USMP
UNIVERSIDAD DE SAN MARTÍN DE PORRES
ESCUELA DE NEGOCIOS

MODELAMIENTO DE LA SEVERIDAD DEL RIESGO OPERACIONAL POR LAS DISTRIBUCIONES $\mathcal{G}\,\&\,\mathcal{H}$

Por Eduardo Court M

Profesor de la Escuela de Post Grado de la USMP

1 Definición de las leyes $\mathcal{G}\&\mathcal{H}$

Propuestas por J. Tuckey, las leyes $\mathcal{G}\&\mathcal{H}$ fueron estudiadas por Hoaglin & Peters (1979), Martinez & Iglewics compararon las técnicas de estimación de estos parámetros, Dutta & Perry analizaron la aplicación del modelo a los riesgos operacionales. El hecho de que estas leyes sean una simple transformación de la ley normal centrada reducida, facilita los cálculos tales como el de la evaluación de los cuantiles y los de la simulación Monte Carlo.

En este trabajo, empezaremos por la comprensión de dos familias de estas leyes que son; las leyes $\mathcal{G}$, que permiten describir las variables continuas asimétricas y las leyes $\mathcal{H}$, que puede modelar las colas gruesas. Luego reuniremos estas dos leyes para formar las leyes $\mathcal{G}\&\mathcal{H}$, que nos permitirán describir estructuras de asimetría y de aplanamiento más complejas.

1.1 Asimetría y ley $\mathcal{G}$

Modelamos una variable aleatoria asimétrica X de ley $\mathcal{G}$ y de parámetros A, B y g, con la ayuda de una función monótona de una variable normal centrada reducida[1] Z. Los parámetros A y B tienen en cuenta, respectivamente, la localización y la escala de X. Denominemos $X \sim \mathcal{G}(A, B, g)$.

Escribimos:

$$X = A + B.Y \qquad (1.1)$$

Donde A y B son escalares e Y es una variable aleatoria de ley $\mathcal{G}(0,1,g)$. La media de Y es escogida igual a cero, de forma que A sea la mediana de X.

Escribamos ahora Y como función de $Z, Y = Y(Z)$. Una forma de proceder es introducir una función de modulación G que afecte los valores positivos de Z de manera diferente que los valores negativos.

$$Y = G(Z).Z \qquad (1.2)$$

El caso de $G(z) = 1$ corresponde a la ley normal centrada reducida. Con el fin de representar la simetría asumiremos G tal que $G(-\mathcal{Z}) \neq G(\mathcal{Z})$, para todo $\mathcal{Z} \neq 0$. Luego, para tener en cuenta de que el efecto de asimetría es cerca de la mediana, necesitamos exigir que $G(\mathcal{Z}) \approx 1$ en la cercanía de 0.

---

[1] En análisis de datos, centrar y reducir las variables (normalizar) permite comparaciones independientes de la unidad de medida: *Centrar* una variable consiste en restar su media a cada uno de sus valores iniciales; *Reducir* una variable consiste en dividir todos sus valores por su desviación típica. Una variable centrada reducida tiene: una media nula, y, una desviación típica igual a uno. Así obtenemos: datos independientes de la unidad, o de la escala escogida.

Una familia practica de funciones propuesta por Tuckey y que verifica estas propiedades es definida por:

$$Y_g(\mathcal{Z}) = \frac{e^{gz}-1}{gz}.\mathcal{Z} = \frac{e^{gz}-1}{g} \qquad (1.3)$$

Otros autores también han propuesto diferentes familias de funciones que verifican los enunciados mencionados. Esté capítulo se basa en las leyes $\mathcal{G}$ de Tuckey.

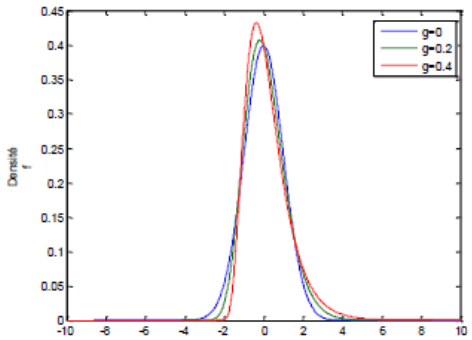
Definición: Una variable aleatoria continua X de ley $\mathcal{G}$ (de Tuckey), y de parámetros A, B y g denominada: $\mathcal{G}(A, B, g)$ se escribe:

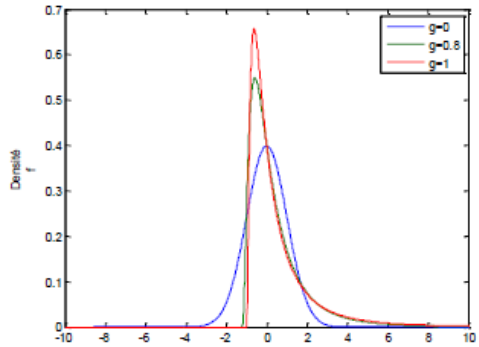
$$X = A + B.\frac{e^{gZ}-1}{g} \qquad (1.4)$$

donde Z es una variable aleatoria centrada reducida, A es la mediana, g el parámetro de asimetría, y B es un parámetro de escala. El caso $g \to 0$ corresponde a la ley normal.

1.1.1 Asimetría para diferentes valores de g

Con el fin de visualizar el efecto del parámetro g sobre la simetría trazamos las densidades de la ley $\mathcal{G}(0,1,g)$ para $g = 0, 0.2, 0.4, 0.8$ y 1





Observamos que los valores negativos de g producen una asimetría (negativa).

1.2 Aplanamiento y leyes $\mathcal{H}$

Con el fin de modelar las leyes de colas mas gruesas que aquellas de la ley normal, introducimos una transformación de esta última que otorga mas peso a los valores extremos.

Modelamos una variable aleatoria asimétrica X de ley $\mathcal{H}$ y de parámetros A, B y h, con la ayuda de una función con una variable normal centrada reducida Z.

Los parámetros A y B tienen en cuenta respectivamente, la localización y la escala de X. Observemos que $X \sim \mathcal{H}(A, B, h)$.

Escribimos:

$$X = A + B.Y \qquad (1.5)$$

De la misma forma que para las leyes $\mathcal{G}$ introducimos una función H que modela el aplanamiento. La elección de H se hace de manera que permita estirar las colas, preservando la simetría. Esto exige que H sea una función par y estrictamente positiva.

$$H(\mathcal{Z}) = e^{\frac{hz^2}{2}} \qquad (1.6)$$

Además, para que el aplanamiento opere, es necesario que H sea creciente para $Z^2 \geq -\frac{1}{h}$.

Una familia simple de funciones, propuesta por Tuckey, que tiene el comportamiento deseado está definida por:

$$Y_h(z) = z.H(z) = z.e^{\frac{hz^2}{2}} \qquad (1.7)$$

Definición: Una variable aleatoria continua X de ley (Tuckey), y de parámetros A, B y h, denominada $\mathcal{H}(A, B, h)$ se escribe:

$$X = A + B.Z.e^{\frac{xz^2}{2}} \qquad (1.8)$$

donde Z es una variable aleatoria normal centrada reducida, A corresponde a la mediana y B a un parámetro de escala; h controla la importancia y la dirección del aplanamiento.

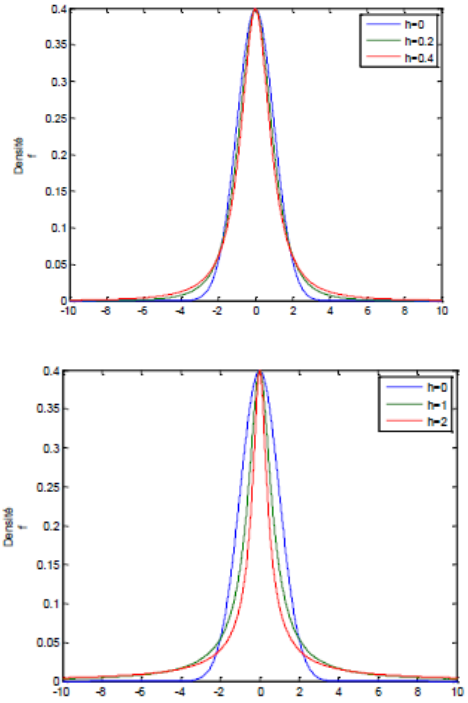
Observaciones:

- El caso $h = 0$ corresponde a una ley normal, y por lo tanto a una ausencia de aplanamiento.
- Un valor negativo de h es una dificultad numérica ya que $Y_h(z)$ no es monótona para $z^2 > -\frac{1}{h}$. Esto se puede observar generalmente en los

resultados luego de la simulación de muestras de la ley $\mathcal{H}$ y al calcular su función de densidad. De todas formas, los datos en riesgo operacional se caracterizan generalmente por colas gruesas, lo que corresponde a un h positivo.

1.2.1 aplanamiento por diversos valores del parámetro h.

Con el fin de visualizar el efecto del parámetro h sobre la asimetría trazaremos las densidades de ley $\mathcal{H}(0,1,h)$ para $h = 0, 0.2, 0.4, 1$ y 2.



1.3 Leyes $\mathcal{G}$ & $\mathcal{H}$

Para poder reunir las leyes $\mathcal{G}$ y las leyes $\mathcal{H}$, es necesario permitir a la distribución resultante de ser a la vez asimétrica y escalonada. En efecto, tratar simultáneamente estos dos aspectos además de la localización y de la escala nos permite mas flexibilidad.

Para combinar estos dos aspectos, usaremos de nuevo la multiplicación, escribiendo como hemos visto antes:

$$X = A + B.Y$$
$$Y(z) = z.G(z)H(z) \qquad (1.9)$$

Tomamos la misma elección de funciones particulares, introducidas por Tuckey, de las leyes $\mathcal{G}$ y las leyes $\mathcal{H}$ definidas líneas arriba, y escribimos:

$$Y(z) = \left(\frac{e^{gz} - 1}{g}\right) e^{\frac{hz^2}{2}} \qquad (1.10)$$

Por las razones explicadas antes, escogemos tratar solo los h positivos o nulos.

Definición: Una variable aleatoria continua X de ley $\mathcal{G}\&\mathcal{H}$ y de parámetros A, B y g denominada $\mathcal{G}\&\mathcal{H}(A, B, g, h)$ se escribe:

$$X = A + B.\left(\frac{e^{gZ} - 1}{g}\right)e^{\frac{hZ^2}{2}} \qquad (1.11)$$

donde Z es una variable aleatoria normal centrada reducida, A corresponde a la mediana, g al parámetro de asimetría, h al parámetro de aplanamiento y B es un parámetro de escala.

1.3.1 Algunas leyes obtenidas por diferentes combinaciones de A, B, g y h

Las leyes $\mathcal{G}\&\mathcal{H}$ por su flexibilidad ofrecen un enorme potencial de modelaje. De acuerdo con la parte anterior, estas leyes nos permiten aproximar varias distribuciones teóricas. En efecto, mas de doce leyes univariadas, entre ellas, Logo-student, Weibull, logo-normal, etc. Pueden ser aproximadas eligiendo de manera apropiada los parámetros g y h.

Ley logo normal:

La ley logo normal es un caso particular de una distribución $\mathcal{G}\&\mathcal{H}$, cuando g es constante y positivo. Así, es posible encontrar las relaciones entre los parámetros de las dos distribuciones.

Si X es una variable aleatoria de distribución $\mathcal{LN}(\mu, \sigma)$ y $Z \sim \mathcal{N}(0,1)$ entonces tenemos: $X = e^{\mu + \sigma Z}$

entonces:

$$X = e^{\mu} + \sigma e^{\mu}\left(\frac{e^{\sigma Z} - 1}{\sigma}\right) \qquad (1.12)$$

Por identificación con relación a una variable aleatoria T de ley $\mathcal{G}\&\mathcal{H}(A, B, g, 0)$

$$T = A + B\left(\frac{e^{\sigma Z} - 1}{\sigma}\right) \qquad (1.13)$$

deducimos que X es de ley $\mathcal{G}\&\mathcal{H}(e^{\mu}, \sigma e^{\mu}, \sigma, 0)$.

De la misma manera, obtenemos para algunas distribuciones de probabilidad los valores de los parámetros A, B, g y h de las leyes $\mathcal{G}\&\mathcal{H}$ que les corresponden.

Distribución	Parámetros	$\mathcal{G}\&\mathcal{H}$			
		A	B	g	h
Cauchy	$\mu, \sigma > 0$	μ	σ	0	1
Normal	μ, σ	μ	σ	0	0
Logo-normal	μ, σ	e^{μ}	σe^{μ}	σ	1

1.3.2 Análisis de flexibilidad: Skewness-Kurtosis (asimetría-aplanamiento)

Además de la caracterización por su número de parámetros, las leyes de probabilidad también se pueden caracterizar por la evaluación del grosor de sus colas y su asimetría. En el contexto de las distribuciones de severidad, la cola corresponde a la parte que se encuentra por encima de un cierto umbral “alto”. Una distribución es llamada de cola gruesa si la probabilidad de caer en una pérdida grande es elevada. Hay varias formas de definir una ley de cola gruesa.

Dutta y Perry usan una forma práctica para caracterizar una distribución, ellos se basan en un estudio de sus momentos de orden k. Los dos primeros representan la localización y la escala, el tercero mide la asimetría (skewness) de la ley, y el cuarto caracteriza el grosor de la cola o el aplastamiento (kurtosis) de la distribución.

Los dos últimos se definen respectivamente por:

$$S = \frac{\mu_3}{\mu_2^{3/2}}$$

y

$$K = \frac{\mu_4}{\mu_2^2}$$

donde:

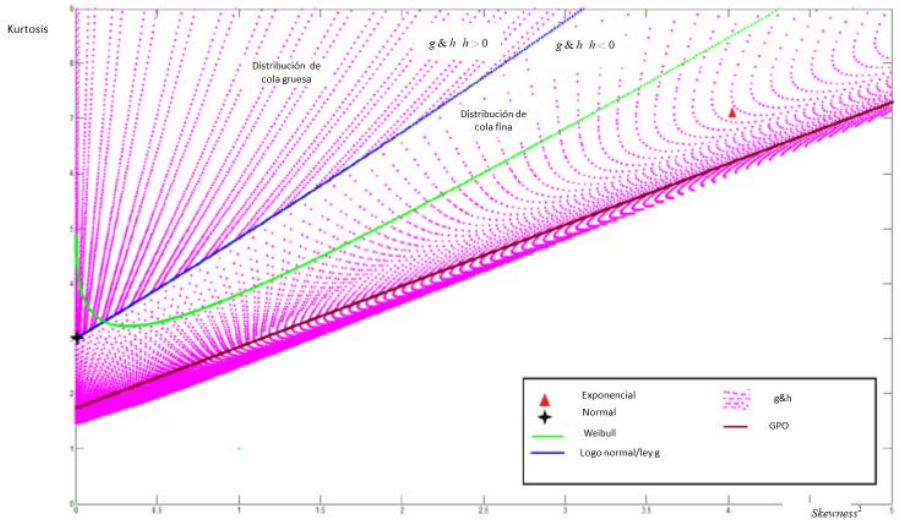
$$\mu = E\left(X\right) \text{ y } \mu_k = E\left\{\left(X-\mu\right)^k\right\} \qquad (1.14)$$

Si:

$K <$: Distribución platikurtica
$K = 3$: Distribución mesokurtica
$K > 3$: Distribución leptokurtica

Una forma interesante de ver la flexibilidad de una ley de probabilidad es la de dibujar una serie de puntos sobre un eje Skewness^2-Kurtosis. Esto nos mostrará la ubicación de los pares que la distribución puede representar para diferentes valores de estos parámetros. Cuanto mayor sea el área barrida por la ley, esta será mas flexible. Obtendremos entonces un punto, una curva o una superficie según si el skewness o la kurtosis son funciones de cero, o de varios parámetros.

En la figura siguiente, podemos ver claramente que las leyes $\mathcal{G}\,\&\,\mathcal{H}$ pueden escanear una superficie importante de la pareja Skewness^2-Kurtosis.



Ellas permiten también, por su parámetro de aplastamiento h separar las distribuciones de cola gruesa y aquellas de cola fina. h=0 se constituye en una frontera entre distribuciones de cola fina y cola gruesa.

1.3.3 Función de repartición

Sea X una variable aleatoria de distribución $\mathcal{G}\&\mathcal{H}(A,B,g,h)$. La función de repartición X se escribe:

$$F(t)=\Phi\left[Y^{-1}\left(\frac{t-A}{B}\right)\right], \qquad \forall t\in\mathbb{R} \qquad (1.15)$$

Donde Φ es la función de repartición de una $N(0,1)$ e Y^{-1} la recíproca de la función de Tuckey:

$$Y(z)=\left(\frac{e^{gz}-1}{g}\right)e^{hz^2/2}$$

En efecto:

$$\begin{aligned}F(t)&=P(X\le t)\\&=P(A+B.Y\le t)\\&=P\left(Y\le\frac{t-A}{B}\right)\\&=P\left(Z\le Y^{-1}\left(\frac{t-A}{B}\right)\right)\\&=\Phi\left[Y^{-1}\left(\frac{t-A}{B}\right)\right]\end{aligned}$$

1.3.4 Densidad

Por derivación de lo que obtuvimos anteriormente, obtenemos la densidad de X

$$f(t)=\frac{\varphi\left[Y^{-1}\left(\frac{t-A}{B}\right)\right]}{B.Y'\left[Y^{-1}\left(\frac{t-A}{B}\right)\right]},\qquad \forall t\in\mathbb{R} \tag{1.16}$$

Donde φ es la función de densidad de una $\mathcal{N}(0,1)$ e $Y'(z)=e^{gz+hz^2/2}+hz\left(\frac{e^{gz}-1}{g}\right)e^{hz^2/2}$

1.3.5 Función de repartición inversa (función cuantil)

Demostramos que, para un nivel de cuantil α, la función cuantil se obtiene aplicando la misma transformación que define una variable aleatoria de distribución $\mathcal{G}\&\mathcal{H}$, es decir:

Para $0<\alpha<1$ tenemos:

$$F^{-1}(\alpha)=A+B.\left[Y\left(\Phi^{-1}(\alpha)\right)\right] \tag{1.17}$$

1.3.6 Espesor de la cola

Con el fin de determinar para que valores de los parámetros una ley $\mathcal{G}\&\mathcal{H}$ es de cola gruesa o fina usaremos la siguiente propiedad:

Propiedad: si X es una variable aleatoria de distribución $\mathcal{G}\&\mathcal{H}(A,B,g,h)$ entonces la ley de X tiene una variación regular respecto al índice $\frac{1}{h}$.

Para simplificar desarrollamos la demostración para $A=0$ y $B=1$, i.e $X\sim\mathcal{G}\&\mathcal{H}(0,1,g,h)$.

$$\begin{aligned}\lim_{x\to\infty}\frac{xf(x)}{1-F(x)}&=\lim_{x\to\infty}\frac{x\varphi\left[Y^{-1}(x)\right]}{\left[1-\Phi\left[Y^{-1}(x)\right]\right].Y'\left[Y^{-1}(x)\right]}\\&=\lim_{x\to\infty}\frac{Y(u).\varphi(u)}{\left[1-\Phi(u)\right].Y'(u)}\quad ,u:=Y^{-1}(x)\\&=\lim_{x\to\infty}\frac{\left(e^{gu}-1\right)\varphi(u)}{\left[1-\Phi(u)\right].\left[ge^{g}+hu\left(e^{gu}-1\right)\right]}\\&=\frac{1}{h}\end{aligned} \tag{1.18}$$

La ley $\mathcal{G}\&\mathcal{H}$ es entonces de cola gruesa para h estrictamente positivo.

Observación: Las leyes $\mathcal{G}\,\&\,\mathcal{H}$ (por lo tanto, logo normal) que no son de variación regular y por lo tanto no son de cola gruesa según la distribución de Karamata tienen limitaciones para ajustar las distribuciones de riesgo operacional. La introducción de un parámetro adicional h permite modelar esta particularidad.

1.3.7 Momentos de orden n

El momento de orden n de una distribución $\mathcal{G}\,\&\,\mathcal{H}(A,B,g,h)$ para $g \neq 0$ y $0 \leq h \leq \frac{1}{n}$ está dado por:

$$E\left(X^{n}\right) = \sum_{i=0}^{n} \binom{n}{i} A^{n-i} B^{i} \frac{\sum_{r=0}^{i} (-1)^{-1} \binom{i}{r} e^{\frac{\{(i-r)g\}^2}{\{2(1-ih)\}}}}{g^{i}\sqrt{1-ih}} \qquad (1.19)$$

De donde deducimos la esperanza y la varianza:

$$E\left(X\right) = A + B.\left[\frac{1}{g\sqrt{1-h}}\left(e^{g^2/\{2(1-h)\}} - 1\right)\right] \text{ ,para } h < 1$$

$$\text{var}\left(x\right) = B^2.\left[\frac{1}{g^2\sqrt{1-2h}}\left(e^{2g^2/\{2(1-2h)\}} - 2e^{g^2/\{2(1-2h)\}} + 1\right) - \frac{1}{g^2(1-h)}\left(e^{g^2/\{2(1-h)\}} - 1\right)^2\right],$$

$$\text{para } h < \frac{1}{2} \qquad (1.20)$$

Teniendo en cuenta la complejidad de los momentos, los Skewness y las kurtosis no pueden tener formas interesantes. Aunque siempre es posible obtenerlas numéricamente gracias a la ecuación anterior.

1.4 Estimación de los parámetros

1.4.1 Teniendo en cuenta el umbral de recolección

Las pérdidas solo se recaudan a partir de un cierto umbral H, lo que afecta la estimación de los parámetros porque la distribución empírica es distinta de la verdadera distribución. Por lo tanto, tenemos que conectar la distribución real a su distribución empírica teniendo en cuenta la densidad condicional:

$$\widetilde{f}_{\theta|H}(t) = \frac{f_\theta(t)}{\int_H^{+\infty} f_\theta(u)\,du} \text{II}_{\{t \geq H\}} = \frac{f_\theta(t)}{1 - f_\theta(H)} \text{II}_{\{t \geq H\}}$$

Reemplazando la densidad y la función de repartición por sus valores, se obtiene:

$$\widetilde{f}_{\theta|H}(t) = \frac{\varphi\left[Y^{-1}\left(\frac{t-A}{B}\right)\right]}{B.Y'\left[Y^{-1}\left(\frac{t-A}{B}\right)\right] * \left[1 - \Phi\left[Y^{-1}\left(\frac{H-A}{B}\right)\right]\right]} \text{II}_{\{t \geq H\}} \qquad (1.21)$$

Con: $Y'(\theta) = B\left\{e^{gz+hz^2/2} + hz\left(\frac{e^{gz}-1}{g}\right)e^{hz^2/2}\right\}$

Deducimos la función de repartición y la función del cuantil:

$$\tilde{f}_{\theta|H}(t) = \frac{F_\theta(t) - F_\theta(H)}{1 - F_\theta(H)} \mathrm{II}_{\{t \geq H\}} \qquad (1.22)$$

Entonces:

$$\tilde{F}_{\theta|H}(t) = \frac{\Phi\left[Y^{-1}\left(\frac{t-A}{B}\right)\right] - \Phi\left[Y^{-1}\left(\frac{H-A}{B}\right)\right]}{1 - \Phi\left[Y^{-1}\left(\frac{H-A}{B}\right)\right]} \mathrm{II}_{\{t \geq H\}} \qquad (1.23)$$

y

$$\tilde{F}_{\theta|H}^{-1}(\alpha) = F^{-1}\left[\left[1 - \Phi\left[Y^{-1}\left(\frac{H-A}{B}\right)\right]\right] * \alpha + \Phi\left[Y^{-1}\left(\frac{H-A}{B}\right)\right]\right]$$

Entonces:

$$\tilde{F}_{\theta|H}^{-1}(\alpha) = F^{-1}\left[\alpha + (1-\alpha).\Phi\left[Y^{-1}\left(\frac{H-A}{B}\right)\right]\right] \qquad (1.24)$$

Con

$$F^{-1}(\alpha) = A + B.\left[Y\left(\Phi^{-1}(\alpha)\right)\right]$$

1.4.2 Método inter-cuantil (IQ)

Empecemos por describir un enfoque simple y práctico, presentado por J. Drouin, que nos permite estimar los parámetros A, B, g y h. Consideramos $X_{g,h}$ y Z como las variables aleatorias de las distribuciones $\mathcal{G}\&\mathcal{H}(A,B,g,h)$ y $\mathcal{N}(0,1)$. Denotamos como x_p y z_p a sus cuantiles de nivel p.

Tendremos:

$$x_p = A + B\left(\frac{e^{gz_p}-1}{g}\right)e^{hz_p{}^2/2} \text{(1)}$$

De donde deducimos por simetría de la ley normal:

$$x_{1-p} = A + B\left(\frac{e^{-gz_p}-1}{g}\right)e^{hz_p{}^2/2} \text{(2)}$$

(2) / (1) nos da, con $x_{0.5} = A$, la mediana de $X_{g,h}$:

$$\frac{x_{1-p}-x_{0.5}}{x_p-x_{0.5}}=\frac{x_{-p}-1}{e^{gz_p}-1}=-e^{-gz_p}$$

Entonces, para $0<p<0.5$ tenemos:

$$g_p=\frac{1}{z_p}\ln\left(\frac{x_{1-p}-x_{0.5}}{x_{0.5}-x_p}\right) \qquad (1.25)$$

En este resultado podemos ver claramente que g depende de p. Hoaglin sugiere elegir el parámetro g igual a la mediana de los g_p.

Ya tenemos definidos A y g, veamos ahora B y h.

(1)-(2) nos da:

$$x_p-x_{1-p}=B\left(\frac{e^{gz_p}-e^{-gz_p}}{g}\right)e^{hz_p{}^2/2}$$

$$\ln\left(\frac{g\left(x_p-x_{1-p}\right)}{e^{gz_p}-e^{-gz_p}}\right)=\ln\left(B\right)+\frac{hz_p{}^2}{2} \qquad (1.26)$$

Dado que las distribuciones de pérdidas en riesgo operacional son positivamente asimétricas $\left(g>0\right)$ y leptokúrticas, es mas apropiado explicar el término de la izquierda con la ayuda de un semi-spread superior (UHS), obtenido solo de la fórmula (2). Esto es sumamente interesante ya que nos focaliza sobre la cola derecha de la distribución:

$$UHS=\frac{g\left(x_{1-p}-x_{0.5}\right)}{e^{gz_p}-1} \qquad (1.27)$$

Los valores de B y h son estimados haciendo una regresión lineal de $\ln\left(UHS\right)$ en función de $\frac{z_p{}^2}{2}$. El estimador B es igual a la exponencial de la ordenada al origen y el de h es igual a la pendiente de la regresión.

Hay que tener en cuenta que este es un método usado en la práctica ya que se basa en un resultado gráfico y depende de varias hipótesis difíciles de verificar. Este procedimiento da resultados aberrantes dado que las muestras usadas no provienen de una distribución $\mathcal{G}\&\mathcal{H}$. Sin embargo, dejamos la elección de usar el resultado de este procedimiento como un punto de partida de los algoritmos de optimización de los métodos de optimización que hemos visto, teniendo en cuenta que son fórmulas cerradas que de todas formas traerán una influencia positiva en los programas de optimización.

1.4.3 El método de máxima verosimilitud (ML)

El método de máxima verosimilitud es el más conocido en la estimación paramétrica, es el que tiene las mejores propiedades teóricas.

Recordemos que si la función de verosimilitud $L(\theta;x)$ admite un único máximo al punto $\hat{\theta}(x)$, entonces la aplicación $x \to \hat{\theta}(x)$ es llamada de máxima verosimilitud y $\hat{\theta}(X)$ es el estimador de máxima verosimilitud de θ.

$$\hat{\theta} = \arg\max_{\theta} L(\theta; X). \qquad (1.28)$$

Es preferible maximizar el logaritmo de la verosimilitud:

$$\hat{\theta} = \arg\max_{\theta} \ln\left(L(\theta; X)\right).$$

La densidad de probabilidad de una variable aleatoria $X \sim \mathcal{GH}(A,B,g,h)$ teniendo en cuenta el umbral H:

$$X = A + B\left(\frac{e^{gz}-1}{g}\right)e^{hZ^2/2} (3)$$

$$f_{X|H}\left(x\middle|A,B,g,h\right) = f_{Z|H}\left(z\middle|A,B,g,h\right)\left|\frac{dz}{dx}\right|$$

$$= \frac{f_{Z|H}\left(z\middle|A,B,g,h\right)}{\left|\frac{dz}{dx}\right|}$$

$$(4) \qquad = \frac{\varphi\left[Y^{-1}\left(\frac{x-A}{B}\right)\right]}{B.Y'\left[Y^{-1}\left(\frac{x-A}{B}\right)\right] * \left[1-\Phi\left[Y^{-1}\left(\frac{H-A}{B}\right)\right]\right]} \mathrm{II}_{\{x\geq H\}} \qquad (1.29)$$

Esta densidad no da como resultado una forma analítica explícita en función de x y por consecuencia se debe de analizar numéricamente.

El cálculo consiste ante todo en evaluar la recíproca de La función Y en $\frac{x-A}{B}$, luego hay que sustituir las soluciones obtenidas en la ecuación (4). Estando dada la muestra independiente e idénticamente distribuida $x_1, x^{2,}..., x_n$, la verosimilitud de cifras bajo la ley $\mathcal{G}\,\&\,\mathcal{H}$ es:

$$L_X\left(A,B,g,h\middle|x\right) = \prod_{i=1}^{n} \frac{\varphi\left[Y^{-1}\left(\frac{x_i-A}{B}\right)\right]}{B.Y'\left[Y^{-1}\left(\frac{x_i-A}{B}\right)\right] * \left[1-\Phi\left[Y^{-1}\left(\frac{H-A}{B}\right)\right]\right]} \mathrm{II}_{\{x\geq H\}} \qquad (1.30)$$

Cuando $h>0$, la verosimilitud se puede maximizar usando procedimientos numéricos. Si usamos mathlab, la función *fminsearch* permite efectuar el cálculo con la ayuda del algoritmo de Nelder-Mead.

El parámetro A ya no puede ser estimado inmediatamente por la mediana de la muestra. La maximización se efectúa entonces con cuatro parámetros: AB, g y h. Sin embargo, el hecho de combinar esta optimización con la inversión mencionada líneas arriba convierte al proceso en complejo y lento.

Cuando h es negativa, x deja de ser una función monótona de z, y obtener este indicador se convierte en un proceso difícil. Igual en el caso de la ley logo normal ($h=0$) puede ponerse delicado de manejar porque los intervalos de confianza de h son centrados alrededor de cero.

Hemos implementado numéricamente el método de máxima verosimilitud. Pero la función Y de Tuckey que se usa no se puede invertir de manera analítica, esto provoca muchas dificultades ya que aun con una versión optimizada de esta inversión el método es muy lento para muestras medianamente grandes. Las complicaciones provienen del hecho de que el algoritmo de Nelder-Mead, de maximización de la verosimilitud que se usa invoca esta inversión (n veces para una muestra de tamaño n) en cada iteración. Además, tenga en cuenta el umbral de la recolección, donde la adición de un factor $1/\left(1-F_{\theta}\left(H\right)\right)$ a todas las contribuciones de las observaciones de la verosimilitud, nos deja un término que puede convertirse en muy grande para algunos valores de los parámetros, lo que nos complica considerablemente la optimización. El método de máxima verosimilitud no es práctico para ajustar una distribución $\mathcal{G}\,\&\,\mathcal{H}$ sobre muestras de mediano a gran tamaño, o para hacer estudios de tipo Monte Carlo que requieren de una gran cantidad de estimaciones.

1.4.4 El método de los momentos generalizados (GMM)

Este método consiste en determinar el vector de parámetros que minimice la distancia entre los momentos teóricos y los momentos empíricos. Las condiciones de momentos para el orden p son definidas por:

$$G_p\left(\theta\right)=\frac{1}{n}\sum_{k=1}^{n}\left(\left(x_k\right)^p-E\left(X^p\right)\right) \qquad (1.31)$$

Donde x_k es la k^{iesima} perdida, y $E\left(X^p\right)$ es el momento de orden p, este último depende de θ.

Martinez & Iglewicks obtuvieron un momento de orden p con relación al origen, para $h\le\frac{1}{p}$ y $g\ge 0$:

$$E\left(X^p\right)=\sum_{i=0}^{p}\binom{p}{i}A^{p-i}B^i\frac{\sum_{r=0}^{i}\left(-1\right)^{-r}\binom{i}{r}e^{\left\{(i-r)g\right\}^2/\left\{2(1-ih)\right\}}}{g^i\sqrt{1-ih}} \qquad (1.32)$$

El método GMM presenta varias propiedades interesantes, particularmente la de convergencia, que permite que los resultados no sean tan aberrantes como los del método de máxima verosimilitud.

La restricción $h \leq \frac{1}{p}$ da lugar a un orden de momento máximo $p \leq \frac{1}{h}$. Esta restricción es difícil de respetar en riesgo operacional, ya que en el trabajo de Dutta y Perry 2007, se encuentra que $h \in [0.1; 0.4]$ lo que corresponde a $p \in [2.5; 10]$, y ya sabemos que requerimos de por lo menos cuatro momentos para estimar nuestros parámetros; la fórmula (1.32) puede no ser válida en al menos el 15% de los casos.

Observemos experimentalmente que para $g \geq 1$, los momentos de orden superior o igual a 1 son altamente variables con relación a los momentos empíricos, esto convierte en impreciso el método, además, según Dutta & Perry 2007, en riesgo operacional $g \in [1.7; 2.3]$.

Este método es además sensible al umbral de recopilación H, lo que provocará que las restricciones mencionadas sean más difíciles de respetar.

Las limitaciones de la GMM lo convierten en un método inutilizable para estimar los parámetros de la distribución $\mathcal{G}\&\mathcal{H}$ aplicadas a riesgo operacional.

1.4.5 Método de la distancia-cuantil (QD)

La forma, -transformada de la ley normal, - de la ley $\mathcal{G}\&\mathcal{H}$ la hace adecuada para métodos de estimación que consisten en minimizar una distancia particular entre cuantiles empíricos y cuantiles teóricos.

La idea de base es la de escoger los parámetros del modelo que convierten en mínima esta distancia.

Para una muestra de n pérdidas $\xi_1, ..., \xi_n$, nos interesamos en la distancia cuadrática entre k cuantiles empíricos $\hat{q}(p_1), ..., \hat{q}(p_k)$ y k cuantiles teóricos $F_\theta^{-1}(p_i)$, donde F_θ^{-1} es la función de repartición inversa y θ es el vector de parámetros de la ley a ajustar. La distancia se escribe:

$$Q^2(\theta, p, \omega) = \sum_{i=1}^{k} \omega_i \left(\hat{q}(p_i) - F_\theta^{-1}(p_i) \right)^2 \qquad (1.33)$$

Donde $\left(\omega_i = \frac{1}{\hat{q}(p_i)^2} \right)_{i=1..k}$ es un vector de ponderación y $(p_i)_{i=1..k}$ es el vector de los niveles de cuantiles a ajustar, con $0 < p_1 < \cdots < p_k < 1$.

Los cuantiles empíricos $\hat{q}$ se construyen a partir del vector de pérdidas de la muestra, en función del vector $p = (p_i, ..., p_k)'$. El i^{iesimo} cuantil empírico

corresponde a la i^{esima} pérdida de la muestra clasificada $\xi_1^*,...,\xi_n^*$ si el número de $n * p_i$ es entero, y a una interpolación lineal entre las dos pérdidas mas cercanas en caso contrario.

El término de ponderación ω_i sirve para limitar la inestabilidad numérica relacionada a los términos de ajuste de los cuantiles extremos en la ecuación de la distancia.

El objetivo de la estimación es encontrar el parámetro $\theta = \hat{\theta}_{QD}$ minimizando la distancia $\theta^2(\theta, p, w)$, con el objetivo de tener una buena adecuación entre las pérdidas observadas (cuantiles empíricos) y las pérdidas estimadas por el modelo paramétrico (cuantiles teóricos).

En el marco de la ley $\mathcal{G}\&\mathcal{H}$, y el vector de parámetros a ajustar es por lo tanto $\theta = (A, B, g, h)$, y la función cuantil es:

$$F^{-1}(\alpha) = A + B.\left[Y\left(\Phi^{-1}(\alpha)\right)\right] \qquad (1.34)$$

con

$$Y(z) = \left(\frac{e^{gz} - 1}{g}\right) e^{hz^2/2}$$

<u>Integración del umbral</u>

El tomar en cuenta el umbral de la muestra H, se traduce en una transformación de los niveles de los cuantiles a ajustar. En efecto, el cuantil de nivel α de la muestra trunca corresponde a un nivel de cuantil α^H de la distribución teórica completa.

Nosotros tenemos:

$$\widetilde{F}_{\theta|H}{}^{-1}(\alpha) = F^{-1}\left[\alpha + (1-\alpha).\Phi\left[Y^{-1}\left(\frac{H - A}{B}\right)\right]\right]$$

Deducimos la relación entre α^H y α:

$$\alpha^H = \alpha + (1-\alpha).F_\theta(H) \qquad (1.35)$$

En consecuencia, modificamos los cuantiles teóricos a ajustar, lo que implica que la ecuación de la distancia a minimizar es:

$$Q^2(\theta, p, \omega) = \sum_{i=1}^{k} \frac{1}{\hat{q}(p_i)^2}\left(\hat{q}(p_i) - F_\theta^{-1}\left[p_i + (1-p_i)F_\theta(H)\right]\right)^2 \quad (1.36)$$

Con: $F_\theta(H) = \Phi\left[Y^{-1}\left(\frac{H - A}{B}\right)\right]$

Sin embargo, esta fórmula necesita el conocimiento del parámetro $\theta(A, B.g.h)$, con el fin de evaluar el nivel del cuantil desfasado α^H. Esto hace que la optimización sea imposible. Para resolver este problema V. Leherisse & A. Renaudin usan un estimador $\tilde{\theta}$ de θ en el término corrector $F_\theta(H)$.

El estimador $\tilde{\theta}$ se obtiene por minimización de una distancia-cuantil teniendo en cuenta el umbral de recolección al desplazar linealmente los cuantiles empíricos hacia el origen:

$$\hat{q}^H(p_i) = \hat{q}(p_i) + (1 - p_i)H$$

Esta modificación permite usar los cuantiles teóricos de la distribución $\mathcal{G}\&\mathcal{H}$ no truncada y por consiguiente evitar todo problema numérico en la minimización de la distancia:

$$\tilde{\theta} = \arg\min_\theta \sum_{i=1}^{k} \frac{1}{\hat{q}(p_i)^2} \left(\left(\hat{q}(p_i) - (1 - p_i).H \right) - F_\theta^{-1}(p_i) \right)^2 \qquad (1.37)$$

Una vez que se ha determinado de esta forma $\tilde{\theta}$, el estimador final $\hat{\theta}_{QD}$ del parámetro de la distribución de severidad se obtiene minimizando la siguiente distancia:

$$\hat{\theta}_{QD} = \arg\min_\theta \sum_{i=1}^{k} \frac{1}{\hat{q}(p_i)^2} \left(\hat{q}(p_i) - F_\theta^{-1}\left(p_i + (1 - p_i) F_{\tilde{\theta}}(H) \right) \right)^2 \qquad (1.38)$$

El método de la distancia-cuantil, aplicado a la ley logo normal nos da resultados interesantes en relación a otros métodos de estimación. Luego vamos a comparar los diferentes métodos de estimación ya descritos, incluso aquellos que hemos descartado, con el fin de juzgar su precisión, estabilidad y tiempo de cálculo de forma cuantitativa.

Si quisiéramos ir más lejos, según Dutta & Perry 2007, en casos raros, el hecho de tomar g y h constantes, no permite tener un buen ajuste. En esos casos, las g y h pueden ser tomadas como funciones polinomiales de z_p^2. Una forma mas general de la ley $\mathcal{G}\&\mathcal{H}$:

$$X_{g.h} = A + B\left(\frac{e^{gz \cdot Z} - 1}{g_z} \right) e^{h_z \cdot (Z^2/2)} \qquad (1.39)$$

Con g_Z y h_Z como las funciones polinomiales en Z^2.

Por ejemplo:

$$g_Z = \gamma_0 + \gamma_1 Z^2 + \gamma_2 Z^4$$
$$h_Z = \eta_0 + \eta_1 Z^2 + \eta_2 Z^4 + \eta_3 Z^6$$

Ya quedó claramente demostrado que solo la generalización del parámetro h es interesante para tener mejores ajustes y que a partir de cuatro coeficientes para h no se gana nada en adecuación.

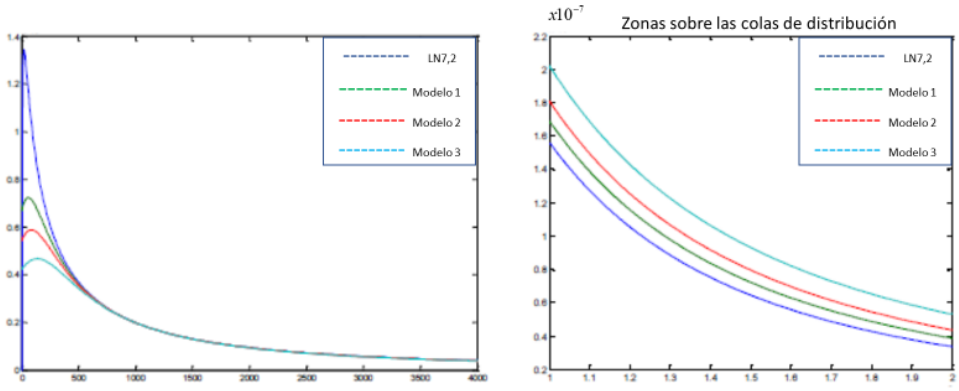
1.4.6 Propiedades teóricas de los métodos de estimación

1.4.6.1 Sobre cifras no sesgadas

En una primera etapa, nosotros comparamos los estimadores IQ, QD, ML y GMM en un ambiente simulado. Simulamos las muestras según una ley $\mathcal{G}\&\mathcal{H}(A,B,g,h)$ para diferentes valores de $\theta=(A,B,g,h)$. Las estimaciones de los parámetros se promedian sobre $N=1{,}000$ simulaciones con el objetivo de presentar las figuras de los casos más comunes en riesgo operacional. Elegiremos (A,B,g) a partir de parámetros (μ,σ), de la ley $\mathcal{LN}(\mu,\sigma)$, comúnmente observados y tomaremos un parámetro diferente h para cada modelo con el fin de representar el efecto de cola gruesa que caracteriza a esta categoría de riesgos. Usaremos los juegos de parámetros de abajo y cada muestra está constituida de $n=5000$ pérdidas:

modelo	A	B	g	h
1	Exp(7)	2*exp(7)	2	0.05
2	Exp(7)	2*exp(7)	2	0.1
3	Exp(7)	2*exp(7)	2	0.2

Veamos también el gráfico de densidad de diferentes modelos



Claramente podemos notar cuando observamos las colas de distribución que las pérdidas del orden de 10^5 son de más en más probables cuando vamos del modelo logo normal al modelo 3 (h=0.2).

Sesgo relativo

El sesgo del estimador $\hat{\theta}_n$ de θ es definido por:

$$\mathbb{E}\left(\hat{\theta}_n\right)-\theta$$

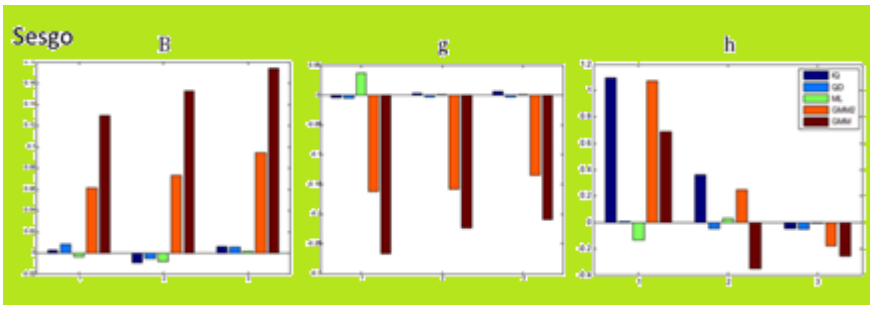
Esta esperanza será calculada promediando los estimadores $\hat{\theta}_n^1,...,\hat{\theta}_n^N$ obtenidos sobre N estimaciones distintas:

$$b\left(\hat{\theta}_n\right)=\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_n^i-\theta \qquad (1.40)$$

De esta forma se determinan los valores del sesgo para cada uno de los métodos y para los tres parámetros B, g y h, ya que el parámetro A es estimado por la mediana de la muestra y por consiguiente es la misma sin importar cuál es el método.

Llamaremos al método GMM[2] a aquel método de momentos generalizados que usa solo los dos primeros momentos. Es decir, que, para estimar nuestros tres parámetros, solo usamos la media y la varianza.

Las figuras de abajo representan los sesgos relativos $b\left(\hat{\theta}_n\right)/\theta$ para cada uno de los parámetros B, g y h.



La existencia de restricciones difíciles de respetar en los momentos teóricos hace que los momentos solo sean útiles para algunos valores de los parámetros, esto dificulta la maximización porque durante la optimización, se comparan los valores “falsos” de la función objetivo.

Los resultados de la gráfica lo muestran, de hecho, los sesgos son mayores usando los tres primeros momentos (GMM) que usando solo los dos primeros (GMM[2]). Esto prueba que los momentos de orden superior a dos no se pueden usar, y, por consiguiente, el método será menos preciso para los datos truncos (que implican la necesidad de un cuarto momento).

Los sesgos de los métodos ML y QD los encontramos relativamente débiles, así como el comportamiento estable de este último, porque no se ve afectado por la falta de datos en la cola de distribución (modelo 1), a diferencia del LD que muestra un mayor sesgo para la misma distribución.

Precisión

En esta parte presentamos dos indicadores para evaluar la precisión de las estimaciones obtenidas por diferentes métodos. El primero determinado por

cada una de las estimaciones obtenidas para cada uno de los tres modelos simulados, es la raíz cuadrada del error cuadrático medio relativo (relative root square error, R-RMSE):

$$R-RMSE\left(\hat{\theta}_n\right)=\frac{1}{\theta}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_n^i-\theta\right)^2} \qquad (1.41)$$

El segundo indicador retoma la idea del primero usando una escala logarítmica:

$$L-RMSE\left(\hat{\theta}_n\right)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\ln\frac{\hat{\theta}_n^i}{\theta}\right)^2} \qquad (1.42)$$

Esta escala logarítmica da mas peso a los errores de sub estimación que a los errores de sobre estimación (contrariamente al primer estimador cuyos pesos son simétricos).

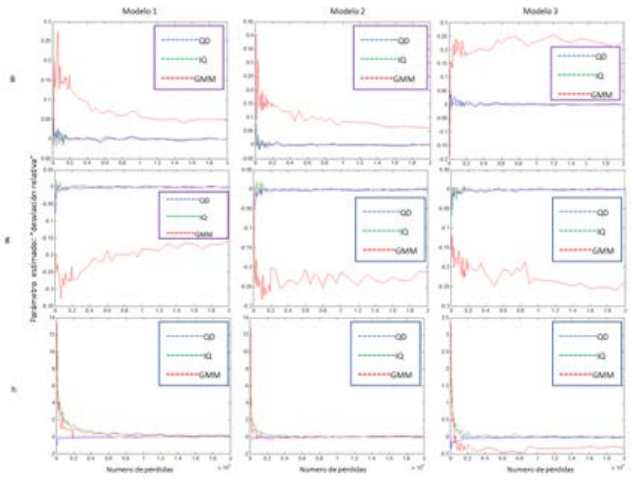
Los resultados de estos estimadores nos llevan a privilegiar el método QD que muestra una mejor precisión que los otros métodos.

Convergencia

Veamos ahora el comportamiento de las estimaciones, para cada uno de los métodos expuestos anteriormente, en función del número de pérdidas de la muestra.

Para hacer esto debemos de construir muestras que tengan un número de pérdidas que vayan de 50 a 20,000. Luego estimamos los parámetros de esas muestras con los diferentes métodos comparados.

Y obtenemos los gráficos de mas abajo para los diferentes modelos: (no pudimos comparar el método ML por su complejidad temporal).



Los resultados gráficos validan lo visto antes. En efecto, el método GMM converge muy lentamente para los parámetros B y g, y no converge del todo para el parámetro h, lo que legitima que rechacemos el GMM como método de

estimación para las leyes $\mathcal{G}\&\mathcal{H}$. Hay que hacer notar que, para las muestras de talla pequeña, solo el QD se comporta de manera aceptable.

El método de momentos generalizados es muy rápido, ya que no hace intervenir en cada iteración la minimización de la distancia entre momentos empíricos y momentos teóricos; el cálculo de estos últimos se hace usando una fórmula cerrada. Sin embargo, los resultados de la convergencia nos permiten confirmar que este método es inestable.

Los métodos de cuantiles nos dan resultados interesantes, ya sea en términos de sesgo, de precisión o de convergencia. Además, estos métodos usan la función cuantil de la distribución $\mathcal{G}\&\mathcal{H}$, que no es otra cosa que la transformación de la función cuantil de la ley normal centrada reducida. Están por lo tanto bien adaptados a la estimación de los parámetros de la ley $\mathcal{G}\&\mathcal{H}$.

En lo que sigue, solo nos quedaremos con el método QD, con el cual desarrollaremos mas en profundidad este tema. Para efectos de comparación usaremos el método ML, a pesar de su complejidad temporal.

1.4.6.2 Con datos truncos

En esta parte vamos a comparar los comportamientos de los métodos ML y QD, con datos truncos a la izquierda al 20%. Mantenemos los mismos modelos que en la parte anterior. Por otro lado, reducimos el número de pérdidas por muestra a 1000 (800 después de los truncos) para acercarnos a cifras reales. Esta vez, promediamos en $N=100$ simulaciones con el objetivo de poder obtener resultados por el método de máxima verosimilitud que es extremadamente lento. Empecemos por introducir algunas precisiones sobre la forma como la data numérica va a ser usada.

Máxima verosimilitud

La función a maximizar es:

$$L_X=\left(A,B,g,h\middle|x\right)=\prod_{i=1}^{n}\frac{\varphi\left[Y^{-1}\left(\frac{x_i-A}{B}\right)\right]}{B.Y'\left[Y^{-1}\left(\frac{x_i-A}{B}\right)\right]\cdot\left[1-\Phi\left[Y^{-1}\left(\frac{H-A}{B}\right)\right]\right]}\mathrm{II}_{\{x\geq H\}} \quad (1.43)$$

Es mejor minimizar:

$$L=\log\left(L_X\left(A.B,g,h\middle|x\right)\right)$$

$$L=\sum_{i=1}^{n}\left[\log\left(\varphi\left[Y^{-1}\left(\frac{x_i-A}{B}\right)\right]\right)-\log\left(B.Y'\left[Y^{-1}\left(\frac{x_i-A}{B}\right)\right]\right)-\log\left(1-\Phi\left[Y^{-1}\left(\frac{H-A}{B}\right)\right]\right)\right] \quad (1.44)$$

La inversa Y^{-1} debe ser calculada n veces a cada iteración, les sugerimos apuntar en una hoja los métodos numéricos usados para efectuar esta operación. El algoritmo que elegimos para hacer esta optimización es el “trust-

region-modified-Dogleg Algorithm (TRMDA)". Sin embargo, este algoritmo no garantiza que el mínimo encontrado sea un mínimo global. Siempre se puede combinar este algoritmo a un método de recolección simulado con el fin de aumentar las posibilidades de encontrar un mínimo global.

Distancia-cuantil

Este procedimiento se hace en dos etapas:

- Con la ayuda del algoritmo "simplex" de Nelder-Mead hacemos una primera aproximación de θ, aproximando el valor de la función de repartición al punto H (umbral de recolección) a través de $\left(\hat{q}(p_i)+(1-p_i)H\right)$ igual como se mostró líneas arriba.
- Usamos el resultado obtenido en el punto anterior para aproximar $F_\theta(H)$ y minimizamos, siempre con el mismo algoritmo, la distancia:

$$\hat{\theta}_{QD} \arg\min_{\theta} \sum_{i=1}^{k} \frac{1}{\hat{q}(p_i)^2}\left(\hat{q}(p_i) - F_\theta^{-1}\left(p_i + (1-p_i)F_{\tilde{\theta}}(H)\right)\right)^2 \qquad (1.45)$$

Los resultados nos confirmaran lo que venimos afirmando a propósito del método QD en el marco de datos no sesgados, de hecho, es más estable y menos sesgado que el método ML.

Como conclusiones podemos mencionar que es claramente visible la calidad de ajuste que nos dan las distribuciones $\mathcal{G}\&\mathcal{H}$ por la flexibilidad que tienen; sin embargo, para obtener resultados más sólidos se requiere una mayor cantidad de datos ya que con pocos datos en la cola, el parámetro que mide el aplanamiento ya no es confiable y puede generar una carga de capital aleatoriamente explosiva.

Hemos podido ver la divergencia de algunos métodos de estimación y la inestabilidad de otros en un contexto de cifras reales muy específicas que presentaban malas propiedades estadísticas, el método de estimación de los parámetros de severidad usado debe ser, ante todo, robusto. El método de la distancia entre cuantiles estudiado presenta buenos resultados y parece adaptarse bien a la problemática de estimación de la severidad del riesgo operacional y es fácil de poner en práctica para distribuciones transformadas de leyes como la ley $\mathcal{G}\&\mathcal{H}$.

La ley $\mathcal{G}\&\mathcal{H}$ se puede extender al análisis de otros riesgos. Basta revisar la literatura financiera para hallar existe mucha data con estructuras de asimetría y de aplanamiento complejos; por ejemplo, los rendimientos de bolsa. Podemos modelar periodos largos de comportamientos de un índice. Otro uso importante de la ley $\mathcal{G}\&\mathcal{H}$ es el modelamiento del comportamiento de las tasas de interés a corto plazo, que puede ser aplicado para la obtención de una fórmula explicita para evaluar opciones europeas.

1.5 Algoritmos y funciones

1.5.1 Recíproca de la función Y de Tuckey

La mayoría de resultados del modelaje por las leyes $\mathcal{G}\,\&\,\mathcal{H}$ se componen de la recíproca Y^{-1} de la función de Tuckey que se muestra:

$$Y(z)=\left(\frac{e^{gz}-1}{g}\right)e^{hz^2/2} \qquad (1.46)$$

Esta función no admite reciprocidad en forma analítica, para este libro, hemos probado varios métodos numéricos con el fin de elegir la forma mas optima en tiempo de cálculo y de precisión para calcularla.

1.5.1.1 Presentación de los algoritmos

Hagamos una pequeña aproximación a los algoritmos usados para aprovechar la recíproca de Y. Para mayores detalles pueden consultar la ayuda de Mathlab. El objetivo de la operación es encontrar z que minimice la función objetivo:

$$f(z)=\left|Y-\left(\frac{e^{gz}-1}{g}\right)e^{hz^2/2}\right| \qquad (1.47)$$

Evidentemente lo que estamos buscando es que este mínimo se acerque lo más posible a cero.

a) Simplex de Nelder – Mead (NMA)

Principio:

El método de Nelder – Mead es un algoritmo de optimización no lineal que fue publicado en 1965. Es un método numérico heurístico que busca minimizar una función continua en un espacio de varias dimensiones. El algoritmo explota el concepto de simplex[2] que es un politopo[3] de $N+1$ vértices en un espacio con N dimensiones. Inicialmente, a partir de dicho simplex sufre transformaciones simples durante las iteraciones: el se deforma, se desplaza y se reduce progresivamente hasta que sus vértices se acerquen aun punto donde la función es localmente mínima.

Algoritmo

Sea N la dimensión del espacio donde la función objetivo f toma sus valores. El algoritmo comienza con la definición de un simplex no degenerado escogido en este espacio. Por iteraciones sucesivas, el proceso consiste en determinar el punto del simplex donde la función es máxima con el fin de sustituirla por el reflejo de este punto con relación al centro de gravedad de los N puntos restantes. Si el valor de la función en este nuevo punto es menor a los otros valores tomados en otros

[2] Un símplex es la envoltura convexa de un conjunto de (*n* + 1) puntos independientes afines en un espacio euclídeo de dimensión *n* o mayor, es decir, el conjunto de puntos tal que ningún *m*-plano contiene más que (*m* + 1) de ellos. Se dice de estos puntos que están en posición general.

[3] En geometría politopo significa, la generalización a cualquier dimensión de un polígono bidimensional, o un poliedro tridimensional.

puntos, el simplex se estira en esa dirección. De lo contrario, se supone que el ritmo local de la función es un valle, y el simplex se reduce mediante una similitud centrada en el punto simplex donde la función es mínima.
Mas exactamente:
1.- Elección de $N+1$ puntos del espacio a N dimensiones desconocidas, formando un simplex: $\{x_1, x_2, ..., x_{N+1}\}$,
2.- Calculo de los valores de la función f en esos puntos, re indexación de los puntos de forma de tener $f(x_1) \leq f(x_2) \leq ... \leq f(x_{N+1})$. Es suficiente conocer el primero y los dos últimos.
3.- Calculo de x_0, centro de gravedad de todos los puntos excepto x_{N+1}
4.- Cálculo de $x_r = x_0 + (x_0 - x_{N+1})$ (reflexión de x_{N+1} con relación a x_0)
5.- si $f(x_r) < f(x_N)$, cálculo de $x_e = x_0 + 2(x_0 - x_{N+1})$ (estiramiento del simplex). Si $f(x_e) < f(x_r)$ reemplazo de x_{N+1} por x_e, sino, reemplazo de x_{N+1} por x_r. Regreso a la etapa 2.
6.- si $f(x_N) < f(x_r)$, cálculo de $x_c = x_{N+1} + \frac{1}{2}(x_0 - x_{N+1})$ (contracción del simplex). Si, $f(x_c) < f(x_N)$ reemplazo de x_{N+1} por x_c y retorno a la etapa 2, si no, ir a la etapa 7.
7.- Similitud de reporte de $\frac{1}{2}$ y de centro x_1: reemplazo de x_i por $x_1 + \frac{1}{2}(x_i - x_1)$ para $i \geq 2$. Retorno a la etapa 2.

Ventajas

- Generalidad: una función continua (sin evaluar sus derivadas)
- Eficiencia para una función no diferenciable.
- La interpretación geométrica subyacente
- La seguridad de obtener una serie decreciente de valores.

Desventajas:

- Se aplica mal o con dificultad cuando el dominio de definición de la función es complejo o cuando el mínimo buscado se encuentra en un vecindario del límite.
- Los datos arbitrarios de un simplex inicial.
- Una degradación de rendimientos cuando la dimensión N aumenta.
- El riesgo de que los simplex obtenidos sucesivamente tengan tendencia a degenerar (a pesar que la experiencia muestra que es difícil que esto ocurra.

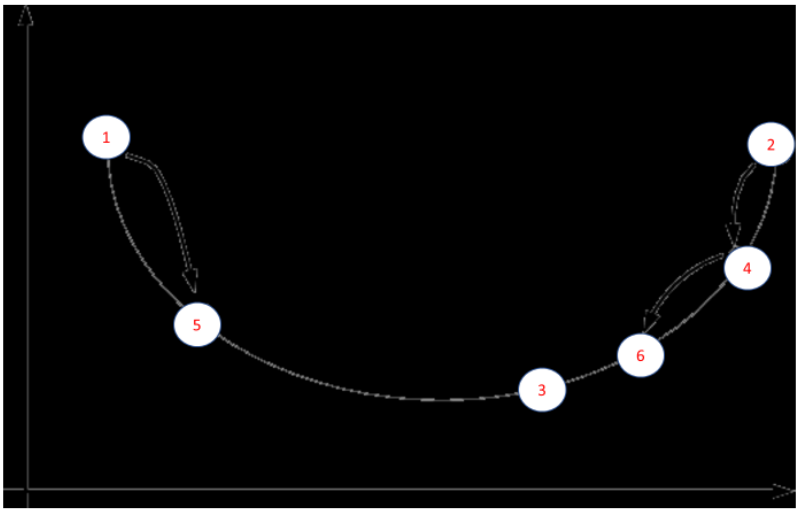
El algoritmo se usará con la función *fminsearch* de mathlab.

b) Búsqueda por la sección de oro, Golden search algorithm (GSA)

Principio

Para obtener un valor aproximado de una raíz de una función es suficiente con enmarcarlo en un intervalo $[a,b]$. Esto es insuficiente para caracterizar un mínimo. En este caso requerimos tres puntos (a,b,c). Un intervalo tal que $a<b<c$ y, $f(b)<f(a), f(b)<f_c$ caracteriza un mínimo en el intervalo $[a,c]$. En otras palabras, para encuadrar un mínimo necesitamos un punto triple tal que el punto central presente un valor de f inferior a aquellos de los bornes del intervalo.

El principio de esta búsqueda se muestra en la siguiente figura.



El mínimo se encuadra al origen por 1,3,2. La función es evaluada en 4, que reemplaza a 2. Enseguida en 5 que reemplaza a 1, luego en 6 que reemplaza a 4. La regla es guardar un punto central para el cual el valor de f es menor al de los bornes. Luego de esta serie de iteraciones, el mínimo estará encuadrado en 5,3,6. El principio es análogo al usado en la búsqueda de una raíz por dicotomía.

El único punto delicado, es el definir un método para elegir un nuevo punto de encuadre en el intervalo (a,b,c) inicial. Supongamos que b sea una fracción w del segmento $[a,c]$, entonces

$$\frac{b-a}{c-a}=w \Leftrightarrow \frac{c-b}{c-a}=1-w$$

Si elegimos un nuevo punto x alejado de una fracción z con relación a b

$$\frac{x-b}{c-a}=z$$

Mostramos simplemente que el nuevo punto x es el simétrico de b en su intervalo de origen, es decir: $|b-a|=|x-c|$. Esto significa que el punto x en el más grande de los segmentos $[a,b]$ y $[b,c]$. Aun falta definir la posición del punto x al interior de ese segmento. El valor w viene de una

etapa anterior de cálculo, y si suponemos que esta es óptima, entonces z debe ser elegida de la misma forma. Esta similitud de escala implica que x debe de estar situada en la misma fracción del segmento $[b,c]$ (si $[b,c]$ es el segmento más largo) que b con respecto al segmento $[a,c]$. Esto nos conduce a la relación:

$$\frac{z}{1-w}=w$$

Si combinamos esta ecuación con la definición de z llegamos a la ecuación cuadrática:

$$w^2-3w \Leftrightarrow w=\frac{3-\sqrt{5}}{2}\approx 0.38197$$

Este resultado significa que el intervalo de encuadramiento (a,b,c) en su punto central b está situado a una distancia tal que:

$$\frac{b-a}{c-a}=w\approx 0.38197$$

o

$$\frac{c-b}{c-a}=1-w\approx 0.61803$$

Estas fracciones son las de la media dorada o la sección dorada que se supone tiene ciertas propiedades estéticas de acuerdo a los pitagóricos del mundo antiguo. De ahí debe su nombre este método.

El algoritmo *fminbnd* describe una implementación optimizada de este método. Y como el valor buscado se supone que representa un logro de una variable aleatoria normal centrada reducida que tiene una probabilidad de $\left(1-9.8*e^{-10}\right)$, de estar en el intervalo $[-6,6]$, no es necesario buscar valores fuera de ese intervalo.

Ventajas

- Generalidad: una función continua (sin evaluar sus derivadas)
- Simplicidad para ejecutarlo
- Eficiencia para una función no derivable
- La seguridad de obtener una serie decreciente de valores
- Buen comportamiento cuando la dimensión N aumenta

Desventajas

- Se aplica mal o con dificultad cuando el dominio de la función es complejo o el mínimo buscado se encuentra en un vecindario del límite.
- Convergencia lenta cuando la solución se encuentra cercana a la frontera.

c) Región de confianza, Trust-Dogleg Algorithm (TRDA)

La función objetivo f se remplaza por una función de modelo cuadrático m_k en una cierta región alrededor de un punto x_k dado. Los dos primeros

términos involucrados en la función del modelo m_k de cada iterado x_k son idénticos a los dos primeros términos de la serie de Taylor de f alrededor de x_k. Entonces tendremos:

$$m_k(p) = f_k + \nabla f_k^T + \frac{1}{2} p^T B_k p \qquad (1.48)$$

donde $f_k = f(x_k)$, ∇f_k es la gradiente de f al punto x_k y B_k es una matriz simétrica que representa la hessiana[4] de f al iterado x_k o una aproximación a este último. Las regiones de confianza agregan una restricción en la longitud del paso al problema de optimización inicial. Lo que buscamos es la solución de cada sub problema de la forma:

$\min_{p \in \mathbb{R}^n} m_k(p)$ bajo la restricción $\|p\| \leq \Delta_k \quad (*)$

donde $\Delta_k > 0$ es el radio de la región de confianza.

Vista previa del algoritmo

Para un punto x_k y un radio Δ_k dados, determinamos la eficacia de la función modelada por la relación siguiente:

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p)} \qquad (1.49)$$

Esta relación se usa como criterio de actualización del radio Δ_k de la región de confianza. Observemos que el denominador es necesariamente positivo, ya que p_k es la solución de $(*)$, lo que significa que m_k decrece. Si $\rho_k < 0, f(x_k + p_k) > f(x_k)$, el paso debe ser rechazado. Si ρ_k es positivo, aceptamos el paso y el radio de la región de confianza se actualiza según los valores tomados por ρ_k. Si ρ_k es cercano a 1, las funciones modeladas y los objetivos están de acuerdo en este paso. Es mas interesante aun, aumentar el radio de la región de confianza para la siguiente iteración con el fin de obtener pasos mas consistentes. Para terminar, si ρ_k es cercano a 0, m_k no representa correctamente a f. Entonces debemos disminuir el radio de confianza.

<u>El punto de Cauchy</u>[5]

De manera similar a los métodos de investigación por líneas, la determinación de los pasos óptimos no es una condición necesaria para obtener una convergencia global. Aunque en principio, uno busca una solución óptima del sub problema $(*)$, es suficiente encontrar una solución próxima a p_k dentro de la región de confianza, que produzca una

[4] La **matriz hessiana** o **hessiano** de una función *f* de n variables, es la matriz cuadrada de *n* × *n*, de las segundas derivadas parciales.

[5] Una **sucesión de Cauchy** es una sucesión tal que, para cualquier distancia dada, por muy pequeña que sea, siempre se puede encontrar un término de la sucesión tal que la distancia entre dos términos cualesquiera posteriores es menor que la dada. Es importante no confundirlo con las sucesiones en las que la distancia entre dos términos consecutivos es cada vez menor, pues estas no son convergentes necesariamente.

“reducción suficiente” de la función modelada. Esta reducción se puede obtener por el método del punto de Cauchy p_k^c:

Algoritmo

1.- Encontrar un vector p_s^k, solución del problema $(*)$ alineado, sea

$p_s^k = \arg\min_{p\in\mathbb{R}^n}\left(f_k + \nabla f_k^T p\right)$ bajo la restricción $\|P\| \le \Delta_k$

2.- Calcular la escalar $\tau_k > 0$ que minimice $m_k\left(\tau p_k^s\right)$, sea

$\tau_k = \arg\min_{\tau>0} m_k\left(\tau p_k^s\right)$ bajo la restricción $\left\|\tau p_k^s\right\| \le \Delta_k$

3.- Finalmente, obtenemos el punto de Cauchy, sea

$p_k^s = \tau_k p_k^s$

Mejoramiento del punto de Cauchy por el método Dogleg

A pesar que el punto de Cauchy p_k^s proporciona una reducción suficiente de la función modelada m_k para producir una convergencia global y que el costo del calculo es pequeño, es interesante buscar una mejor solución de aproximación de $(*)$. En efecto, el punto de Cauchy se define como el punto que minimiza m_k a lo largo de la pendiente más grande $-\nabla f_k$. Es simplemente la implementación del método de descenso mas fuerte con una elección particular de la longitud de cada paso. Consideramos tres métodos para encontrar una solución aproximada a $(*)$. En este capítulo nos focalizaremos sobre trabajos de una sola iteración. Elevamos entonces el índice k de Δ_k, p_k y m_k para simplificar la notación. El sub problema $(*)$ se convierte en:

$\min_{p\in\mathbb{R}^n} m(p) = f + g^T p + \frac{1}{2} p^T B_p$ bajo la restricción $\|p\| \le \Delta (**)$ (1.50)

donde g es la gradiente de f. A la solución la denominamos $p*(\Delta)$ para mostrar la dependencia de Δ.

En la metodología de Dogleg examinamos el efecto del radio Δ de la región de confianza de la solución de $p*(\Delta)$ del sub problema $(**)$. Si B es definida positiva, siempre se ha observado que el minimizador sin restricciones de m es el paso integral $p^B = -B^{-1}g$. Cuando este punto es admisible para $(**)$, es evidentemente una solución, tendremos:

$p*(\Delta) = p^B$ (1.51)

Cuando Δ es minúsculo, la restricción $\|p\| \le \Delta$ garantiza que el término cuadrático de m tenga un pequeño efecto sobre la solución de $(**)$. La solución $p(\Delta)$ es casi la misma que obtendríamos minimizando la función lineal $f + g^T p$ sobre $\|p\| \le \Delta$ entonces tendremos:

$$p^*(\Delta) \cong -\Delta \frac{g}{\|g\|} \text{ cuando } \Delta \text{ es pequeño} \qquad (1.52)$$

Para los valores intermedios de Δ, la solución $p^*(\Delta)$ sigue una trayectoria curva. La m' del método de Dogleg encuentra una solución aproximada reemplazando esta trayectoria curva por un camino formado por dos segmentos. El primero va desde el origen hacia el minimizador a lo largo de la dirección de la pendiente más grande definida por:

$$p^U = \frac{(g^T g)}{g^T B g} \qquad (1.53)$$

El segundo segmento va de p^U hacia p^B. Formalmente, escribimos esta trayectoria como $\tilde{p}(\tau)$ para $\tau \in [0,2]$:

$$\tilde{p}(\tau) = \begin{cases} \tau p^U \text{si } 0 \le \tau \le 1 \\ p^U + (\tau - 1)(p^B - p^U) \text{ si } 1 \le \tau \le 2 \end{cases} \qquad (1.54)$$

Trust-Region-Modified-Dogleg Algorithm (TRMDA)
Anteriormente habíamos descrito el algoritmo (TRDA) el cual se implementa en la función *fsolve* de mathlab.

Sin embargo, esta función toma un vector de función objetivo para minimizar y, de acuerdo con diversas pruebas existentes, el tamaño de este vector afecta directamente el tiempo de cálculo, así como la convergencia de la optimización. Así que decidimos buscar el tamaño óptimo para que el procedimiento sea el más eficiente. Los test efectuados por diversos investigadores han llegado a que el tamaño óptimo del vector es de 50. Por lo tanto, el algoritmo que implementamos divide el vector inicial en pequeños vectores de 50 y aplicamos la función *fsolve* para cada uno de ellos. Estas optimizaciones son mutuamente independientes, por lo que las debemos efectuar paralelamente con la ayuda de la función loop For optimizada de mathlab (par f or).

Ventajas

- Facilidad para la puesta en práctica
- Seguridad de obtener una serie decreciente de valores
- Buen rendimiento cuando la dimensión N aumenta
- Convergencia cuadrática
- La convergencia no se afecta en el caso de soluciones vecinas a la frontera
- Costo (tiempo de cálculo) reducido

Desventajas

- Difícil de aplicar cuando el dominio de definición de la función es complejo o que el mínimo buscado se sitúe cerca de la frontera.
- La evaluación de las derivadas de la función objetivo.

1.5.1.2 Metodología

En lugar de calcular el error sobre la función Y, la calcularemos sobre la función X con:

$$X(z) = A + B*Y(z) \qquad (1.55)$$

Esta operación tiene por objetivo no sub estimar el error de la inversión de Y, ya que un error e sobre Y es equivalente a un error $B*e$ sobre X. Procedemos de la siguiente manera:

Procedimiento

- Simular una muestra X^0 de talla n según una ley $\mathcal{G}\&\mathcal{H}(A,B,g,h)$. Asumimos $(,B,g,h) = (e^7; 2*e^7; 2; 0.4)$
- Calcular $Y^0(z) = \dfrac{X^0(z) - A}{B}$
- Encontrar z^i por el método M_i
- Calcular $X^i(z) = A + B*Y(z^i)$
- Calcular el error cuadrático $EQ_i = \sum_{k=1}^{n} \left(X_k^i - X_k^0\right)^2$
- Rehacer las operaciones N veces y calcular el error cuadrático medio EQM_i (la media de los EQ_i)

Este test nos permite retener la inestabilidad del GSA cuando h es negativa y la divergencia del TRDA cuando n es grande, elegimos usar el TRMDA que es estable y que da los mejores resultados ya sea en tiempo de cálculo o en precisión.

1.5.2 Ley normal inversa Gaussiana (NIG)

En esta parte expondremos brevemente una ley generalizada distinta a la distribución $\mathcal{G}\&\mathcal{H}$, la ley de Wald. Esta ley es conocida por su uso extensivo en el modelaje de la severidad del riesgo operacional.

Ley inversa gaussiana (IG)

1.5.2.1 Definición

Empecemos definiendo la ley inversa gaussiana o ley de Wald, como una ley a dos parámetros cuya NIG es una transformación. En teoría de probabilidades y en estadística, la ley inversa gaussiana es una ley de probabilidad continua con dos parámetros cuyo soporte es $[0,\infty[$. El término “inverso” no debe ser mal interpretado, la ley es inversa en el siguiente sentido: El valor del movimiento browniano a un tiempo fijo es de ley normal, a la inversa, el tiempo en el cual el movimiento browniano con una derivada positiva llega a un valor establecido (fijado) es de ley inversa gaussiana.

La densidad de su probabilidad está dada por:

$$f(\chi,\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp \frac{-\lambda(\chi-\mu)^2}{2\mu^2\chi} \mathrm{II}_{[0,\infty]}(\chi) \qquad (1.56)$$

Donde $\mu > 0$ es su esperanza y $\lambda > 0$ es un parámetro de forma.

Cuando λ tiende al infinito, la ley inversa gaussiana se comporta como una ley normal, tiene varias propiedades similares con este último.

La función generadora de los acumulados (logaritmo de la función característica) de la ley inversa gaussiana es la inversa de la de la ley normal.

Para indicar que una variable aleatoria X es de ley inversa gaussiana de parámetros μ usamos la notación $X \sim IG(\mu, \lambda)$.

1.5.2.2 Máxima verosimilitud

Consideremos el modelo dado por:

$$X \sim IG(\mu, \lambda), \qquad i = 1, 2, ..., n$$

La función de verosimilitud se escribe:

$$\mathcal{L}(\chi; \mu, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \sqrt{\sqrt{\prod_{i=1}^{n} \frac{1}{x_i^3}}} \exp\left(n * \frac{\lambda}{\mu} - \frac{\lambda}{2\mu^2} \sum_{i=1}^{n} \chi_i - \frac{\lambda}{2} \sum_{i=1}^{n} \frac{1}{\chi_i}\right) \quad (1.57)$$

Resolviendo esta ecuación, obtenemos los siguientes estimadores:

$$\hat{\mu} = \overline{X}_n, \quad \hat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\chi_i} - \frac{1}{\hat{\mu}}\right)\right)^{-1} \qquad 1.58)$$

Observamos que:

$$\hat{\mu} \sim IG(\mu, n\lambda), \quad \frac{n}{\hat{\lambda}} \sim \frac{1}{\lambda} \chi^2_{n-1} \qquad (1.59)$$

1.5.2.3 Integración del umbral de recolección

Consideremos ahora la verosimilitud condicional con el fin de tener en cuenta el umbral de recolección H:

$$\mathcal{L}_{X|H}(\chi | \mu, \lambda) = \frac{\left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \sqrt{\prod_{i=1}^{n} \frac{1}{\chi_i^3}} \exp\left(n * \frac{\lambda}{\mu} - \frac{\lambda}{2\mu^2} \sum_{i=1}^{n} \chi_i - \frac{\lambda}{2} \sum_{i=1}^{n} \frac{1}{\chi_i}\right)}{1 - F(H; \mu, \lambda)} \quad (1.60)$$

Reemplazamos el valor de la función de repartición por su valor:

$$\mathcal{L}_{X|H}(\chi | \mu, \lambda) = \frac{\left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \sqrt{\prod_{i=1}^{n} \frac{1}{\chi_i^3}} \exp\left(n * \frac{\lambda}{\mu} - \frac{\lambda}{2\mu^2} \sum_{i=1}^{n} \chi_i - \frac{\lambda}{2} \sum_{i=1}^{n} \frac{1}{\chi_i}\right)}{1 - \Phi\left(\sqrt{\frac{\lambda}{H}}\left(\frac{H}{\mu} - 1\right)\right) - \exp\left(\frac{2\lambda}{H}\right) \Phi\left(-\sqrt{\frac{\lambda}{H}}\left(\frac{H}{\mu} + 1\right)\right)} \quad (1.61)$$

1.5.2.4 Convergencia del método de máxima verosimilitud

	Parámetros conocidos		Estimación sin umbral		Estimación con un umbral de 1000			
	$\log(\mu)$	$\log(\lambda)$	$\log(\hat{\mu})$	$\log(\hat{\lambda})$	$\log(\hat{\mu})\|H$	$\log(\hat{\lambda})\|H$	$F_{\mu,\lambda}(H)$	$F_{\hat{\mu},\hat{\lambda}}(H)$
M1	10	5	10,00	4,99	10,08	5,21	0,70	1
M2	11	6	11,00	6,00	11,00	6,03	0,52	0
M3	12	7	12,02	7,00	12,03	7,02	0,30	0
M4	12,5	7	12,53	7,50	12,53	7,51	0,18	0
M5	13	8	13,00	8,00	12,98	7,99	0,08	0

1.5.3 Ley NIG

1.5.3.1 Definición

Si Y sigue una ley $IG\left(\delta,\sqrt{\alpha^2-\beta^2}\right)$ y $X|Y$ sigue una ley $N(\mu+\beta Y,Y)$ entonces X es de ley $NIG(\alpha,\beta,\mu,\delta)$ de densidad:

$$f_{NIG}(x;\alpha,\beta,\mu,\delta)=\frac{\alpha}{\pi}\exp\left(\delta\sqrt{\alpha^2-\beta^2}-\beta\mu\right)\frac{K_1\left(\alpha\delta\sqrt{1+\left(\frac{x-\mu}{\delta}\right)^2}\right)}{\sqrt{1+\left(\frac{x-\mu}{\delta}\right)^2}}\exp(\beta x) \quad (1.62)$$

donde $\alpha\in\mathbb{R},\ \alpha>0,\delta>0,\ \mu\in\mathbb{R},\ 0<|\beta|<\alpha$ y K_1 es la función de Bessel modificada de tercera especie con el índice 1. El caso límite $\alpha\to\infty$ corresponde a la ley normal.

La distribución NIG tiene dos propiedades interesantes:

- Propiedad de escala

$$X\sim NIG(\alpha,\beta,\mu,\delta)\ \Leftrightarrow\ cX\sim NIG\left(\frac{\alpha}{c},\frac{\beta}{c},c\mu,c\delta\right) \quad (1.63)$$

- Propiedad de circunvolución

$$NIG(\alpha,\beta,\mu_1,\delta_1)*NIG(\alpha,\beta,\mu_2,\delta_2)\Leftrightarrow NIG(\alpha,\beta,\mu_1+\mu_2,\delta_1+\delta_2) \quad (1.64)$$

1.5.3.2 Método de los momentos

Los estimadores por el método de los momentos fueron calculados en el trabajo de Eriksson, Forsberg, y Ghysels:

Sea $X\sim NIG(\alpha,\beta,\mu,\delta)$ denominando $\overline{X}_n,\overline{V}_n^2,\overline{S}_n,\overline{K}_n$ a la media, la varianza, la asimetría y la kurtosis de la muestra, tendremos entonces:

$$\widehat{\alpha} = 3\sqrt{\rho}\,(\rho-1)^{-1}\overline{V}_n\left|\overline{S}_n^{-1}\right|$$

$$\widehat{\beta} = 3(\rho-1)^{-1}\overline{V}_n\overline{S}_n^{-1}$$

$$\widehat{\mu} = \overline{X}_n - 3\rho^{-1}\overline{V}_n\overline{S}_n^{-1}$$

$$\widehat{\delta} = 3\rho^{-1}(\rho-1)^{1/2}\,\overline{V}_n\left|\overline{S}_n^{-1}\right| \qquad (1.65)$$

1.5.4 Optimización por recocido simulado

Al contrario de los algoritmos de optimización que hemos visto antes, el algoritmo de recocido simulado permite la búsqueda de un mínimo global. Esa es su principal ventaja, y su desventaja es que usa mucho tiempo en el cálculo, mayor que el que usan los métodos por iteraciones (no probabilísticos).

1.5.4.1 Principio

Este método viene de un principio de la termodinámica usado en metalurgia para mejorar la calidad de un sólido, se inspira en la evolución de este último hacia una posición de equilibrio luego de su enfriamiento.

Asumimos que tenemos un sistema físico a la temperatura T. Hacemos la hipótesis que S puede tener un número innumerable de estados $i \in \mathbb{N}$. A cada estado i asociamos un nivel de energía E_i. Denominamos X al estado del sistema. Tendremos entonces, la distribución que caracteriza al equilibrio térmico (ley de Boltzmann):

$$P_T(X=i) = \frac{1}{Z(T)}e^{\frac{E_i}{k_B T}} \qquad (1.66)$$

Donde k_B es la constante de Boltzmann y Z es una función de normalización dada por:

$$Z(T) = \sum_{i\in\mathbb{N}} \exp\left(-\frac{E_i}{k_B T}\right) \qquad (1.67)$$

Si i y j son dos estados. Definimos ΔE como su diferencia de energía. Tenemos entonces:

$$\frac{P_T(X=i)}{P_T(X=j)} = e^{\frac{\Delta E}{k_B T}} \qquad (1.68)$$

Podemos deducir que sí $\Delta E > 0$, el estado j es mas probable que el estado i e inversamente.

Sin embargo, la relación de probabilidades depende también de $k_B T$, y si este término es grande frente a ΔE, los estados de i y de j son igualmente probables.

Hagamos entonces una analogía entre el sistema físico y el problema de minimización:

- Energía del sistema $\leftrightarrow$ costo de una solución
- $k_B T \leftrightarrow$ parámetro de control denominado T

Sobre esta analogía vamos a introducir una implementación del método de recocido simulado, el algoritmo de Metropolis. La idea es la de efectuar un movimiento según una distribución de probabilidad que dependa de la calidad de distintos vecinos:

- Los mejores vecinos tienen una probabilidad más elevada
- Los menos buenos tienen una probabilidad más débil

El parámetro T (de temperatura) varía durante la investigación: T es elevado al inicio, luego disminuye y termina por tender a cero.

1.5.4.2 El algoritmo de Metropolis

Si (S, f) es un problema de optimización, e i, j dos soluciones, introducimos el criterio de aceptación de Metropolis:

$$P_T(\text{aceptar } j) = \begin{cases} 1 \text{ si } \Delta f \geq 0 \\ \text{sino } e^{-\frac{\Delta f}{T}} \end{cases} \qquad (1.69)$$

Donde T es nuestro parámetro de control y $\Delta f = f(j) - f(i)$.

El algoritmo es definido a partir de un estado inicial $i = i_0$ dado, y de la repetición de dos etapas. Además, vamos a bajar la temperatura con una frecuencia regular.

La primera etapa es llamada desplazamiento. Se trata de generar una solución admisible j a partir de i. En una segunda etapa, ponemos en juego el criterio de aceptación con el fin de decidir si vamos a mantener j o no.

Hay varias reglas de desplazamiento, acá solo mostraremos dos de las más conocidas:

- La regla de Cerny: Construimos un punto de referencia $\{e_1, ..., e_n\} \in \mathbb{R}^n$ con e_1 apuntando en la dirección del mejor resultado obtenido. Elegimos una dirección d de manera aleatoria entre $e_1, ..., e_n$ con una probabilidad más grande para e_1, de ahí nos desplazamos aleatoriamente siguiendo a d.
- La regla uniforme: Dibujamos al azar un vector u en el cual cada uno de sus componentes sigue una ley uniforme sobre $[-1,1]$. Nos movemos de $\delta j = q.u$, donde q es un paso fijo.

En lo que concierne al enfriamiento podemos considerar alguna de las reglas siguientes:

- Reducir T a $(1-\varepsilon)T$ todos los desplazamientos m, ε y m los elegimos por experiencia, hasta $T \simeq 0$.
- Fijamos un número total K (muy grande) de desplazamiento y reducimos T cada $\left(\frac{K}{N}\right)$ desplazamientos (N veces) haciendo: $T_k = T_0\left(1-\frac{k}{N}\right)$ para $k = 1...N$, usualmente elegimos $\alpha = 1, 2$ o 4.

Algoritmo:

Inicio

Generar una configuración inicial $x_0; x = x_0$

$T := T_0$

Repetir

Nb_movimientos:=0

Para $i := 1$ a Nb_iter

Generar un vecino x' de x

Calcular $\Delta = f(x') - f(x)$

Si CritMetropolis (Δ, T), entonces

$x := x'$

$\text{Nb\_movimientos}++;$

Tasa_aceptación$:= i / (\text{Nb_movimientos})$

$T := \text{Disminucióntemperatura}(T);$

Hasta $< CritStop >$

Regresar a la mejor configuración encontrada

Fin.

CritStop es un criterio de parada:

- La tasa de aceptación se vuelve muy pequeña
- f deja de evolucionar.

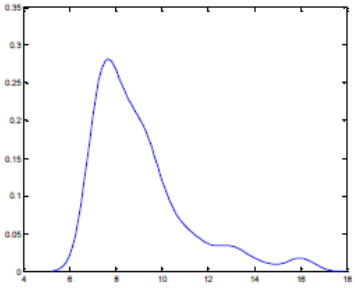
El inconveniente de este algoritmo es que es muy lento, por lo que para la optimización de una función con cuatro variables puede ser inutilizable. Además, la elección de los parámetros se debe de hacer en cada llamada a la función, lo que hace difícil poder automatizar el proceso de estimación de los parámetros; sin embargo, en Mathlab podemos encontrar una versión optimizada de este

algoritmo lo que nos permite elegir de manera automática los valores buenos para los parámetros de temperatura, número de iteraciones, etc.

1.5.5 Tratamiento de pérdidas aberrantes

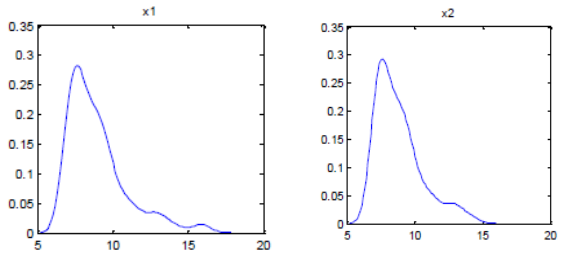
1.5.5.1 Reducción de la varianza

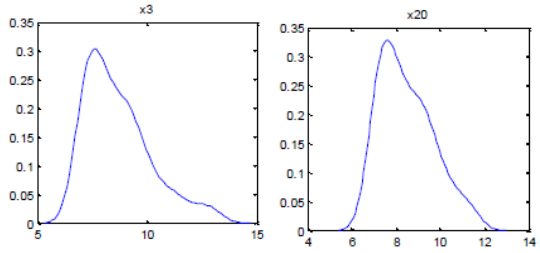
En las páginas precedentes hemos visto que podemos encontrar valores de carga en capital extremadamente grandes (aberrantes), estos valores pueden ser explicados por una distribución empírica de pérdidas muy heterogéneas. Para convencernos mejor, dibujamos la densidad de la distribución estimada por el método de las pérdidas de registro de Kernel.



Observamos la presencia de un segundo pico de amplitud sustancial en la cola de la distribución (alrededor de e^{16}), es a causa de esto que la ley logo normal trata de modelar esta categoría y que la ley $\mathcal{G}\&\mathcal{H}$ llega un poco mejor al elegir el parámetro g grande para presentar la varianza de las cifras y un parámetro h importante para caracterizar su cola gruesa. A pesar de que este modelamiento es mejor que el que se puede obtener con una ley logo normal en términos de adecuación, este no refleja el nivel de riesgo real de la banca. Es entonces indispensable suavizar los datos de severidad antes del modelamiento, las pérdidas, una vez separadas deben ser tratadas como distintos escenarios.

Tratemos en una primera etapa de separar las pérdidas aberrantes con un criterio simplista, aquel de la reducción de la varianza, y evaluemos el impacto de esta operación sobre el valor del CaR (cargas en capital). El método evalúa las varianzas de las muestras obtenidas eliminando una pérdida diferente en cada iteración, y elegida la muestra minimiza la varianza. Veamos los gráficos obtenidos eliminando 1,5,10 y 20 pérdidas de la muestra de origen.



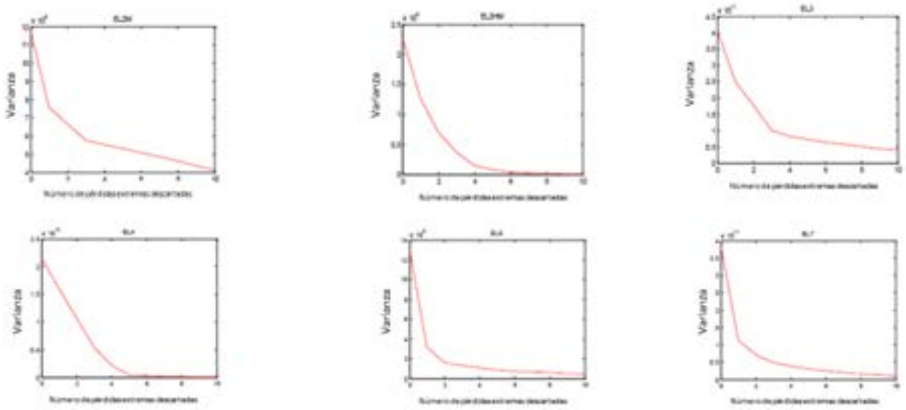


Calculemos ahora, luego de la estimación de los parámetros, los CaR correspondientes a cada caso.

	X_1	X_5	X_{10}	X_{20}
CaR	28,769,917,730	285,834,814	18,286,814	1,158,456

1.5.5.2 Elección del número de pérdidas a externalizar

Observamos que en la mayoría de casos las perdidas eliminadas son pérdidas extremas, para elegir los valores que se alejan mucho del perfil de riesgo trazamos la evolución del criterio (la varianza en un primer momento) de la muestra en función del número de pérdidas descartadas:



Hay que observar que en ciertas categorías las dos pérdidas mas grandes de la muestra no son compatibles con el resto de datos (dos últimas de abajo a la derecha) en la medida en que ellas se alejan mucho de la muestra.