MOMENTOS DA CLASSE DE DISTRIBUIÇÕES MISTURAS DA ESCALA NORMAL TRUNCADAS

Aluno: Eraldo Barbosa dos Anjos Filho [1]

Orientador: Prof. Aldo William Medina Garay[2]

Resumo

Neste trabalho desenvolvemos os procedimentos computacionais, baseados nos resultados obtidos por Genç (2013) e Garay et al. (2017), que permitem obter os quatros primeiros momentos da classe de distribuições Misturas da Escala Normal (SMN), duplamente truncadas, que tem como casos particulares as distribuições Normal truncada, t de Student truncada, Slash truncada, Normal Contaminada truncada e Pearson tipo VII truncada. Assim, a partir de diversas manipulações algébricas, implementamos e elaboramos o pacote TSMN, o qual já está disponível, para seu uso livre, no repositório CRAN (https://CRAN.R-project.org/package=TSMN).

Palavras-chave: Momentos, misturas da escala normal, misturas da escala normal truncadas.

Abstract

In this work, we develop the computational procedures, based on the results obtained by Genç (2013)) and Garay et al. (2017), that allow to obtain the first four moments of the class of distributions. Normal Scale Mixtures (SMN), double truncated, has as particular cases the Truncated Normal, Student t truncated, Truncated Slash, Truncated Normal, and Truncated Type VII Pearson distributions. From various algebraic manipulations, we have implemented and developed the TSMN package, which is now available for free use in the CRAN repository (https://CRAN.R-project.org/package=TSMN).

Keywords: Moments, scale mixtures of normal distributions, scale mixtures of normal distributions truncated.

[1] Universidade Federal de Pernambuco (UFPE), ebdaf1@de.ufpe.br

[2] Universidade Federal de Pernambuco (UFPE), agaray@de.ufpe.br

Introdução

A distribuição normal é muito utilizada na estatística clássica. No entanto dados com possíveis pontos atípicos podem gerar distribuições de cauda pesada, distorcida ou multi-modal. Nesse caso, podem ocorrer violações na suposição de normalidade, fazendo com que as inferências não sejam robustas. Para lidar com o problema das observações atípicas varias propostas foram feitas na literatura. Por exemplo, Lange et al. (1989) discutiram o uso da distribuição Student-t em modelos de regressão multivariada. Neste caso, o parâmetro de graus de liberdade é a escolha natural para controlar a curtose. Outra solução proposta para trabalhar com um conjunto de dados com pontos atípicos é utilizando a distribuição mistura da escala normal.

Objetivo

Neste trabalho desenvolvemos e implementamos a metodologia para se obter os quatros primeiros momentos da classe de distribuições Misturas da Escala Normal (SMN), duplamente truncadas, de uma forma rápida e pratica.

Material e Método

Misturas da Escala Normal (SMN): Ao longo deste trabalho, $X \sim N(\mu, \sigma^2)$ denota uma variável aleatória X com distribuição normal com média μ e variância σ^2 e $\phi(\cdot|\mu, \sigma^2)$denota sua função de densidade de probabilidade (fdp). $\phi(\cdot)$ e $\Phi(\cdot)$ indicam, respectivamente, a fdp e a função de distribuição acumulada (fda) da distribuição normal padrão. Em geral, usamos a convenção tradicional que denotou uma variável aleatória (ou um vetor aleatório) por uma letra maiúscula e sua realização pelas minúsculas correspondentes. $X^\top$ é a transposição de X. $X \perp Y$ indica que as variáveis aleatórias X e Y são independentes. Começamos definindo as distribuições de SMN, através de sua formulação hierárquica, e então apresenta mos algumas propriedades adicionais

Definição 1: A variável aleatória X segue uma distribuição SMN com parâmetro de locação μ e parâmetro de escala $\sigma^2 > 0$, se possuir a seguinte representação:

$$X = \mu + U^{-\frac{1}{2}} Z, \tag{1}$$

em que $Z \sim N(0, \sigma^2)$, U é uma variável aleatória positiva com fda $H(\cdot|\nu)$ e ν é um escalar ou vetor de parâmetros indexados a U. Se utiliza a notação: $X \sim SMN(\mu, \sigma^2, \nu)$e quando $\mu = 0$ e $\sigma^2 = 1$ se tem a chamada distribuição SMN padrão.

Da equação (1), tem-se que $X|U=u \sim N(\mu, u^{-1}\sigma^2)$. Integrando em U, na densidade conjunta de X e U, obtêm-se a seguinte densidade marginal:

$$f_{SMN}(x|\mu,\sigma^2,\nu) = (2\pi\sigma^2)^{-\frac{1}{2}}\int_0^\infty u^{\frac{1}{2}} exp\left\{-\left(\frac{u}{2\sigma^2}\right)(x-\mu)^2\right\} dH(u|\nu),$$

em que $H(.|\nu)$ é a fda de U e determina a forma da distribuição SMN. U e $H(.|\nu)$ são denominados fator escala e distribuição de mistura, respectivamente.

Definição 2: Seja $X \sim SMN(\mu,\sigma^2,\nu)$ e os valores fixos $a<b$, tal que $P(a<X<b)>0$. A variável aleatória Y segue uma distribuição truncada SMN, no intervalo (a,b), se seguir a mesma distribuição de $X|X \in (a,b)$. Neste caso se escreve $Y \sim TSMN_{(a,b)}(\mu,\sigma^2,\nu)$.

De uma consequência da definição (2), a função de densidade da variável aleatória $Y \sim TSMN_{(a,b)}(\mu,\sigma^2,\nu)$ é dada por:

$$f_{TSMN}(y|\mu,\sigma^2,\nu;(a,b)) = \frac{f_{SMN}(y|\mu,\sigma^2)}{F_{SMN}\left(\frac{b-\mu}{\sigma}\right) - F_{SMN}\left(\frac{a-\mu}{\sigma}\right)}, a<y<b,$$

em que $F_{SMN}(.)$ denota a fda da distribuição SMN padrão.

Momentos das TSMN: Dos resultados obtidos por Genç (2013) e Garay et al. (2017), define-se o seguinte teorema:

Teorema 1: Seja $X \sim SMN(0,1,\nu)$ com fator escala U e distribuição de mistura $H(|\nu)$, tem-se que para $a<b$, $s=\{1,2,3,4\}$, $E(X^s|X \in <a,b>)$ é dado por:

$$\begin{aligned}
E[X|X \in <a,b>] &= \tau(a,b)\left[E_\phi\left(-\frac{1}{2},a\right) - E_\phi\left(-\frac{1}{2},b\right)\right]\\
E[X^2|X \in <a,b>] &= \tau(a,b)[E_\Phi(-1,b) - E_\Phi(-1,a)]\\
&+ \tau(a,b)\left[aE_\phi\left(-\frac{1}{2},a\right) - bE_\phi\left(-\frac{1}{2},b\right)\right]\\
E[X^3|X \in <a,b>] &= 2\tau(a,b)\left[E_\phi\left(-\frac{3}{2},a\right) - E_\phi\left(-\frac{3}{2},b\right)\right]\\
&+ \tau(a,b)\left[a^2E_\phi\left(-\frac{1}{2},a\right) - b^2E_\phi\left(-\frac{1}{2},b\right)\right]\\
E[X^4|X \in <a,b>] &= 3\tau(a,b)\left[\left(E_\Phi(-2,b) - E_\Phi(-2,a)\right)\right]\\
&+ 3\tau(a,b)\left[\left(aE_\phi\left(-\frac{3}{2},a\right) - bE_\phi\left(-\frac{3}{2},b\right)\right)\right]\\
&+ \tau(a,b)\left[a^3E_\phi\left(-\frac{1}{2},a\right) - b^3E_\phi\left(-\frac{1}{2},b\right)\right]
\end{aligned}$$

em que

$$\tau(a,b) = \frac{1}{F_{SMN}(b) - F_{SMN}(a)};$$

$$E_{\phi}(r,h) = E\left[U^r \phi\left(hU^{\frac{1}{2}}\right)\right] = \int_0^{\infty} u^r \phi\left(hu^{\frac{1}{2}}\right) dH(u|\nu);$$

$$E_{\phi}(r,h) = E\left[U^r \Phi\left(hU^{\frac{1}{2}}\right)\right] = \int_0^{\infty} u^r \Phi\left(hu^{\frac{1}{2}}\right) dH(u|\nu).$$

Casos particulares:

Distribuição Pearson Tipo VII: Neste caso se considera $U \sim Gamma\left(\frac{\nu}{2}, \frac{\delta}{2}\right)$, com $\nu > 0$ e $\delta > 0$, em que $Gamma(a,b)$ denota a distribuição gama com media $\frac{a}{b}$. A densidade de uma variável aleatória X, definida em (1), é dada por:

$$f_{PVII}(x|\nu,\delta) = \frac{1}{B\left(\frac{\nu}{2}, \frac{1}{2}\right)\sqrt{\delta}} \left(1 + \frac{x^2}{\delta}\right)^{-\frac{\nu+1}{2}}$$

em que, $\delta > 0$ e $\nu > 0$ são os parâmetros de forma e $B(a,b)$ representa a função beta. Se utiliza a notação $X \sim PVII(0,1;\nu,\delta)$. Neste caso se tem

$$E_{\Phi}(r,h) = \frac{\Gamma\left(\frac{\nu+2r}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\delta}{2}\right)^{-r} F_{PVII}(h|\nu+2r,\delta),$$

$$E_{\phi}(r,h) = \frac{\Gamma\left(\frac{\nu+2r}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi}} \left(\frac{\delta}{2}\right)^{\frac{\nu}{2}} \left(\frac{h^2+\delta}{2}\right)^{\frac{-(\nu+2r)}{2}},$$

onde $\Gamma(a)$ é a função gamma e $F_{PVII}(\cdot)$ é a fda da distribuição Pearson Tipo VII. Quando $\delta = \nu$ se tem a distribuição t-Student com ν graus de liberdade. Também, se tem a distribuição Cauchy quando $\delta = \nu = 1$.

Distribuição Slash: Para esta caso a distribuição do fator escala U é $Beta(\nu,1)$, com $\nu > 0$. A densidade da variável aleatória X, definida em (1), é dada por:

$$f_{sl}(x|\nu) = \nu\int_0^1 u^{\nu-1}\phi(xu^{\frac{1}{2}})du$$

Se utiliza a notação $X \sim SL(0,1;\nu)$. Para este caso se tem

$$E_\Phi(r,h) = \left(\frac{\nu}{\nu+r}\right) F_{SL}(h|\nu+r);$$

$$E_\phi(r,h) = \frac{\nu}{\sqrt{2\pi}}\left(\frac{h^2}{2}\right)^{-(\nu+r)} \Gamma\left(\nu+r,\frac{h^2}{2}\right),$$

em que $\Gamma(a,b) = \int_0^b e^t t^{a-1} dt$ é a função gamma incompleta. Verificar Lema 6 em Genç (2013), $F_{SL}(\cdot)$ é a fda da distribuição Slash.

Distribuição Normal Contaminada: U é a variável aleatória discreta com dois estados, 1 ou γ. Para este caso a função de probabilidade de U é

$$U = \begin{cases} \gamma & \text{com probabilidade } \varepsilon \text{ ;} \\ 1 & \text{com probabilidade } 1-\varepsilon, \end{cases}$$

seguindo imediatamente a densidade da variável aleatória X, definido em (1),

$$f_{CN}(x|\varepsilon,\gamma) = \varepsilon\phi(x|0,\gamma^{-\frac{1}{2}}) + (1-\varepsilon)\phi(x)$$

então, se tem

$$E_\Phi(r,h) = \gamma^r F_{CN}(h|\epsilon,\gamma) + (1-\gamma^r)(1-\epsilon)\Phi(h);$$

$$E_\phi(r,h) = \epsilon\gamma^r\phi(h\sqrt{\gamma}) + (1-\epsilon)\phi(h)$$

em que $F_{CN}(\cdot)$ é a fda da distribuição Normal Contaminada.

Aplicação

O *Tail Conditional Expectation* (TCE) ou *Expected Shortfall* é uma medida útil para gerenciamento de risco, a qual calcula a média das piores perdas, dado que a perda excedeu um valor particular x_q, isto é

$$TCE(q) = E(X|X > x_q),$$

em que x_q representa o quantil da ordem q da distribuição da perda X. Esse valor limite é chamado de Valor em Risco (VaR) Artzner et al. (1999).

O $TCE(q)$ é de fato a média da distribuição de X truncada por $x = x_q$. Assumindo que X segue a distribuição t-Student, então $TCE(q)$ será

$$TCE(q) = \mu + \frac{\Gamma\left(\frac{\nu-1}{2}\right)\nu^{\frac{\nu}{2}}\sigma}{2\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)\left[1 - F_{t(\nu,\mu,\sigma)}(x_q)\right]}\left[\nu + \left(\frac{x_q-\mu}{\sigma}\right)^2\right]^{-\frac{(\nu-1)}{2}}, \nu > 1. \qquad (2)$$

Para $n > 2$, nós temos a seguinte expressão

$$TCE(q) = \mu + \frac{\sqrt{\frac{\nu+1}{\nu}}\sigma f_{t(\nu-2)}\left(\sqrt{\frac{\nu-2}{\nu}}z_q\right)}{1 - F_{t(\nu)}(z_q)}$$

onde $z_q = \frac{x_q - \mu}{\sigma}$ Landsman e Valdez (2003).

Seguindo a aplicação apresentada por Genç (2013), temos um conjunto de dados reais chamado de perdas ocasionadas por furacões, que representa o total de danos causados por 35 furacões entre os anos de 1949 e 1980 Klugman et al. (2012). Na Tabela 1 temos os dados, os valores estão na 10^7\$.

Tabela 1 – Perdas Ocasionadas por Furacões Expressos em 10^7\$

0.06766	0.07123	0.10562	0.14474
0.15351	0.16983,	0.18383,	0.19030,
0.25304	0.29112,	0.30146,	0.33727,
0.40596	0.41409,	0.47905,	0.49397,
0.52600	0.59917,	0.63123,	0.77809,
1.02942	1.03217	1.23680	1.40136
1.92013	1.98446	2.27338	3.29511
3.61200	4.21680	5.13586	5.45778
7.50389	8.63881	16.38000	

Fonte: Klugman et al. (2012).

A diferença do enfoque apresentado por Genç (2013) neste trabalho calcularemos os valores do TCE considerando diferentes distribuições da classe SMN truncadas, como a distribuição t de Student e para a distribuição Slash, serão utilizadas as estimativas de máxima verossimilhança, $\boldsymbol{\nu = 3.198035}$, $\boldsymbol{\mu = 2.049004}$ e $\boldsymbol{\sigma^2 = 1.9000434}$. Como sugerido por Genç (2013), o valor de $\boldsymbol{x_q}$ foi obtido através da função $\boldsymbol{qTF}$ do pacote gamlss do sistema R.

As estimativas do $\boldsymbol{TCE}$ foram obtidas utilizando a equação (2) e estão apresentadas na Tabela 2.

Tabela 2 – Quantis e estimativas de $TCE(q)$ para vários q.

q	x_q	$\hat{TCE}_T$	$\hat{TCE}_{Slash}$
0.5	2.049004	3.530573	-
0.75	3.094951	4.011771	3.579064
0.90	4.269292	4.809666	4.475003
0.95	5.212652	5.723768	5.227567
975	6.286011	6.964605	6.110005
0.99	8.008423	9.189713	7.678551
999	14.913166	18.906600	15.380544

Resultados e Discussão

Os resultados obtidos por Genç (2013) e Garay et al. (2017) foram implementados e como resultado foi elaborado o pacote do sistema R: TSMN, Dos Anjos Filho e Garay (2017), o qual pode ser livremente instalado. O pacote inclui duas funções relacionadas com as misturas da escala normal truncadas.

- Retornar os primeiros quatro momentos teóricos das distribuições TSMN implementas.
- Gerar uma amostra aleatória a partir das distribuições TSMN implementas.

Conclusão

Este trabalho generaliza os resultados obtidos por Kim (2008), tanto para os casos com e sem truncamento, pois apresenta uma classe de distribuições mais robustas que a distribuição t de Student e utiliza as distribuições sem truncamento, quando o limite inferior e limite superior converge para o infinito. Os momentos obtidos como resultado de uma das funções do pacote TSMN, fornece ganho de tempo e facilidade na obtenção de estatísticas como media, desvio padrão, curtose e assimetria.

Referências

ARTZNER, P.; DELBAEN, F.; EBER, J. M.; HEATH, D. **Coherent measures of risk**. Mathematical finance, v. 9, n. 3, p. 203-228, 1999. ISSN 1467-9965.

GARAY, A. M.; LACHOS, V. H.; BOLFARINE, H.; CABRAL, C. R. **Linear censored regression models with scale mixtures of normal distributions.** Statistical Papers, v. 58, n. 1, p. 247-278, 2017.

DOS ANJOS FILHO, E. B.; GARAY, A. M. **TSMN: Truncated Scale Mixtures of Normal Distributions**. R CRAN, https://CRAN.R-project.org/package=TSMN, 2017.

GENÇ, A. İ. **Moments of truncated normal/independent distributions**. Statistical Papers, v. 54, n. 3, p. 741-764, 2013.

KIM, H.-J. **Moments of truncated Student-t distribution**. Journal of the Korean Statistical Society, v. 37, n. 1, p. 81-87, 2008.

KLUGMAN, S. A.; PANJER, H. H.; WILLMOT, G. E. **Loss models: from data to decisions**. John Wiley & Sons, 2012. ISBN 0470391332.

LANDSMAN, Z. M.; VALDEZ, E. A. **Tail conditional expectations for elliptical distributions**. North American Actuarial Journal, v. 7, n. 4, p. 55-71, 2003. ISSN 1092-0277.

LANGE, K. L.; LITTLE, R. J.; TAYLOR, J. M. **Robust statistical modeling using the t distribution**. Journal of the American Statistical Association, v. 84, n. 408, p. 881-896, 1989.

Anexo - Código R - Pacote “Truncated Scale Mixtures of Normal Distributions“

```
#install.packages("TSMN")

library(TSMN)

hurricane <- c(6766, 7123, 10562, 14474, 15351, 16983, 18383, 19030, 25304, 29112, 30146, 33727, 40596, 41409, 47905, 49397, 52600, 59917, 63123, 77809, 102942, 103217, 123680, 140136, 192013, 198446, 227338, 329511, 361200, 421680, 513586, 545778, 750389, 863881, 1638000)

hurricane <- hurricane/(10^5)

mu = mean(hurricane)

sigma2 = 1.9000434

nu = 3.198035

#install.packages("gamlss")

library(gamlss)

q <- c(0.5, 0.75, 0.90, 0.95, 0.975, 0.99, 0.999)

xq <- qTF(q, mu, sigma=sqrt(sigma2), nu)

TCE_hat <- c()

for(i in 1:length(xq)){
```

```
aux <-  TSMNmoments(mu = mu, sigma2 = sigma2, nu = nu, lower = xq[i], upper = Inf, dist = "T")

TCE_hat[i] <- aux$EY1}

TCE_Slash_hat <- c()

for(i in 1:length(xq)){

aux <-  TSMNmoments(mu = mu, sigma2 = sigma2, nu = nu, lower = xq[i]-1, upper = Inf, dist = "Slash")

 TCE_Slash_hat[i] <- aux$EY1}

final <- data.frame(xq, TCE_hat, TCE_Slash_hat)
```