Distribucions amb cues pesades en dades financeres

Joan del Castillo
Universitat Autònoma de Barcelona

Resum

En aquest treball proposem un nou enfocament metodològic per a estudiar el pes de les cues de la distribució marginal d'un procés de preus. El nostre enfocament es basa a considerar, en primer lloc, la simetria d'aquesta distribució; en segon lloc, la utilització del coeficient de variació per a mesurar el pes de les cues; i en tercer lloc, un gràfic de la raó d'atzar. La conclusió final és que les dades financeres s'ajusten bé a un model amb cues semipesades.

1. Introducció

El model bàsic per estudiar l'evolució dels mercats financers és avui el model anomenat de Black i Scholes, després del seu destacat treball sobre valoració de derivats de 1973 (tot i que en realitat el model era utilitzat per Samuelson el 1965). És ben conegut que aquest model és només una primera aproximació a la realitat i, per tant, en presenta diferents desviacions. El model suposa que els logaritmes dels preus d'un actiu financer segueixen un passeig aleatori, és a dir, les diferències dels logaritmes, o rendibilitats contínues, són una successió de variables normals, independents, idènticament distribuïdes.

El model de Black i Scholes és prou aproximat per estudiar dades dia a dia d'un actiu de renda variable, s'utilitza també en models de tipus de canvi i ja presenta serioses dificultats per a models de renda fixa. Hi ha estudis economètrics al voltant de les tres hipòtesis bàsiques del model: normalitat, increments estacionaris i increments independents. Recordem també que el model inclou la hipòtesi d'eficiència, que ha estat àmpliament debatuda i contrastada. Per a estudis economètrics sobre dades dia a dia vegeu Taylor (1986) o Campbell, Lo i MacKinlay (1997).

Nosaltres ens centrarem en les desviacions de la normalitat i més concretament en el fet empíric que hom observa més valors extrems, més valors petits i menys valors moderats dels que caldria esperar d'una distribució normal. Aquest fet, que és observable en dades dia a dia, augmenta progressivament en estudiar intervals de temps menors, que avui arriba fins a les dades *tic-a-tic*. Un extens treball sobre dades intradia és de Guillaume *et al.* (1997). En el seu treball afirmen que les cues tenen un decreixement polinomial amb $\alpha \approx 3{,}5$, la qual cosa implica la no-existència de moments de quart ordre. Aquest fet està també recollit a Shiryaev (2000).

La mesura més habitual, en dades financeres, per estudiar del pes de les cues és la curtosi, basada en el moment d'ordre quatre. Duffie i Pan (1997) discuteixen la variació de la curtosi segons l'horitzó temporal i el model que hom suposa. En models de tipus de canvi, en dades observades cada 30 minuts, s'observen curtosis entre 20 i 60 (Müller *et al.*, 1998). Hull i White (1998) estudien el *VaR* (*value at risc*) en dades diàries de canvi de moneda, on el pes de les cues és superior al cas de la normal, observant curtosis moderades (al voltant de 4) i proposen com a model alternatiu la mixtura de distribucions normals.

La metodologia *VaR*, basada en la hipòtesi de normalitat, s'interessa per probabilitats a l'interval 1 % a 5 %. Podem dir que en finances interessen els valors extrems, però no molt extrems o catastròfics. No hi ha, doncs, problemes amb dades setmanals, en canvi, si progressivament estudiem intervals més petits, diaris o dades intradia, és quan apareixen les cues pesades.

Les dades que hem escollit per il·lustrar aquest treball són el futur sobre l'Ibex 35, l'index de la borsa española. Les dades es prenen cada 5 minuts al llarg de tot un mes. Hi observem una curtosi de 39,7.

2. El problema de les cues

Quan parlem de distribucions amb cues pesades, probablement, pensem que no es compleixen les previsions que permet fer la hipòtesi de normalitat. Per exemple, pot fallar la previsió que una desviació de la mijana superior a 1,65 vegades la distribució estàndard només es produeix un 95 % de vegades. No obstant això, el problema de les cues pesades en finances ha estat enfocat de diverses maneres, depenent de les hipòtesis, dels objectius o de les simplificacions que hom assumeixi.

Un dels enfocaments àmpliament estudiat parteix de les hipòtesis d'increments independents i estacionaris, sense suposar normalitat o continuïtat de les trajectòries. Aquest punt de vista dóna lloc a l'estudi dels processos de Lévy, en què les distribucions marginals considerades han de ser infinitament divisibles. Com a exemples podem citar els models de variància Gamma, de Madan i Seneta (1990), i la distribució Normal Inversa Gaussiana (NIG), de Barndorff-Nielsen (1998).

Un altre enfocament parteix del fet que la volatilitat no és constant. Així, les distribucions marginals dels preus apreixeran com a mixtures de distribucions. En aquest punt, podem citar la utilització de les distribucions hiperbòliques d'Eberlein i Keller (1995) i el treball de Barndorff-Nielsen i Shephard (2001).

Un tercer enfocament trasllada al camp financer el punt de vista de les assegurances, on el problema de les cues és a cops més seriós i cal prendre precaucions extremes. Així es recupera el punt de vista del teorema de Fisher-Tippett, sobre la distribució del màxim i, especialment, les caracteritzacions de Von Misses dels seus dominis d'atracció. Aquest punt de vista està perfectament detallat a Embrechts, Klüppelberg i Mikosch (1997).

A diferència dels anteriors, el nostre punt de vista segueix, en certa mesura, el de Hull i White (1998). Creiem com ells que per estudiar el *VaR* es pot buscar un model senzill, sense pensar en la dinàmica del procés. No obstant això, no ens basem en la mixtura de normals, sinó que plantegem models alternatius basats en la distribució de Pareto i la distribució Normal truncada.

Per tal de precisar els conceptes introduirem un cert vocabulari. Sigui x una variable aleatòria contínua amb densitat $f(x)$ i funció de distribució $F(x)$. Denotarem la funció de supervivència per $\overline{F}(x) = 1 - F(x)$.

Habitualment estarem interessats a determinar un valor v que només sigui superat per x en una probabilitat determinada, p, com ara 95 %. És a dir $\overline{F}(v) = p$. Quan parlem del pes de la cua dreta d'una distribució, volem dir que ens interessa conèixer com varia $\overline{F}(v)$, per a valors *grans* de v. Alternativament el pes de la cua esquerra correspon al comportament de $F(v)$ quan v és petit. Al llarg d'aquest treball ens situarem en el primer cas, el segon es redueix al primer canviant el signe del fenomen en estudi.

2.1. *Desigualtat de Txebixev i moments*

Estem acostumats a la interpretació de la desviació típica, o de la variància, que se segueix de la utilització de la distribució normal. No obstant això, en un context més ampli, hem de començar amb la interpretació que se segueix de la desigualtat de Txebixev. Aquesta ens diu que, si suposem que x és una variable aleatòria positiva amb esperança finita, per a tota constant positiva, v, tenim

$$\Pr\{x > v\} \leq \frac{E[x]}{v}. \tag{1}$$

Habitualment aquest resultat s'aplica en una variable aleatòria qualsevol amb variància, $V[x] = \sigma^2$, finita. Aplicant (1) a $|x - \mu|$, amb $E[x] = \mu$, tenim la desigualtat clàssica

$$\Pr\{|x - \mu| > v\} \leq \frac{\sigma^2}{v^2}.$$

Aquests resultats ens diuen com decreix *el pes de les cues.* En el primer cas, com a molt el pes de la cua decreix com la funció v^{-1}. En el segon cas, el decreixement és com v^{-2}.

L'existència de moments d'ordre superior ens dóna més control sobre aquest decreixement. De forma anàloga, si el moment centrat (o no centrat) d'ordre k, μc_k, és finit

$$\Pr\left\{\left|x-\mu\right|>v\right\}\leq\frac{\mu c_k}{v^k}. \tag{2}$$

Veiem així que el decreixement de la cua és com v^{-k}, si existeix el moment d'ordre k.

Podem fer un pas més suposant que x té funció generatriu de moments, és a dir, $M(t)=E[e^{tx}]$ és finita. Suposarem per simplificar que $x>0$, aleshores

$$\Pr\left\{x>v\right\}=\Pr\left\{e^{t\,x}>e^{t\,v}\right\}\leq M(t)\,e^{-t\,v} \tag{3}$$

Veiem que en aquest cas el decreixement de la funció de supervivència és exponencial. Aquest és el comportament típic de la distribució exponencial en què $\bar{F}(x)=e^{-x/\mu}$, on μ és el seu valor esperat.

2.2. *Classificació segons el pes de les cues*

El model bàsic que constitueix la distribució normal direm que és un model amb *cues lleugeres*. Realment el decreixement de les cues en la distribució normal és encara més ràpid que l'observat a (3), donat que en aquest cas fins i tot x^2 té funció generatriu de moments. Estendrem el nom de cues lleugeres en aquells models amb un decreixement similar o més ràpid que el de la distribució normal.

Anomenarem models amb cues *semipesades* els que tenen un comportament com a (3). Entre aquests models cal destacar, a més de la distribució exponencial, els models NIG de Barndorff-Nielsen (1998) i el models hiperbòlics d'Eberlein i Keller (1995) utilitzats per modelar dades financeres.

Els models amb un decreixement polinomial de les cues, similiar a (2), els anomenarem models de *cues pesades*. El cas més característic d'aquests models de cues pesades és la distribució de Pareto, que estudiarem en l'apartat 6. Altres distribucions que es troben en aquest grup són les distribucions *estables*, àmpliament tractades en finances des de Mandelbrot (1965) i les distribucions de *variació regular*, vegeu Embrechts, Klüppelberg i Mikosch (1997).

Aquesta classificació no és exhaustiva. Per exemple, la distribució Lognormal té moments de tots els ordres, però no funció generatriu; es troba, doncs, entre les de cues pesades i les semipesades. Podríem solucionar el problema ampliant la definició, anomenant distribució amb cues pesades aquella que no té funció generatriu de moments. No és necessari, però, anar més enllà i amb els tres models definits ja en tindrem prou.

Cal tenir més compte en altres situacions. Per exemple, una distribució de Pareto truncada, diguem a $[0, 10^6]$, a efectes pràctics no la distigirem d'una Pareto sense truncar, tot i que una és una distribució amb cues pesades i l'altra no. Un altre cas delicat el dóna la mixtura de normals que presenta els problemes de les cues pesades essent un model de cues lleugeres com la normal, vegeu Hull i White (1998). En realitat es tracta d'aclarir quines probabilitats estem disposats a negligir, o quin rang de valors de $\bar{F}(x)$ ens interessa realment. Diferents problemes donaran lloc a diferents respostes. En assegurances interessen els valors molt extrems, els valors catastròfics, que tot i ser molt improbables donen lloc a pèrdues molt elevades que no es poden menystenir. Quan calculem el *VaR*, no estem interessats en valors realment extrems, sinó que ens movem en el rang de l'1 % al 5 %, aproximadament.

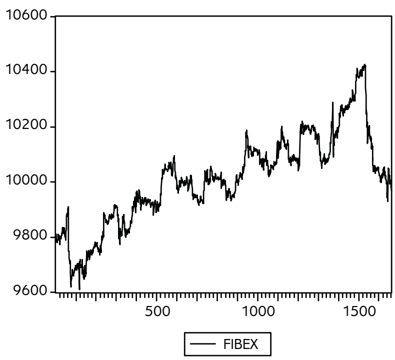
3. Normalitat, truncament i simetria

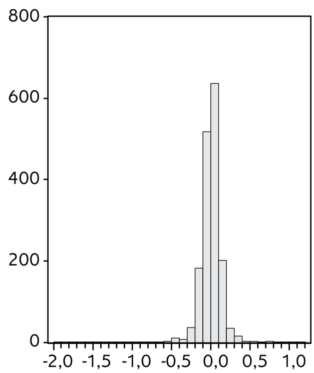
Per tal d'il·lustrar el nostre punt de vista utilitzarem l'evolució del futur de l'Ibex 35, a intervals de 5 minuts, al llarg de tot un mes (del 19/04/99 al 18/05/99). Les dades de futurs són el resultat d'operacions dutes a terme habitualment només per professionals.

Les observacions que estudiarem formen una mostra $\{x_1, x_2, ..., x_n\}$ de rendibilitats contínues, expressades en tant per cent, obtinguda de les cotitzacions P_t, d'acord amb el model de Black i Scholes, segons

$$x_t = 100 * \ln\left[\frac{P_t}{P_{t-1}}\right].$$

En total disposem de $n = 1661$ rendibilitats. Els preus inicials i finals són $P_1 = 9794$ i $P_{1662} = 9967$, la variació mensual ha estat doncs de 173 punts. Un histograma de la mostra presenta una elevada concentració al voltant de zero i una apreciable simetria. La curtosi és elevada (39,7) i un test de bondat d'ajust rebutja la hipòtesi de normalitat, tant el test χ^2_n com el de Jarque-Bera.





Series: REND	
Sample 2 1662	
Observations 1661	
Mean	0,001054
Median	0,000000
Maximum	1,141382
Minimum	-1,952670
Std. Dev.	0,144767
Skewness	-1,909680
Kurtosis	39,78043
Jarque-Bera	94634,60
Probability	0,000000

Els valors de les rendibilitats oscil·len entre ±0,2 % i només un 3,5 % dels valors que s'escapen a aquests límits. En un sol interval de temps s'han observat variacions comparables a la variació mensual. En 90 ocasions el preu no canvia.

La mitjana ($\overline{x} = 0{,}001054$) és poc informativa, representa només la variació mensual a través de

$$\overline{x} = \frac{100}{n} \ln \left[\frac{P_{n+1}}{P_1} \right].$$

Resulta més interessant, per conèixer el comportament dels operadors, estudiar la distribució dels valors absoluts de les observacions, o millor encara, la distribució de les pujades d'una banda, i la distribució de les baixades de l'altra. Així la mitjana de les pujades s'interpreta com a la rendibilitat que cal oferir al comprador per realitzar un venda de futur en un període de 5 minuts.

Les rendibilitats molts cops presenten distribucions força simètriques. En el cas present, aquest fet s'observa amb l'histograma i el valor petit del coeficient de biaix. No obstant això, diferents idees econòmiques suggereixen fer la distinció entre les dues cues de les distribucions. D'una banda, tenim el diferent comportament del mercat en períodes alcistes o baixistes, es puja més poc a poc que no pas es baixa. En l'estudi dels riscos de mercat preocupen les baixades i no els guanys exessius. El càlcul del *VaR* presenta una asimetria, ja que aquest representa el 5-percentil esquerre. En el risc de crèdit la asimetria és ja manifesta.

Donada una variable x, amb densitat $f(x)$, definim x_+ com al seu truncament positiu. Aleshores x_+ és una variable positiva amb densitat $f_+(x) = 2f(x)$, per a $x > 0$. Denotarem l'esperança per $\mu_+ = E[x_+]$ i la variància per $\sigma_+^2 = V[x_+]$. Per estudiar els valors negatius de x, considerarem $x_- = (-x)_+$, i així ens reduïm a estudiar variables positives. Naturalment sota la hipòtesi de simetria les distribucions de x_+ i de x_- coincideixen.

Des del punt de vista estadístic hom pot contrastar la simetria amb diferents testos, per exemple amb un test del tipus Kolomogorov-Smirnov. En qualsevol cas, veurem que estudiant les dues cues per separat hom té més informació que una suposada simetria. Suposem ara que x és una variable aleatòria amb densitat $f(x)$ simètrica, és a dir, $f(x) = f(-x)$. La primera conseqüència de la simetria és $E[x] = 0$ i, aleshores, $V[x] = E[x^2] = \sigma^2$, si suposem l'existència dels primers moments.

Proposició 1. *Per a una variable aleatòria simètrica,* x, *es compleixen les relacions:*

1. $E[|x|] = E[x_+] = E[x_-] = \mu_+$.
2. $V[x] = E[x^2] = E[x_+^2] = E[x_-^2] = \sigma^2 = \mu_+^2 + \sigma_+^2$

Prova. $E[|x|] = \int_{-\infty}^{\infty} |x| \; f(x)\, dx = 2\int_0^{\infty} x\, f(x)\, dx = \int_0^{\infty} x\, f_+(x)\, dx = \mu_+$.
$V[x] = E[x^2] = \int_{-\infty}^{\infty} x^2\, f(x)\, dx = 2\int_0^{\infty} x^2\, f(x)\, dx = E[x_+^2] = \mu_+^2 + \sigma_+^2$

La proposició anterior ens porta a fer algunes observacions. En primer lloc veiem que μ_+ s'interpreta com a la mitjana de les desviacions en valor absolut de x. Així és una mesura de la dispersió d'x, igual que ho és la desviació estàndard, σ. També observem que $\sigma_+^2 = V[x] - E[|x|]^2$

és la diferència de dues mesures de dispersió. Veurem que el pes de les cues està relacionat amb aquestes desviacions en valor absolut, que estan definides simplement amb l'existència del primer moment.

Paral·lelamemt al concepte de truncament positiu es pot desenvolupar el concepte de simetrització. Si y és una variable que pren valors positius, amb densitat, $f(y)$, podem definir la seva simetrització, y_s, com una variable amb densitat $f_s(y) = \frac{1}{2} f(|y|)$. Naturalment, donada una variable simètrica qualsevol, x, en fer truncament i simetrització, $(x_+)_s$, en resulta una variable amb la mateixa distribució que x. Així doncs, no és cap restricció treballar amb distribucions sobre els positius, perquè en cas de simetria recuperem la distribució inicial. Si no hi ha simetria, aleshores clarament és molt millor treballar per separat amb la part positiva i la part negativa.

Donada una variable x amb distribució normal, $N(\mu, \sigma^2)$, direm que el seu truncament positiu, x_+, té distribució normal truncada, $NT(\mu, \sigma^2)$. Aquest model el desenvolupem amb més detall a la secció 5.

Quan intentem aproximar unes dades financeres per a una distribució normal, estem aproximant-ne les parts positives i negatives per a distribucions normals truncades. En aquest cas, els dos truncaments són complementaris, perquè junts constitueixen una distribució normal. També es pot procedir a aproximar separadament cada cua per a una distribució normal truncada.

Les dades del nostre exemple es poden aproximar per a una distribució normal amb $\mu = 0{,}001054$ i $\sigma = 0{,}144767$, si bé ja hem dit que l'aproximació és dolenta. Podem intentar aproximar les parts positives i negatives per separat, mitjançant distribucions normals truncades. Estimant els paràmentres per màxima versemblança, ens trobem que, en ambdós casos, les equacions no tenen solució. Aquest fet ha estat estudiat àmpliament a Castillo (1994). Tal com enunciem al teorema 3 les equacions de versemblança no tenen solució quan el coeficient de variació empíric és més gran que 1. Aquest és el nostre cas, els coeficients de variació són $cv[x_+] = 1{,}14$ i $cv[x_-] = 1{,}39$, respectivament per a la part positiva i per a la part negativa de la mostra.

Treballar separadament amb les parts positives i negatives, tal com hem dit, és més flexible. En aquest exemple, però, ens fa veure que l'ajust a cues lleugeres com la de la normal és francament inapropiat. L'estimador màxim versemblant només es pot trobar com a un cas límit, dintre de la família exponencial, que donarà un millor ajust que les distribucions normals.

4. El coeficient de variació com a mesura del pes de les cues

El coeficient de variació és una eina usual en estadística per comparar variabilitats en relació a les mitjanes. Té sentit en distribucions centrades lluny de zero, perquè la mitjana apareix en el denominador. Nosaltres l'utilitzarem només en distribucions que prenen valors positius que, en general, seran la part positiva de distribucions simètriques amb moda a zero. Com a exemples bàsics de les distribucions en què estem pensant podem considerar la distribució exponencial, Exp(μ), o el truncament positiu d'una normal simètrica $NT(0, \sigma^2)$. Cal fer notar que en aquests casos l'esperança és més una mesura de dispersió que no pas de posició. En realitat en aquests exemples l'esperança és un paràmetre d'escala.

Definició 2. *Donada una variable, x_+, amb variància finita, σ_+^2 i esperança no nul·la, μ_+, el seu coeficient de variació és donat per*

$$cv[x_+] = \frac{\sigma_+}{\mu_+}.$$

Considerem ara x_+ com a la part positiva d'una distribució simètrica, x, amb variància $V[x] = \sigma^2$. En aquest cas, d'acord amb la proposició 3, tindrem

$$1 + cv^2[x_+] = \frac{\sigma^2}{\mu_+^2}. \tag{4}$$

veiem així que el coeficient de variació serveix per comparar dues mesures de dispersió d'x, com són la variància i la desviació en valor absolut.

Els primers avantatges del coeficient de variació són que no té unitats i que és invariant per canvi d'escala. Per exemple, si les rendibilitats estudiades s'expressen en tant per cent, tant per un o tant per mil, el coeficient de variació és sempre el mateix.

Un altre avantatge, en relació al problema de les cues pesades, és que el coeficient de variació només suposa l'existència del moment d'ordre 2. Així es pot utilitzar encara que les estimacions de Guillaume *et al.* (1997) sobre el pes de la cua ($\alpha \approx 3{,}5$) siguin correctes. En canvi, la curtosi necessita l'existència de moment d'ordre 4, que en el cas anterior no es lícit suposar.

Volem fer veure que el coeficient de variació, o el seu quadrat, és una bona mesura del pes de les cues, entenent com a tal la distribució entre valors petits, extrems i mitjans. En primer lloc veiem quin valor té en dos exemples bàsics. Si x_+ segueix una distribució exponencial, Exp(μ), és ben sabut que l'esperança i la variància són exactament μ i així el coeficient de variació val 1. Recordem que en aquest cas la simetritzada de x_+, es coneix com a distribució de Laplace.

Suposem ara que x_+ segueix una distribució normal truncada, $NT(0, \sigma^2)$. És a dir, suposem que x_+ és la part positiva d'una distribució, x, normal $N(0, \sigma^2)$. Aleshores de la proposició 3,

$$E\big[|x|\big] = E[x_+] = \frac{2}{\sqrt{2\pi}\sigma}\int_0^\infty x \ \exp\left[-\frac{x^2}{2\sigma^2}\right]dx = \frac{2\sigma}{\sqrt{2\pi}}$$

d'on, utilitzant (4), deduïm que

$$cv(x_+) = \sqrt{\frac{\pi - 2}{2}} \approx 0{,}75551$$

Aquest és el valor que aproximadament hem de trobar pel coeficient de variació, si les nostres dades segueixen una distribució normal centrada en zero, amb qualsevol variància. El grau d'aproximació depèn, naturalment, de la distribució asimptòtica del coeficient de variació empíric.

Distribució asimptòtica del coeficient de variació

Per tal de simplificar, farem servir ara la notació $c = cv(x_+)$, per a una variable qualsevol amb moments $\mu_k = E[x_+^k]$. Anomenarem c_n al coeficient de variació empíric en una mostra de mida n. És a dir, dient $\bar{x}$ a la mitjana mostral i s a la desviació estàndard, tenim

$$c_n = \frac{s}{\bar{x}} .$$

A partir de la distribució asimptòtica dels dos primers moments de la mostra, aplicant el mètode de linealització, o mètode delta, s'obté que

$$\sqrt{n}(c_n - c) \xrightarrow{d} N(0, \sigma_c^2) ,$$

on la variància asimptòtica és

$$\sigma_c^2 = \left\{ \left(3 - 4\frac{\mu_3}{\mu_1^3} + \frac{\mu_4}{\mu_1^4} \right) + \left(10 - \frac{\mu_3}{\mu_1^3} \right) c^2 + 11\, c^4 + 4\, c^6 \right\} / (4\, c^2) .$$

L'expressió anterior ens mostra que la distribució del coeficient de variació està relacionada amb el moment d'ordre quatre, que a la vegada està relacionat amb la curtosi.

En casos concrets la variància asimptòtica es simplifica notablement, així tenim en una distribució exponencial

$$\sqrt{n}(c_n - 1) \xrightarrow{d} N(0,1) . \tag{5}$$

En una distribució normal simètrica i positivament truncada, $NT(0, \sigma^2)$, la distribució asimptòtica és

$$\sqrt{n}\left(c_n - \sqrt{\frac{\pi - 2}{2}} \right) \xrightarrow{d} N\left(0, \frac{\pi^2(\pi - 3)}{4(\pi - 2)} \right) \approx N(0,\ 0{,}3) .$$

5. La distribució normal truncada

Donada una variable aleatòria contínua, x, amb densitat $f(x)$ i una constant a, la distribució de x truncada a $x > a$, té densitat

$$f(x|x>a)=\frac{f(x)}{\Pr\{x>a\}}\,.$$

Quan x té distribució normal, $N(\mu,\sigma^2)$, tindrem

$$\Pr\{x>a\}=1-\Phi\left(\frac{a-\mu}{\sigma}\right)=1-\Phi\,(\alpha)$$

on $\Phi\,(\cdot)$ és la funció de distribució de la normal estàndard i $\alpha=\frac{\alpha-\mu}{\sigma}$. L'esperança i la variància de la distribució normal truncada en el punt a s'expressen com a

$$\begin{aligned} E[x \mid x>a] &= \mu+\sigma\lambda(\alpha) \\ V[x \mid x>a] &= \sigma^2\big(1-\lambda(\alpha)^2+\alpha\lambda(\alpha)\big) \end{aligned} \tag{6}$$

on $\lambda(\alpha)=\frac{\phi(\alpha)}{1-\Phi(\alpha)}$ s'anomena l'invers de la raó de Mills, denotant per $\phi\,(\cdot)$ a la funció de densitat de la normal estàndard. De fet $\lambda(\alpha)$ és la funció d'atzar de la normal estàndard.

La distribució normal truncada ha estat àmpliament estudiada en economia, en relació a les mostres selectives, el treball de Heckman en aquest sentit va merèixer el premi Nobel d'economia l'any 2000. Vegis Heckman (1979) o Greene (1993).

Quan $a = 0$, tal com hem anat fent, direm que x_+ té distribució normal truncada, $NT(\mu,\sigma^2)$ (sense referència al punt de truncament). Des del punt de vista de l'estimació sempre ens podem reduir a aquesta situació. Les equacions de versemblança per a l'estimació del paràmetres s'obtenen igualant esperança i variància amb els corresponents valors de la mostra, $\bar{x}$ i s^2. Per tal de resoldre-les, a partir de (6), escrivim $cv^2(x_+)$ en termes d'$\alpha=-\frac{\mu}{\sigma}$. Aleshores, igualant el resultat al coeficient de variació empíric, c^2, tenim

$$\frac{\alpha^2+1-\alpha\lambda(\alpha)}{(\lambda(\alpha)-\alpha)^2}=1+c^2\,. \tag{7}$$

Obtenint α de l'equació anterior, trobem els estimadors dels paràmetres amb les expressions

$$\begin{aligned} \hat{\mu} &= \frac{\overline{x}\alpha}{\alpha-\lambda(\alpha)} \\ \hat{\sigma} &= \frac{\overline{x}}{\lambda(\alpha)-\alpha}\,. \end{aligned} \tag{8}$$

El següent resultat sobre les equacions de versemblança es troba demostrat a Castillo (1994).

Teorema 3. *Les equacions de versemblança (7) i (8) tenen solució si i només si el coeficient de variació de la mostra és inferior a 1. El cas límit,* c = *1, correspon a la distribució exponencial.*

Realment tenim una relació bijectiva entre el coeficient de variació i el paràmetre α, donada per l'equació (7). Aquest fet permet parametritzar la família de distribucions normal truncada a partir del coeficient de variació, que pren valors a (0, 1), i el paràmetre d'escala, σ. A més, aquesta relació permet demostrar el teorema següent, que es troba a Castillo i Puig (1999).

Teorema 4. *El millor test (UMPU) d'exponencialitat contra la distribució normal truncada té regió crítica de la forma* {c < k}*, on c és el coeficient de variació empírica i* k *és una constant que depèn del nivell de significació.*

6. Distribució de Pareto modificada

Clàssicament la distribució de Pareto es considera com aquella que té per funció de densitat $\frac{\alpha}{\sigma}\left(\frac{\sigma}{x}\right)^{\alpha+1}$, $(x > \sigma)$, on α s'interpreta com el pes de la cua i σ, que determina el suport de la distribució, es pot interpretar com a paràmetre d'escala. L'estimació màxima versemblant d'σ és, en aquest cas, el mínim de la mostra. Pareto havia introduït aquesta distribució a partir d'observar una relació lineal entre $Log[\bar{F}(x)]$ i $Log[x]$.[1]

Per tal d'estudiar dades financeres, on les distribucions són força simètriques al voltant de zero, és millor considerar la modificació:

$$p(x;\alpha,\sigma)=\frac{\alpha\sigma^{\alpha}}{(x+\sigma)^{\alpha+1}}\ ,\ (x>0) \tag{9}$$

que es coneix com a distribució de Pareto de segona espècie, o distribució de Lomax que, per simplificar, seguirem anomenant distribució de Pareto i representarem per $Par(\alpha, \sigma)$. Aquí σ és un veritable paràmetre d'escala, perquè si x té distribució $p(x; \alpha, \sigma)$ aleshores λx té distribució $p(x; \alpha, \lambda\sigma)$. En aquesta distribució modificada es presenta una relació lineal entre $Log[\bar{F}(x)]$ i $Log[x+\sigma]$. El comportament de les cues és exactament el mateix que en el cas de la distribució de Pareto clàssica i, en canvi, el suport és ara sempre el conjunt de valors positius. Estadísticament evitem el problema de treballar amb distribucions de suport variable.

La distribució de Pareto modificada té moments d'ordre k finits, si $k < \alpha$, que es poden calcular a partir de

1. Concretament, si n_x representa el nombre de persones amb renda superior a x, d'un conjunt de N persones, Pareto va observar $Log[n_x] = \alpha\ Log[x] + \beta$. Aleshores $\bar{F}[x] = \frac{n_x}{N} = \beta\ x^{\alpha}$.

$$E\left[(x+\sigma)^k\right]=\frac{\alpha\ \sigma^k}{\alpha-k,}$$

concretament, els dos primers moments són

$$E[x]=\frac{\sigma}{\alpha-1}\ , \qquad E\left[x^2\right]=\frac{2\sigma^2}{(\alpha-1)(\alpha-2)}\ ,$$

cosa que ens permet deduir pel coeficient de variació

$$cv^2[x]=\frac{\alpha}{\alpha-2}>1. \tag{10}$$

Així en una distribució de Pareto el coeficient de variació sempre és superior a 1. Quan α tendeix a 2 el coeficient de variació es fa infinit. Realment si $\alpha \leq 2$ no hi ha moment de segon ordre i aquest estadístic no està definit. També és clar que quan α tendeix a infinit el coeficient de variació s'acosta a 1. L'expressió (10) ens permet també una primera estimació consistent del pes de les cues

$$\tilde{\alpha}=\frac{2c^2}{c^2-1}\ ,$$

on c és el coeficient de variació empíric.

L'estimació màxima versemblant es pot resumir en trobar σ de l'equació

$$\left(\frac{1}{n}\sum\left(\frac{x_i}{\sigma}+1\right)^{-1}\right)\left(1+\frac{1}{n}\sum\log\left(1+\frac{x_i}{\sigma}\right)\right)==1$$

i després calcular

$$\hat{\alpha}=\left(\frac{1}{n}\sum\log\left(1+\frac{x_i}{\sigma}\right)\right)^{-1}$$

Es planteja a cops, però, el problema de la no-existència de solució. Aquest fet està relacionat en el fet que la distribució exponencial és també un cas límit de la distribució de Pareto. El límit s'assoleix quan α tendeix a infinit i, per tant, quan el coeficient de variació tendeix a 1, segons veiem a (10). És doncs natural que en mostres on el coeficient de variació empíric sigui menor que 1 acostumi a fallar la solució de les equacions.

Abans de la utilització de la distribució de Pareto cal fer un test per descartar si la mostra procedeix d'una distribució amb cues més lleugeres, com l'exponencial o, fins i tot, una normal truncada. La nostra proposta és fer un test basat en el coeficient de variació, per assegurar-se que aquest no és inferior a la unitat.

7. Reparametrització de les distribucions

Passem a proposar un model bàsic de distribucions decreixents amb la moda a zero, apropiat a dades financeres. Aquest model no és més que la reunió del model de normals truncades i les distribucions de Pareto, per a $\alpha > 2$. Indexem aquest conjunt de distribucions pel coeficient de variació i un paràmetre d'escala, anomenant-lo

$$NTP(cv,\sigma) = \begin{cases} NT(\mu,\sigma), & \text{si } cv < 1 \qquad \alpha, \sigma \text{ de les equacions (7), (8).} \\ Exp(\sigma), & \text{si } cv = 1 \\ Par(\alpha,\sigma), & \text{si } cv > 1 \qquad \alpha \text{ de l'equació (10).} \end{cases}$$

Disposem així d'una distribució per a cada coeficient de variació entre $(0, \infty)$. Quan $0 < cv < 1$, les nostres distribucions seran normals truncades; $c = 1$ correspon a la distribució exponencial i $1 < cv < \infty$ correspon a distribucions de Pareto.

És important en aquest model contrastar si $cv = 1$. El contrast pot fer-se a partir de (5). Si acceptem la hipòtesi nul·la, parlarem d'un model amb cues semipesades; quan resulti que $cv < 1$, parlarem de cues lleugeres i quan $cv > 1$, direm que el model és de cues pesades. El nostre model pretén una primera aproximació al problema de modelar unes dades financeres, pel que fa a la distribució marginal, i una classificació del pes de les cues d'aquesta distribució. Això no exclou una anàlisi posterior més detallada que utilitzarà unes tècniques o altres segons les primeres conclusions obtingudes.

A Hull i White (1998) s'estudien les taxes de canvi diàries de dotze monedes respecte el dòlar, entre el 4 de gener de 1988 i el 15 d'agost de 1997. Recollim, del seu treball, les dades de les monedes que presenten més i menys curtosi, així com els valors mitjans per a les dotze monedes i l'aproximació que proporciona la distribució normal. En cada cas s'especifica el percentge d'observacions que sobrepassen, en més o en menys, un múltiple de la desviació estàndard, σ.

	Espanya	*Dinamarca*	*mitjana*	*normal*
$Pr\{x > 1\sigma\}$	23,72	25,58	25,04	31,73
$Pr\{x > 2\sigma\}$	4,4	5,4	5,27	4,55
$Pr\{x > 3\sigma\}$	1,53	1,32	1,34	0,27
$Pr\{x > 4\sigma\}$	0,37	0,12	0,29	0,01
$Pr\{x > 5\sigma\}$	0,17	0,0	0,08	0,0
$Pr\{x > 6\sigma\}$	0,08	0,0	0,03	0,0
curtosi	7,75	1,61	3,39	0

Mentre que en la normal només un 0,3 % supera 3σ, en la corona danesa un 1,3 % supera 3σ i en la moneda espanyola un 1,5 % supera aquesta quantitat. Clarament l'aproximació normal és inapropiada. Si treballem amb dades intradia aquest fet s'accentua.

Passem a comparar aquests resultats amb la varietat de models que s'obtenen, amb el model que acabem de proposar, en funció del coeficient de variació.

	Normal	*NTrun*	*NTrun*	*NTrun*	*Exp*	*Par*	*Par*	*Par*
CV	0,76	0,85	0,91	0,94	1	1,11	1,25	1,67
$Pr\{x > 1\sigma\}$	31,73	28,74	27,02	26,00	24,31	22,91	21,41	18,06
$Pr\{x > 2\sigma\}$	4,55	5,48	5,80	5,86	5,91	5,81	5,63	5,06
$Pr\{x > 3\sigma\}$	0,27	0,68	0,98	1,15	1,44	1,61	1,73	1,83
$Pr\{x > 4\sigma\}$	0,01	0,05	0,13	0,19	0,35	0,48	0,60	0,79
$Pr\{x > 5\sigma\}$	0,00	0,00	0,01	0,03	0,08	0,15	0,23	0,38
$Pr\{x > 6\sigma\}$	0,00	0,00	0,00	0,00	0,02	0,05	0,10	0,20

Veiem que el model proposat recull comportaments diversos, que enllacen la normal amb l'exponencial i les distribucions de Pareto de manera contínua. La distribució de les cues depèn d'un sol paràmetre, el coeficient de variació. També podem apreciar que les dades de tipus de canvi són realment properes al model exponencial. Podríem parlar doncs d'un model amb cues semipesades per a les dades financeres.

8. Aproximació no paramètrica

Disposar d'un bon model paramètric és sempre el més pràctic. El problema és que a cops no tenim aquest model. Aleshores hem de recórrer als mètodes no paramètrics. En qualsevol cas, sempre és recomanable contrastar un model paramètric des d'un punt de vista diferent, que és el que farem en aquest apartat.

Introduirem algunes idees de l'anàlisi de la supervivència i la fiabilitat, que creiem importants en relació a l'estudi del pes de les cues i que, per ara, no han estat utilitzades en aquest context. Vegeu Barlow i Proschan (1981) o bé Castillo i Puig (1999) per a més bibliografia.

Per a una variable aleatòria positiva, x, es defineixen la *funció de risc* i la *funció de risc acumulat* com

$$r(x) = \frac{f(x)}{\bar{F}(x)}\,, \qquad R(x) = \int_0^x r(t)\,dt\ .$$

Es defineixen aleshores les classes de distribucions IFR (Increasing Failure Rate) quan la funció de risc és creixent i DFR (Decreasing Failure Rate) quan la funció de risc és decreixent. De la mateixa forma es defineixen les classes IFRA (Increasing Failure Rate in Average) quan $R(x)\ /\ x$ és creixent i DFRA (Decreasing Failure Rate in Average) quan $R(x)\ /\ x$ és decreixent.

Des del punt de vista de l'anàlisi de la supervivència i la fiabilitat, les distribucions importants són les IFR, que corresponen a temps de vida de fenòmens que envelleixen. Les distribucions

DFR corresponen a fenòmens que es rejoveneixen, dels quals hi ha pocs exemples. Els especialistes consideren central el concepte de funció de risc, per la seva interpretació pràctrica, mentre que la funció de risc acumulat és considerada una simple eina auxiliar.

Des del punt de vista estadístic, i més concretament no paramètric, és important distingir entre mètodes basats en estimacions de funcions de distribució i mètodes que utilitzen funcions de densitat. Les funcions de distribució s'estimen amb precisió, a partir de la funció de distribució empírica, mentre que les funcions de densitat s'estimen amb dificultat, quan no es disposa d'un model paramètric apropiat. Per tal d'estimar una derivada cal recórrer a fixar de forma arbitrària l'increment de la variable i aquest és l'origen dels problemes. En el context de l'anàlisi de la supervivència i la fiabilitat, la funció de risc és complicada d'estimar sense un model paramètric, perquè inclou la densitat, mentre que la funció de risc acumulat pot estimar-se amb precisió per la relació

$$R(x) = -\log\left[\overline{F}(x)\right]. \tag{11}$$

El teorema següent relaciona aquests conceptes amb les distribucions de cues pesades. El resultat, que s'aplica a la distribució de Pareto definida a (9), es troba demostrat a Barlow i Proshan (1981).

Teorema 5. *Cada una de les propietats donades a continuació implica la immediata següent:*

1. *El logaritme de la funció de densitat és una funció convexa.*
2. *La distribució és DFR.*
3. *La distribució és DFRA.*
4. *El coeficient de variació és més gran que la unitat* (cv > 1).

Les distribucions DRFA estan molt lligades a les distribucions de cues pesades. Es pot demostrar que presenten sempre *més risc* que les distribucions exponencials; en el sentit que qualsevol inversor, amb la seva pròpia funció d'utilitat, les rebutjaria. Per això, és important poder-les detectar empíricament, cosa que s'aconsegueix amb el resultat següent:

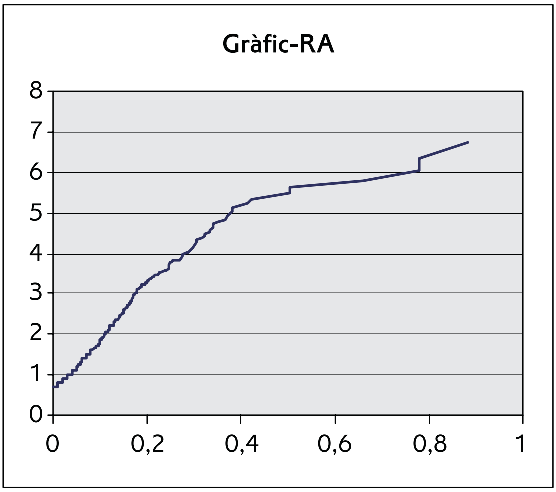
Teorema 6. *Una variable aleatòria positiva amb funció de distribució* $F(x)$ *és DFRA si i només si la seva funció de risc acumulat, donada per* (11) *talla com a molt una recta arbitrària* λx *en un punt, i el tall es produeix de dalt a baix.*

Aquest resultat ens permet proposar una nova eina per detectar el comportament de les cues. El gràfic de la funció de risc acumulat:

$$R_n(x) = -\log\left[\overline{F}_n(x)\right]$$

on $F_n(x)$ és la funció de distribució empírica associada a una mostra. Si la distribució de les dades correspon a un model amb cues pesades hem d'observar essencialment una funció còncava. Per contra, una distribució amb cues lleugeres, com la normal, presentaria un gràfic convex. Les distribucions amb cues semipesades s'han d'aproximar a una recta.

A continuació presentem el gràfic corresponent a les dades estudiades a la secció (3)



En aquest exemple hem de pensar en un model de cues més aviat pesades. Estimaríem, doncs, un model de Pareto. Concretament les dades s'ajusten a una distribució amb $\hat{\alpha} = 16{,}15$, que tot i ser una cua pesada és lluny d'altres estimacions que posen en dubte l'existència dels primers moments.

9. Conclusions

Resumim breument les idees defensades en aquest article. En primer lloc, creiem millor estudiar per separat les dues cues de la distribució. La cua que descriu com són les pujades i la que descriu les baixades. A més que, en general, són diferents, ens ajuden a comprendre el comportament dels agents en situacions alcistes o baixistes.

Hem presentat el coeficient de variació com un eina apropiada per estudiar el comportament de les cues. Aquest estadístic depèn només dels dos primers moments de la distribució, davant dels quatre moments que utilitza la curtosi. És doncs d'utilitat en situacions on està en dubte l'existència del quart moment, com les descrites per diversos autors en relació a les finances.

La distribució normal truncada i la distribució de Pareto presenten com a cas límit la distribució exponencial, que té coeficient de variació igual a 1. Ambdues distribucions es descriuen només amb un paràmetre d'escala i el coeficient de variació. La primera correspon a coeficients de variació menors que la unitat i la segona a coeficients superiors a la unitat. Conjuntament formen un model molt apropiat per descriure el comportament de les cues d'una distribució.

Les eines de l'anàlisi de la supervivència i la fiabilitat són d'utilitat en l'estudi de les cues d'una distribució. Permeten fer una aproximació no paramètrica del problema, que complementa l'estudi paramètric proposat anteriorment.

La nostra moderada experiència amb dades financeres, des de dades de tipus de canvi, fins a dades intradia per a futurs, ens porta a proposar models de cues semipesades, properes a les cues de la distribució exponencial.

10. Bibliografia

BARLOW, R. E.; PROSCHAN, F. (1981). *Statistical theory of reliability and life testing.* Silver Spring, MD.

BARNDORFF-NIELSEN, O. E. (1998). «Processes of normal inverse gaussian type». *Finance Stochastics*, vol. 2, p. 41-61.

BARNDORFF-NIELSEN, O. E.; SHEPHARD, N. (2001). «Non-gaussian Ornstein-Uhlenbeck based models and some of their uses in financial econometrics (with discussion)». *Journal of the Royal Statistical Society*, B, vol. 63, p. 167-241.

BLACK, F.; SCHOLES, M. (1973). «The pricing of options and corporate liabilities». *Journal of Political Economy*, vol. 81, p. 637-654.

BLATTBERG, R.; GONEDES, N. (1974). «A comparison of the stable and student distributions as statistical models for stock prices». *The Journal of Business*, vol. 47, p. 244-280.

CAMPBELL, J.; LO, A.; MACKINLAY, A. (1997). *The econometrics of financial markets*, Princeton: Princeton University Press.

CASTILLO, J. (1994). «The singly truncated normal distribution, a non-steep exponential family». *Annals of the Institute of Mathematical Statistics*, vol. 46, p. 57-66.

CASTILLO, J.; PUIG, P. (1999). «The best test of exponentiality against singly truncated normal alternatives». *Journal of the American Statistical Association*, vol. 94, núm. 446, p. 529-532.

DUFFIE, D.; PAN, J. (1997). «An overview of value at risk». *Journal of Derivatives*, vol. 4, núm. 3, p. 27-39.

EBERLEIN, E.; KELLER, U. (1995). «Hyperbolic distributions in finance». *Bernouilli*, vol. 1, p. 281-299.

EMBRECHTS, P.; KLÜPPELBERG, C.; MIKOSCH, T. (1997). «Modeling extremal events for insurance and finance». *Springer* [Berlín].

GREENE, W. H. (1993). *Econometric analysis.* Nova York: Macmillan Publishing Company.

GUILLAUME, D.; DACOROGNA, M.; DAVE, R.; MÜLLER, U.; OLSEN, R.; PICTET, O. (1997). «From the bird's eye to the microscope: A survey of new stylized facts of the intra-daily foreign exchange markets». *Finance & Stochastics*, vol. 1, núm. 2, p. 95-129

HECKMAN, J. (1979). «Sample selection bias as a specification error». *Econometrica*, vol. 47, p. 153-161.

HOSKING, J.; WALLIS, J. (1987). «Parameter and quantile estimation for the generalized Pareto distribution». *Technometrics*, vol. 29, núm. 3, p. 339-348.

HULL, J.; WHITE, A. (1998). «Value at risc when daily changes in marked variables are not normally distributed». *The Journal of Derivatives*, p. 9-19.

MADAN, D. B.; SENETA, E. (1990). «The variance gamma model for share marked returns». *Journal of Business*, vol. 63, p. 511-524.

MÜLLER, U.; DACOROGNA, M.; PICTET, O. (1998). «Heavy tails in high-frequency financial data». Adler-Felman-Taqqu ed. *A practical guide to heavy tails*. Birkhäuser.

MANDELBROT, B. (1965). «Forecast of future prices, unbiased markets, and martingale models». *Journal of Business*, p. 242-255.

TAYLOR, S. (1986). *Modelling financial time series*. Londres: John Wiley and Sons.

SHIRYAEV, A. (2000). *Essentials of stochastic finance: Facts, models, theory*. World Scientific Publishing.