístics, anomenats T_m que es poden fer servir a diferents nivells. En primera instància es poden fer servir per estimar l'índex del valor extrem a partir de a seva minimització, també es poden fer servir com a tests de bondat d'ajust per a la GPD i, finalment, es poden utilitzar dins d'un algoritme de selecció automàtica de llindars, el qual també serveix per estimar l'índex del valor extrem.

- Les noves eines, *CV-plot* i tests T_m, s'han aplicat a dades de temps real de sistemes encastats, un camp on la teoria de valors extrems està prenent importància. Dels 16 conjunts estudiats, el mètode ha funcionat bé en la majoria del casos i no acaba de reflectir el comportament de la cua en casos particulars on sembla que pot haver una mixtura de distribucions, una de les quals queda bastant centrada a la cua.

- S'han aplicat les noves eines a dades que es troben bastant estudiades a literatura, com dades financeres, sobre assegurances en focs i d'onades, i els resultats obtinguts han resultat semblants als que resulten de fer servir altres mètodes. Fet que destaca la coherència del mètode.

Futures línies de recerca relacionades amb la teoria desenvolupada en aquesta tesi són la comparació de diferents tests d'uniformitat i l'estudi d'estimació de llindars en mixtures de distribucions. En el primer cas, per exemple, l'aplicació del test d'Anderson-Darling per a una distribució uniforme amb paràmetres desconeguts requereix d'una estimació prèvia d'aquests, per contrastar una distribució uniforme a l'interval (0,1), si es fa servir l'estimador ML no és possible calcular l'estadístic corresponent en alguns casos, mentre que el contrast T_m per a la GPD_{-1} no requereix de cap estimació prèvia encara que els paràmetres siguin desconeguts.

Per una altra banda, en el mateix sentit que la teoria asimptòtica s'ha aplicat per trobar intervals de confiança puntuals per al *CV-plot* teòric de la GPD, es podrien buscar els intervals de confiança puntuals per als *CV-plots* teòrics ja estudiats.

Bibliografia

[1] Abella, J., Padilla, M., Castillo, J. & Cazorla, F.J. (2015). Relating Processor Design, Execution Time Distributions and Extreme Value Theory. *Presentat*.

[2] Balkema, A. & de Haan, L. (1974). Residual life time at great age. *Annals of Probability*, *2*, 792-804.

[3] Barndorff-Nielsen, O. E. (1997). *Normal inverse Gaussian distributions and stochastic volatility modelling*. Scandinavian Journal of statistics, *24(1)*, 1-13.

[4] Beirlant, J., Dierckx, G., Goegebeur, Y., & Matthys, G. (1999). Tail index estimation and an exponential regression model. *Extremes*, *2(2)*, 177-200.

[5] Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J., De Waal, D. and Ferro, C.. (2004) *Statistics of Extremes: Theory and applications*. Wiley Series in Probability and Statistics.

[6] Castillo, E. & Hadi, A. S. (1997). Fitting the generalized Pareto distribution to data. *J. Amer. Statist. Assoc*, *92*, 1609-1620.

[7] Castillo, J. del & Daoudi, J. (2009). Estimation of the generalized Pareto distribution. *Statistics & Probability Letters*, *79*, 684-688.

[8] Castillo, J. del, Daoudi, J. & and Lockhart, R. (2014a) Methods to distinguish between polynomial and exponential tails. *Scandinavian Journal of Statistics*, *41*, 382-393.

[9] Castillo, J. del, Daoudi, J. & Serra, I. (2013) The full-tails gamma distribution applied to model extreme values. http://arxiv.org/abs/1211.0130

[10] Castillo, J. & Padilla, M. (2015) 'Modeling extreme values by the residual coefficient of variation', *arXiv:1510.00179 [math.ST]*.

[11] Castillo, J., Padilla, M., and Serra, I. (2014b) Comparison of techniques for extreme values using financial data, *Proceedings of the 21st International Conference on Computational Statistics hosting the 5th IASC World Conference*, COMPSTAT '14, 45-52.

[12] Castillo, J. del & Serra, I. (2015). Likelihood inference for Generalized Pareto Distribution. *Computational Statistics and Data Analysis*, *83*, 116-128.

[13] Cazorla, F.J., Vardanega, T., Quinones, E., and Abella, J. (2013) Upper-bounding Program Execution Time with Extreme Value Theory. *Worst-Case Execution Time workshop*.

[14] Choulakian, V. & and Stephens, M. A. (2001). Goodness-of-Fit for the Generalized Pareto Distribution. *Technometrics*, *43*, 478-484.

[15] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer Series in Statistics.

[16] Cucu-Grosjean, L., Santinelli, L., Houston, M., Lo, C., Vardanega, T., Kosmidis, L., Abella, J., Mezzetti, E., Quinones, E., and Cazorla, F.J. (2012) Measurement-Based Probabilistic Timing Analysis for Multi-path Programs, *Proceedings of the 2012 24th Euromicro Conference on Real-Time Systems*, ECRTS '12, 91-101.

[17] Davison, A. C. (1984). Modelling Excesses Over High Thresholds, with an Application. *Statistical Extremes and Applications*, Springer Netherlands, 461-482.

[18] Davison, A. C. & Smith, R. L. (1990). Models for exceedances over high thresholds. With discussion and a reply by the authors. *J. Roy. Statist. Soc. Ser. B*, *52*, 393-442.

[19] Diebold, F., Schuermann, T. & Stroughair, J. (1998) Pitfalls and Opportunities in the Use of Extreme value Theory in Risk management. *Working papers series*, FIN-98-081.

[20] Drees H., de Haan, L. & Resnick, S. (2000). How to make a Hill plot. *Annals of Statistics*, *28*, 254-274.

[21] Embrechts, P. Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin.

[22] Fisher, R.A. & Tippett, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press *24* 02.

[23] Gnedenko, B. V. (1948). On the local limit theorem of probability theory. *Uspekhi Matematicheskikh Nauk 3(3)*, 187-194.

[24] Gomes, M.I. & Pestana, D. (2007). A sturdy reduced-bias extreme quantile (VaR) estimator. *Journal of the American Statistical Association*, *102(477)*, 280-292.

[25] Gupta, R. C. & Kirmani, S. N. U. A. (2000). Residual coefficient of variation and some characterization results. *Journal of statistical planning and inference*, *91*, 23-31.

[26] Hosking, J.R.M. & Wallis, J.R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, *29*, 339-349.

[27] Leadbetter, M.R., Lindgren, G. & Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics.

[28] McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *Astin Bulletin*, *27(01)*, 117-137.

[29] McNeil, A. J., Frey, R. & Embrechts P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools.* Princeton University Press.

[30] Ocaña, J., Sánchez, M. P., Pla, A. S., & Pla, J. L. C. (2008). On equivalence and bioequivalence testing. *SORT: statistics and operations research transactions*, *32(2)*, 151-176.

[31] Padilla, M., Castillo, J., Abella, J. i Cazorla, F.J. (2015). Execution time distributions in embedded safety-critical systems using extreme value theory. *Presentat.*

[32] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, *3*, 119-131.

[33] Poovey, J. (2007) Characterization of the EEMBC Benchmark Suite. North Carolina State University.

[34] R Development Core Team (2010). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[35] Reiss, R.D. & Thomas, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields.* Birkhäuser.

[36] Resnick, S.I. (1997). Discussion of the Danish data on large fire insurance losses. *Astin Bulletin*, *27*, 139-151.

[37] Slijepcevic, M., Kosmidis, L., Abella, J., Quiñones, E., and Cazorla, F.J. (2014) Time-Analysable Non-Partitioned Shared Caches for Real-Time Multicore Systems, *Proceedings of the 51st Annual Design Automation Conference*, DAC '14, 1-6.

[38] Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, *72*, 67-90.

[39] SoCLib. http://www.soclib.fr/trac/dev.

[40] Song, J. & Song, S. (2012). A quantile estimation for massive data with generalized Pareto distribution. *Computational Statistics and Data Analysis*, *56*, 143-150.

[41] Sullo, P., & Rutherford, D. K. (1977). Characterizations of the power distribution by conditional exceedance. *Proceedings of the Business and Economic Society*, American Statistical Association.

[42] Thompson, M. L., Reynolds, J., Cox, L. H., Guttorp, P., & Sampson, P. D. (2001). A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment*, *35*, 617-630.

[43] Wilhelm, R., Engblom, J., Ermedahl, A., Holsti, N., Thesing, S., Whalley, D., Bernat, G., Ferdinand, C., Heckmann, R., Mitra, T., Mueller, F., Puaut, I., Puschner, P., Staschulat, J., and Stenström, P. (2008) The Worst-case Execution-time Problem - Overview of Methods and Survey of Tools, *ACM Trans. Embed. Comput. Syst.*, Vol. 7, No. 3, pp.36–53.

[44] Zhang, J. & Stephens, M. A. (2009). A new and efficient estimation method for the generalized Pareto distribution. *Technometrics*, *51*, 316-325.