Journal of Applied Statistics
Vol. 00, No. 00, April 2012, 1–17

RESEARCH ARTICLE

A Generalized Quantile Regression Model

Vahid Nassiri[a*] and Ignace Loris[b]

[a]*Department of Mathematics, Vrije Universiteit Brussel, Belgium.* [b]*Département de Mathématique, Université Libre de Bruxelles, Belgium.*

(*Received 00 Month 200x; in final form 00 Month 200x*)

A new class of probability distributions, the so-called connected double truncated gamma distribution, is introduced. We show that using this class as the error distribution of a linear model leads to a generalized quantile regression model that combines desirable properties of both least squares and quantile regression methods: robustness to outliers and differentiable loss function.

Keywords: quantile regression, log-concave density, penalization, soft thresholding, outlier, long-tail.

1. Introduction

The least squares estimation of the parameters of a model, connected with a Gaussian error distribution, has been the most favorite method of fitting models for several decades. Part of the interest in the Gaussian error and least squares estimates comes from their convenient mathematical properties. However, in many situations considering the error *as if* it is Gaussian is very unrealistic and a famous quote in statistics suggests: *"Normality is a myth"*. Many efforts have therefore been made to extend this concept. In this case gaining more realistic results requires the solution of a more difficult problem.

One of the very first competitors of the least squares (LS) method is the least absolute deviations (LAD) method. LS corresponds to minimizing the ℓ_2 norm of the error while LAD uses the ℓ_1 norm instead. As it is well-known, in case of a single variable, mean is the minimizer of the LS expected loss, while median is the minimizer of the LAD expected loss. As a result of its corresponding distribution, LAD is more robust to outliers. However, the ℓ_1 norm is not differentiable which makes the numerical minimization task involved more cumbersome when many variables are involved.

Quantile regression (QR) of [13] is an extension of the traditional LAD regression to an asymmetric version. However, the non-differentiability of the corresponding loss function presents a problem and finding a robust error distribution that is easy to work with in practice is a challenging problem.

The so-called curse of dimensionality also presents a serious problem in recent statistical literature. **This term was introduced by Richard Bellman [3], and has various interpretations in different fields. The common property**

[*] Corresponding author. Email: vnassiri@vub.ac.be

ISSN: 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664763.YYYY.XXXXXX
http://www.tandfonline.com

between all of them is that they refer to high dimensional data. Here, by curse of dimensionality, we mean that the number of variables in the model is much larger than the number of observations. This requires efficient estimation in high dimensional parameter space. As a result sparse modeling has become prominent in recent years, **and [7] have mentioned sparse models as one of the frontiers of statistical research in the 21st century.** In this context, sparsity refers to having many zeros in the estimated vector of parameters. In other words, a model selection is done at the same time as the model fitting. Sparsity can be introduced in the model via a penalty function added to the loss function. The most famous type of sparse models is the Lasso introduced by [21]. Lasso adds the ℓ_1-norm of the parameters vector to the LS loss function. The resulting model is sparse but it is not robust to outliers or asymmetric error.

After choosing a good model and a suitable penalty, solving the penalized problem, is also a challenging aspect. Lasso can be interpreted as a convex minimization problem with a differentiable loss function and a separable **non-differentiable** penalty. Minimization problems of this specific type can be solved efficiently, even in case of a large number of variables [2]. Using the non-differentiable LAD loss instead of the LS loss, together with a ℓ_1 norm penalty, leads to a more difficult minimization problem.

In this paper we introduce a differentiable log-concave probability distribution with heavier tails than the Gaussian distribution and (possibly) with skewness. In fact, the proposed distribution has tails that are at least as heavy as the tails of the error probability distribution used in QR. We therefore also propose a generalized version of the traditional QR, with a convex and differentiable loss function. When used to model errors in a *sparse* regression problem, several of the nice properties of Lasso, i.e. convexity and differentiability of the loss function, will be preserved and the minimization algorithm of [2] can be used.

The symmetric version of the new class of distributions, so-called connected double truncated gamma, is introduced in Section 2. Many of the properties of this class will be discussed. Section 3 will present the asymmetric version of the density. It will be shown how this distribution can be connected with a generalized QR. Section 4.1 is dedicated to introducing the model for generalized QR and its penalized version. An algorithm to solve the penalized problem fast and easily is given in Section 4.2. Finally, the paper is concluded in Section 5.

2. The symmetric connected double truncated gamma distribution

In this section a new class of probability distributions is introduced. The corresponding loss function will produce a generalized quantile regression. The distribution is called **symmetric** connected double truncated gamma distribution (SCDTG). Many of its properties will be studied. An asymmetric version will be introduced in Section 3.

We first introduce some notations and formulas that will frequently be used throughout the text. Firstly

$$\Gamma(s) = \int_0^{+\infty} u^{s-1} e^{-u} \mathrm{d}u \quad \text{and} \quad \Gamma(s,r) = \int_r^{+\infty} u^{s-1} e^{-u} \mathrm{d}u \tag{1}$$

are the usual gamma and upper incomplete gamma functions. We will also use the

CDF of a gamma random variable with parameter α at point x:

$$G_\alpha(x) = \int_0^x \frac{1}{\Gamma(\alpha+1)} u^\alpha e^{-u} \mathrm{d}u. \tag{2}$$

Thus the incomplete gamma function can be computed by:

$$\Gamma(s, r) = (1 - G_{s-1}(r))\Gamma(s). \tag{3}$$

Furthermore the incomplete gamma function satisfies the following equation

$$\begin{aligned}\Gamma(\alpha+i+1,\alpha) = {} & \left(\prod_{l=0}^{p}(\alpha+i-l)\right)\Gamma(\alpha+i-p,\alpha) \\ & + e^{-\alpha}\sum_{l_1=0}^{p}\left(\prod_{l_2=0}^{l_1-1}(\alpha+i-l_2)\right)\alpha^{\alpha+i-l_1}\end{aligned} \tag{4}$$

for $p = 0, 1, 2, \ldots$. The proof is straightforward using $\Gamma(s, x) = (s-1)\Gamma(s-1, x) + x^{s-1}e^{-x}$.

2.1 *Construction and general properties*

A random variable X has Laplace distribution if its density has the form: $f(x) = \frac{1}{2}e^{-|x|}$. The distribution is called Laplace, since it is connected with LAD regression introduced by Laplace (early work on LAD regression was done by Ruggiero Boscovich; see [20] and [8]). Looking at its construction, its other name, double exponential, is more useful. The exponential distribution is a special case of the gamma distribution $f(x) = \frac{1}{\Gamma(\alpha+1)}x^\alpha e^{-x}$, $x \geq 0$ for $\alpha = 0$ (see e.g. [11]). Therefore, a sort of symmetric double gamma density (with PDF equal to $\frac{1}{2\Gamma(\alpha+1)}|x|^\alpha e^{-|x|}$) would provide a generalized version of double exponential for $x \in \mathbb{R}$ and hence, quantile regression.

However we desire a unimodal distribution with log-concave density (hence not only unimodal but also strongly unimodal). Strong unimodality was introduced by [9]: a distribution is strongly unimodal if it is unimodal and its convolution with any other unimodal distribution is unimodal as well. In a very interesting result Ibragimov **[9]** could prove that a distribution is strongly unimodal if and only if its corresponding density is log-concave.

For $\alpha = 0$ (exponential distribution) there is no problem in this respect, since the mode of this distribution is at zero. In general the gamma distribution maximum is located at $x = \alpha$ which is non-zero of any $\alpha > 0$. Therefore, the PDF $\frac{1}{2\Gamma(\alpha+1)}|x|^\alpha e^{-|x|}$ would have two modes, one at $x = \alpha$ and one at $x = -\alpha$, which is not desirable. The construction of the SCDTG tries to solve this problem by following these steps:

(1) truncate gamma from below at $x = \alpha$: $f(x) \propto x^\alpha e^{-x}$, $x \geq \alpha$
(2) Construct double truncated gamma distribution: $f(x) \propto |x|^\alpha e^{-|x|}$, $|x| \geq \alpha$
(3) Shift each side by α toward zero to connect the two sides: $f(x) \propto (\alpha + |x|)^\alpha e^{-(\alpha+|x|)}$
(4) Find the normalizing constant: $f(x) = \frac{1}{2\Gamma(\alpha+1,\alpha)}(\alpha + |x|)^\alpha e^{-(\alpha+|x|)}$

Figure 1 (top) shows these steps graphically. We call the result the symmetric connected double truncated gamma distribution.

The SCDTG is different from to the bilateral Gamma distribution of [12]. The latter is the distribution of $X_1 - X_2$ when X_1 and X_2 are independent and have Gamma distribution.

Probability density function

A random variable X has symmetric connected double truncated gamma distribution with parameter α is written as $X \sim \text{SCDTG}(\alpha)$, if its density function is $f_\alpha(x)$ as follows:

$$f_\alpha(x) = \frac{1}{2\Gamma(\alpha+1,\alpha)}(\alpha+|x|)^\alpha e^{-(\alpha+|x|)} \tag{5}$$

where $\alpha \geq 0$ is the shape parameter. Figure 1 (bottom-left) shows the density for different values of α in comparison with Laplace and Gaussian PDF.

Cumulative distribution function

For the CDF we have $F_\alpha(x) = P(X \leq x) = \int_{-\infty}^{x} f_\alpha(u)\mathrm{d}u$, and one may derive:

$$F_\alpha(x) = \begin{cases} \dfrac{\Gamma(\alpha+1,\alpha-x)}{2\Gamma(\alpha+1,\alpha)} & \text{if} \quad x < 0, \\ 1 - \dfrac{\Gamma(\alpha+1,\alpha+x)}{2\Gamma(\alpha+1,\alpha)} & \text{if} \quad x \geq 0. \end{cases} \tag{6}$$

If working with upper incomplete gamma function is difficult, one may use the CDF of gamma distribution instead by using formula in (3). Figure 1 (bottom-middle) shows the CDF plots in comparison with the Laplace ($\alpha = 0$) and Gaussian CDF. The much heavier tail for larger α's is obvious.

Quantile function

As is the case for the gamma distribution, finding an explicit form of the quantile function is not possible, since one needs to compute the inverse of upper incomplete gamma function. However, using the connection with the gamma CDF and its inverse (which is available in all standard software packages) one may find the quantile function:

$$F_\alpha^{-1}(p_x) = x_{p,\alpha} = \begin{cases} \alpha - G_\alpha^{-1}(1 - \dfrac{2p_x\Gamma(\alpha+1,\alpha)}{\Gamma(\alpha+1)}) & \text{if} \quad p_x < \frac{1}{2}, \\ G_\alpha^{-1}(1 - \dfrac{2(1-p_x)\Gamma(\alpha+1,\alpha)}{\Gamma(\alpha+1)}) - \alpha & \text{if} \quad p_x \geq \frac{1}{2}. \end{cases} \tag{7}$$

where G_α^{-1} is the quantile function of a gamma CDF with shape parameter α. Figure 1 (bottom-right) shows the quantile function for some different α's in comparison with the Laplace and Gaussian cases. Using inverse transform sampling method, one may easily use formula (7) to generate random numbers from SCDTG:

(1) generate u from Uniform(0,1)
(2) $x = F_\alpha^{-1}(u)$
(3) take x as a random number generated from SCDTG with the parameter α.

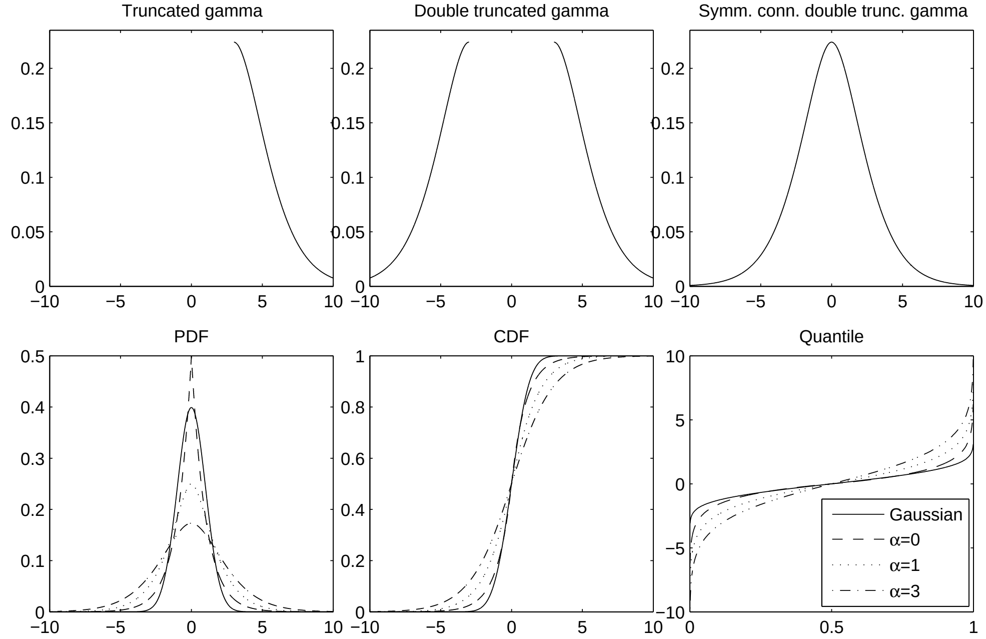

Figure 1. Top: connected double truncated gamma distribution construction. Bottom: PDF (Left), CDF (Middle) and quantile (Right) functions Gaussian and SCDTG for $\alpha = 0, 1, 3, 5$

Moments

As $f_\alpha(x)$ is symmetric around zero (i.e. it is an even function), all its odd moments are equal to zero. For the even moments ($k = 0, 2, 4, \ldots$) one has:

$$\mathbb{E}(X^k) = \frac{1}{\Gamma(\alpha+1,\alpha)} \sum_{i=0}^{k} \binom{k}{i} \alpha^{k-i} (-1)^i \Gamma(\alpha+i+1,\alpha). \tag{8}$$

This result is proven by splitting the integral in two parts:

$$\begin{aligned} \mathbb{E}(X^k) &= \int_{-\infty}^{+\infty} x^k f_\alpha(x) dx \\ &= \frac{1}{c} \left(\int_{-\infty}^{0} x^k (\alpha - x)^\alpha e^{-x} dx + \int_{0}^{\infty} x^k (\alpha + x)^\alpha e^{-x} dx \right) \end{aligned} \tag{9}$$

where $c = 1/2\Gamma(\alpha+1,\alpha)$. Calling the first integral I_1 and the second integral I_2 we make the change of variables $y = \alpha - x$ in I_1 such that: $I_1 = \int_\alpha^\infty (\alpha - y)^k y^\alpha e^{-y} dy$. Using the change of variable $y = \alpha + x$ in I_2, we have: $I_2 = \int_\alpha^\infty (y-\alpha)^k y^\alpha e^{-y} dy$. As for any even k we know that $(\alpha - y)^k = (y - \alpha)^k$, such that: $I_1 + I_2 = 2I_1$. Now using the binomial expansion we find:

$$\mathbb{E}(X^k) = \frac{1}{\Gamma(\alpha+1,\alpha)} \int_\alpha^\infty \left(\sum_{i=0}^{k} \binom{k}{i} \alpha^{k-i} (-y)^i \right) y^\alpha e^{-y} dy$$

and by changing integral and the summation, we have:

$$\begin{aligned}\mathbb{E}(X^k) &= \frac{1}{\Gamma(\alpha+1,\alpha)}\sum_{i=0}^{k}\binom{k}{i}\alpha^{k-i}(-1)^i\left(\int_{\alpha}^{\infty} y^{\alpha+i}e^{-y}dy\right)\\ &= \frac{1}{\Gamma(\alpha+1,\alpha)}\sum_{i=0}^{k}\binom{k}{i}\alpha^{k-i}(-1)^i\Gamma(\alpha+i+1,\alpha)\end{aligned}$$

which proves formula (8).

Using property in (4) for $p = i-1$ in formula (8) gives the following expression:

$$\begin{aligned}\mathbb{E}(X^k) = \sum_{i=0}^{k}\binom{k}{i}\alpha^{k-i}(-1)^i & \\ \times\left(\prod_{l=0}^{i-1}(\alpha+i-l) + \frac{e^{-\alpha}\sum_{l_1=0}^{i-1}\left(\prod_{l_2=0}^{l_1-1}(\alpha+i-l_2)\right)\alpha^{\alpha+i-l_1}}{\Gamma(\alpha+1,\alpha)}\right) & \end{aligned} \tag{10}$$

again for k even.

Characteristic function

The characteristic function (i.e. Fourier transform) of a density, as its name suggests, can characterize that density. $X \sim \text{SCDTG}(\alpha)$ if and only if its characteristic function $\phi_X(t) = \mathbb{E}(e^{itX})$ has the following form:

$$\begin{aligned}\phi_X(t) = \frac{1}{2\Gamma(\alpha+1,\alpha)}&\left(\frac{e^{+i\alpha t}}{(1+it)^{\alpha+1}}\Gamma(\alpha+1,\alpha(1+it))\right.\\ &\left.+\frac{e^{-i\alpha t}}{(1-it)^{\alpha+1}}\Gamma(\alpha+1,\alpha(1-it))\right)\end{aligned} \tag{11}$$

where we have used the analytic continuation of the incomplete gamma to complex numbers.

Properties of the log*-density*

If $X \sim \text{SCDTG}(\alpha)$, then $\log f_\alpha(x)$ is:

$$\log f_\alpha(x) = -\log 2\Gamma(\alpha,\alpha+1) + \begin{cases}\alpha\log(\alpha-x)-(\alpha-x) & \text{if} \quad x<0,\\ \alpha\log(\alpha+x)-(\alpha+x) & \text{if} \quad x\geq 0.\end{cases}$$

Obviously $\log f_\alpha(x)$ is a continuous function. Its derivative is:

$$\frac{\partial}{\partial x}\log f_\alpha(x) = \begin{cases}\dfrac{-\alpha}{\alpha-x}+1 & \text{if} \quad x<0,\\ \dfrac{\alpha}{\alpha+x}-1 & \text{if} \quad x\geq 0.\end{cases}$$

For any $\alpha > 0$ the left and right derivatives at $x = 0$ are equal. We conclude that for $\alpha > 0$ the function $\log f_\alpha$ is differentiable on the whole real line. As expected, for $\alpha = 0$ (the Laplace case), it is not differentiable at zero.

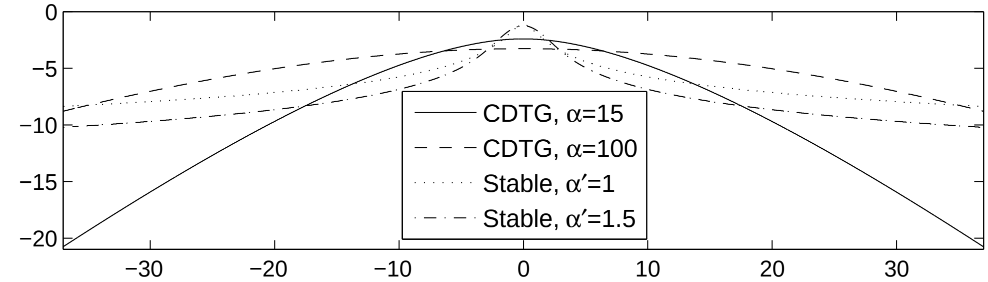

Figure 2. log-PDF of SCDTG with $\alpha = 15$ and $\alpha = 100$ and Stable distribution with $\alpha' = 1$ and $\alpha' = 1.5$. **The location parameter of the latter is** 0**, the scale parameter is** 1 **and the skewness parameter is** 0 **(symmetric case)). The** log**-density of the Stable distribution was calculated using parametrization-**0**, in Definition 1.7, p. 8 of [16].**

The second derivative of $\log f_\alpha$ is:

$$\frac{\partial^2}{\partial x^2} \log f_\alpha(x) = \begin{cases} \dfrac{-\alpha}{(\alpha - x)^2} & \text{if} \quad x < 0, \\ \dfrac{-\alpha}{(\alpha + x)^2} & \text{if} \quad x \geq 0. \end{cases}$$

which is negative $\forall x \in \mathbb{R}$. Therefore, $f_\alpha(x)$ is log-concave.

The boundedness of $-\partial_x^2 \log_\alpha f(x)$ implies that $-\partial_x \log_\alpha f(x)$ is Lipschitz continuous. If we set L the smallest Lipschitz constant of $-\partial_x \log_\alpha f(x)$, then $L = 1/\alpha$ for $\alpha > 0$.

Tails

We have already discussed the CDF of SCDTG(α) which, for a larger α, will have a larger probability in the tails. An interesting point to investigate would be how useful such a weight is. Figure 2 shows the log-PDF of the Stable distribution (see e.g. [18]) for $\alpha' = 1.5$ and $\alpha' = 1$ (which are not log-concave) and SCDTG for $\alpha = 15$ and $\alpha = 100$ (which are log-concave). As one may see, in log-concave cases as the tails of the curve of $\log f_\alpha(x)$ go higher, its peak comes a lot lower. The reason of such behavior lies in the log-concavity of the SCDTG density. Since for such densities there is no point of inflection in the log-density, bringing the tails higher means making the peak lower, which means having a flatter density. As one may see the log-PDF of SCDTG(100) is nearly a horizontal line compared to a log-density with points of inflection such as stables ones.

For example the symmetric stable distribution with $\alpha' = 1$ **(and location parameter** 0**, scale parameter** 1 **and skewness parameter** 0**)** is the standard Cauchy distribution. It has two inflection points at -1 and $+1$. These two points act like joints for the density. This means that the peak of the curve can stay high, while the tails also come higher at the point of inflection. As one may see, while a log-density with $\alpha' = 1$ has much longer tails than a log-density with $\alpha' = 1.5$ **(and identical location, scale and skewness parameter)**, their peaks are almost the same.

This discussion would suggest that, using a log-concave density with longer tail (larger α in our case) would not necessarily perform better than a log-concave density with shorter tails (smaller α in our case) for modeling data in presence of larger outliers. If one is interested in really heavy-tails, log-concave densities are not enough. For this reason we may call such log-concave densities *long-tail* instead of *heavy-tail*.

3. The asymmetric connected truncated double gamma distribution

Consider $f(x)$ a density symmetric around $x = 0$. Given two positive real numbers τ_1 and τ_2 ($\tau_1, \tau_2 > 0$) the skewed density $f_{\tau_1,\tau_2}(x)$ is defined as follows:

$$f_{\tau_1,\tau_2}(x) = \frac{2\tau_1\tau_2}{\tau_1+\tau_2}\begin{cases} f(\tau_1 x) & \text{if} \quad x < 0, \\ f(\tau_2 x) & \text{if} \quad x \geq 0. \end{cases} \tag{12}$$

Obviously, for $\tau_1 = \tau_2 = 1$, f_{τ_1,τ_2} is the same as $f(x)$ and for any $\tau_1 = \tau_2$, f_{τ_1,τ_2} is still symmetric. Larger τ_1 will produce a right skewed density, while larger τ_2 would lead to a left skewed one. Choosing $f(x)$ as the Laplace density and $\tau_1+\tau_2 = 1$, f_{τ_1,τ_2} would be equal to the asymmetric Laplace distribution which is used in quantile regression. In this sense if we use SCDTG distribution of previous section as $f(x)$ we would have a generalization of the asymmetric Laplace and hence a generalized quantile regression. Unlike the traditional quantile regression, this general version has a differentiable loss function for any $\alpha > 0$. It also shows heavier tails than the asymmetric Laplace.

Consider F as the CDF of f in (12), the corresponding CDF of f_{α,τ_1,τ_2} can be obtained as:

$$F_{\tau_1,\tau_2}(x) = \begin{cases} \dfrac{2\tau_2}{\tau_1+\tau_2}F(\tau_1 x) & \text{if} \quad x < 0, \\ \dfrac{\tau_2-\tau_1}{\tau_2+\tau_1} + \dfrac{2\tau_1}{\tau_1+\tau_2}F(\tau_2 x) & \text{if} \quad x \geq 0. \end{cases} \tag{13}$$

And for the quantile function one finds:

$$x_{p_{\tau_1,\tau_2}} = F^{-1}_{\tau_1,\tau_2}(p_x)\begin{cases} \dfrac{1}{\tau_1}F^{-1}(\dfrac{\tau_1+\tau_2}{2\tau_2}p_x) & \text{if} \quad p_x < \frac{\tau_2}{\tau_1+\tau_2}, \\ \dfrac{1}{\tau_2}(\dfrac{\tau_1+\tau_2}{2\tau_1}p_x - \dfrac{\tau_2-\tau_1}{2\tau_1}) & \text{if} \quad p_x \geq \frac{\tau_2}{\tau_1+\tau_2}. \end{cases} \tag{14}$$

For $\tau_1 \neq \tau_2$ we find that $-\log f_{\tau_1,\tau_2}(x)$ is just once differentiable. This won't cause any problem for using convex optimization algorithms, since we just need the first derivative.

THEOREM 3.1 *If $-\log f$ achieves its minimum at $x = 0$ then $-\log f_{\tau_1,\tau_2}(x)$ is convex if and only if $-\log f(x)$ is convex.*

Proof Set $g = -\log f$ and $\tilde{g} = -\log f_{\tau_1,\tau_2}$. Suppose g is convex. It suffices to prove that

$$\tilde{g}(x) \leq \frac{x_2 - x}{x_2 - x_1}\tilde{g}(x_1) + \frac{x - x_1}{x_2 - x_1}\tilde{g}(x_2)$$

for $x_1 \leq x \leq x_2$. We show this inequality in the case $x_1 \leq 0 \leq x_2$ as the cases $x_1 \leq x_2 \leq 0$ and $0 \leq x_1 \leq x_2$ follow trivially from the convexity of g.

Suppose $x \leq 0$ (the case $x \geq 0$ is treated analogously). As $\tilde{g}$ is convex for $x \leq 0$, we have:

$$\tilde{g}(x) \leq \frac{0 - x}{0 - x_1}\tilde{g}(x_1) + \frac{x - x_1}{0 - x_1}\tilde{g}(0).$$

On the other hand, we have that

$$\begin{aligned}\tilde{g}(0) &= \frac{x_2 - 0}{x_2 - x_1}\tilde{g}(0) + \frac{0 - x_1}{x_2 - x_1}\tilde{g}(0)\\ &\leq \frac{x_2 - 0}{x_2 - x_1}\tilde{g}(x_1) + \frac{0 - x_1}{x_2 - x_1}\tilde{g}(x_2)\end{aligned}$$

as $\tilde{g}$ achieves its minimum is $x = 0$. Combing the last two inequalities, we find:

$$\begin{aligned}\tilde{g}(x) &\leq \frac{0 - x}{0 - x_1}\tilde{g}(x_1) + \frac{x - x_1}{0 - x_1}\left[\frac{x_2 - 0}{x_2 - x_1}\tilde{g}(x_1) + \frac{0 - x_1}{x_2 - x_1}\tilde{g}(x_2)\right]\\ &= \frac{x_2 - x}{x_2 - x_1}\tilde{g}(x_1) + \frac{x - x_1}{x_2 - x_1}\tilde{g}(x_2).\end{aligned}$$

The converse follows the same lines. ■

Theorem 3.2 *$-\frac{\partial}{\partial x}\log f_{\tau_1,\tau_2}(x)$ is locally Lipschitz continuous if and only if $-\frac{\partial}{\partial x}\log f(x)$ is locally Lipschitz continuous.*

Proof Consider $k(x) = -\log f_{\tau_1,\tau_2}(x)$ and $l(x) = -\log f(x)$, then:

$$k'(x) = \frac{\partial}{\partial x}(-\log f_{\tau_1,\tau_2}(x)) = \begin{cases} \tau_1 l'(\tau_1 x) & \text{if} \quad x < 0,\\ \tau_2 l'(\tau_2 x) & \text{if} \quad x > 0.\end{cases}$$

Consider $x < 0$, $\|k'(x) - k'(y)\| = \tau_1\|l'(\tau_1 x) - l'(\tau_1 y)\| \leq \tau_1 L\|\tau_1 x - \tau_1 y\| = \tau_1^2 L\|x - y\|$. Where L is the Lipschitz constant of l'. Therefore, $L\tau_1^2$ is the Lipschitz constant for k'. The same is true for $x > 0$. In this case the Lipschitz constant for k' will be $L\tau_2^2$. Therefore, one finds for the Lipschitz constant of $k'(x)$ the value $\max(L\tau_1^2, L\tau_2^2)$ (for any real x). The converse is shown analogously. ■

3.1 *The loss function and generalized quantile regression*

If we consider the PDF in equation (5) and the PDF in equation (12), then the random variable X will have asymmetric connected double truncated gamma distribution with parameters α, τ_1 and τ_2, written as $X \sim \mathrm{ACDTG}(\alpha, \tau_1, \tau_2)$, if its PDF has the following form:

$$f_{\alpha,\tau_1,\tau_2}(x) = \frac{1}{2\Gamma(\alpha+1,\alpha)}\frac{2\tau_1\tau_2}{\tau_1+\tau_2}\begin{cases}(\alpha - \tau_1 x)^{\alpha} e^{-(\alpha - \tau_1 x)} & \text{if} \quad x < 0,\\ (\alpha + \tau_2 x)^{\alpha} e^{-(\alpha + \tau_2 x)} & \text{if} \quad x \geq 0.\end{cases} \tag{15}$$

Therefore, the corresponding loss function, $T(x - b) = -\log f(x - b)$, is:

$$T_{\alpha,\tau_1,\tau_2}(x - b) = \begin{cases}\alpha - \tau_1(x - b) - \alpha\log(\alpha - \tau_1(x - b)) & \text{if} \quad x < b,\\ \alpha + \tau_2(x - b) - \alpha\log(\alpha + \tau_2(x - b)) & \text{if} \quad x \geq b.\end{cases} \tag{16}$$

The following theorem shows how using this loss function will give a generalized version of the famous quantile regression of [13].

Theorem 3.3 *Consider the loss function in (16) with the density $f_{\alpha,\tau_1,\tau_2}(x)$, if $\tau_1 = 1 - \tau_2$ and $0 < \tau_2 < 1$ then $\lim_{\alpha\to 0}\arg\min_b \mathbb{E}(T_{\alpha,\tau_1,\tau_2}(X - b)) = x_{\tau_2}$ where x_{τ_2} is the $100\tau_2\%$ quantile of $F_{\alpha,\tau_1,\tau_2}(x)$, the CDF of (15).*

Proof By definition one has (skipping subscript α,τ_1,τ_2 in f for simplicity):

$$\begin{aligned}\mathbb{E}(T_{\alpha,\tau_1,\tau_2}(X-b)) &= \int_{-\infty}^{+\infty} T_{\alpha,\tau_1,\tau_2}(x-b)f(x)\mathrm{d}x \\ &= \int_{-\infty}^{b} (\alpha-\tau_1 x+\tau_1 b-\alpha\log(\alpha-\tau_1 x+\tau_1 b))\, f(x)\mathrm{d}x \\ &\qquad + \int_{b}^{+\infty} (\alpha+\tau_2 x-\tau_2 b-\alpha\log(\alpha+\tau_2 x-\tau_2 b))\, f(x)\mathrm{d}x.\end{aligned}$$

Now setting $\tau_1 = 1-\tau_2$ and simplifying integrals, one finds:

$$\mathbb{E}(T_{\alpha,\tau_1,\tau_2}(X-b)) = \alpha - \int_{-\infty}^{b} xf(x)\mathrm{d}x + \tau_2\mathbb{E}(X) + bF(b) - \tau_2 b$$
$$-\alpha\left(\int_{-\infty}^{b} \log(\alpha-x+b+\tau_2(x-b))f(x)\mathrm{d}x + \int_{b}^{+\infty}\log(\alpha+\tau_2 x-\tau_2 b)f(x)\mathrm{d}x\right).$$

The derivative of $\mathbb{E}\left(T_{\alpha,\tau_1,\tau_2}(X-b)\right)$ with respect to b is:

$$\begin{aligned}0 &= \frac{\partial}{\partial b}\mathbb{E}\left(T_{\alpha,\tau_1,\tau_2}(X-b)\right) \\ &= -\tau_2 + F(b) - \alpha\int_{-\infty}^{b}\frac{(1-\tau_2)f(x)\mathrm{d}x}{\alpha-(1-\tau_2)(x-b)} + \alpha\int_{b}^{+\infty}\frac{\tau_2 f(x)\mathrm{d}x}{\alpha+\tau_2(x-b)}\end{aligned} \tag{17}$$

Now, as

$$\begin{aligned}\alpha\int_{-\infty}^{b}\frac{(1-\tau_2)f(x)\mathrm{d}x}{\alpha-(1-\tau_2)(x-b)} &= \int_{-\infty}^{b}\frac{\alpha(1-\tau_2)f(x)\mathrm{d}x}{\alpha-(1-\tau_2)(x-b)} \\ &= \int_{-\infty}^{b}\frac{(\alpha-(1-\tau_2)(x-b))(1-\tau_2)f(x)\mathrm{d}x}{\alpha-(1-\tau_2)(x-b)} \\ &\qquad + \int_{-\infty}^{b}\frac{(1-\tau_2)(x-b)(1-\tau_2)f(x)\mathrm{d}x}{\alpha-(1-\tau_2)(x-b)} \\ &= \int_{-\infty}^{b}(1-\tau_2)f(x)\mathrm{d}x \\ &\qquad + \int_{-\infty}^{b}\frac{(1-\tau_2)(x-b)(1-\tau_2)f(x)\mathrm{d}x}{\alpha-(1-\tau_2)(x-b)},\end{aligned}$$

we find that:

$$\lim_{\alpha\to 0}\alpha\int_{-\infty}^{b}\frac{(1-\tau_2)f(x)\mathrm{d}x}{\alpha-(1-\tau_2)(x-b)} = \int_{-\infty}^{b}(1-\tau_2)f(x)\mathrm{d}x + \int_{-\infty}^{b}-(1-\tau_2)f(x)\mathrm{d}x = 0,$$

and analogously for the $\alpha\to 0$ limit of the second integral in (17).

Letting $\alpha \to 0$ in (17) one finds:

$$F(b) = \tau_2 \Rightarrow \int_{-\infty}^{b} F(x)dx = \tau_2 \Rightarrow b = x_{\tau_2},$$

where $x_{\tau_2} = 100\tau_2\%$ quantile of $F(x)$. ■

Solving (17) for $\alpha > 0$ would give a generalized version of the quantiles as a measure of centrality.

4. The generalized quantile regression model and its penalized version

In Theorem 3.3 we have seen that using the ACDTG as the error distribution will produce a generalized version of the quantile regression. Here we will introduce the model and estimate its parameters. First we may consider the simplest model which is a line through the origin. It will be extended to a general case later on in this section.

4.1 *Simple regression model through origin*

First we consider the simple one variable model $y = \beta x + \epsilon$, where y is the response variable, x is the regressor and $\epsilon \sim \text{SCDTG}(\alpha)$. The loss is as follows:

$$T(\beta) = \sum_{i=1}^{n} |y_i - \beta x_i| - \alpha \sum_{i=1}^{n} \log(\alpha + |y_i - \beta x_i|).$$

The minimizer of $T(\beta)$ is called $\hat{\beta}$ which is the estimator of the parameter β. Clearly the above loss function has a term $-\alpha \sum_i \log(\alpha + |y_i - \beta x_i|)$ extra, as compared to the LAD loss function.

As T is differentiable (for $\alpha > 0$), the minimum of the above function is reached at the root (with respect to β) of the following equation:

$$\sum_{i=1}^{n} \frac{\hat{\beta} x_i^2 - x_i y_i}{\alpha + |y_i - \hat{\beta} x_i|} = 0. \tag{18}$$

As is well-known, the minimizer of the LS loss function is the root of $\sum_i \hat{\beta} x_i^2 - x_i y_i = 0$. Thus, the difference of these two is obvious. Although solving equation (18) is not as straightforward as solving the LS case, standard methods can solve it numerically.

We have seen the similarities and differences of this loss function and its minimizer with LS and LAD. To study the performance of this method compared to LAD and LS, we have performed a simulation study. The model $y = 3x + \epsilon$ was used. The error distribution is made as a 5%-contaminated Gaussian distribution with the standard deviation (STD) equal to 10% of the STD of βx. For the 5%-contaminated part, two scenarios are used: firstly, 5% is generated from Gaussian with mean randomly chosen in the interval $[0, 15]$ and secondly 5% is generated from Gaussian with mean randomly chosen in the interval $[15, 30]$. Two sample sizes have been considered (100 and 20), to study the performance for both large and small sample sizes. For each case the LS and LAD estimates are computed. Also the estimator based on the generalized quantile regression is computed for the α resulting in the best β (assuming we know the real $\beta = 3$). In addition, to study the performance of the generalized version as an approximation of the traditional quantile regression the value $\alpha = 0.0001$ is considered. For each case, 500 replications are made. Table 1 shows the results.

As one may see for large or small sample sizes there exists an α for which the GQR estimates becomes consistent. Also GQR with $\alpha = 0.0001$ provides an almost

Table 1. Simulation results for the simple regression estimates $\hat{\beta}$ comparing the ordinary least squares (OLS), least absolute deviations (LAD), generalized quantile regression (GQR, for the best value of α), and approximate quantile regression (AQR, for $\alpha = 0.0001$) as an approximation of LAD. The line with (α) gives the value of the best α found from GQR.

5%- contaminated $[0, 15]$

	sample size 20			sample size 100		
Method	Mean	Median	STD	Mean	Median	STD
OLS	2.9975	2.9946	0.0757	2.9280	2.9313	0.2251
LAD	2.9983	2.9913	0.0909	2.9975	2.9972	0.0394
GQR	3.0000	3.0000	0.0001	3.0000	3.0000	0.0001
(α)	(10.2007)	(10.2800)	(5.6195)	(7.9164)	(6.7100)	(5.9778)
AQR	2.9983	2.9913	0.0909	2.9975	2.9972	0.0394

5%- contaminated $[15, 30]$

	sample size 20			sample size 100		
Method	Mean	Median	STD	Mean	Median	STD
OLS	3.0061	3.0053	0.0774	3.1397	3.1712	0.4911
LAD	3.0055	3.0056	0.0972	3.0034	3.0014	0.0413
GQR	3.0000	3.0000	0.0001	3.0000	3.0000	0.0001
(α)	(9.7420)	(9.4300)	(5.9614)	(6.7390)	(4.8700)	(5.7760)
AQR	3.0055	3.0056	0.0972	3.0034	3.0014	0.0413

exact approximation of the QR estimates. As we have discussed, it seems for a larger outlier, some smaller (not very small) α's give better results.

4.2 *The general model*

Using ACDTG as the error distribution leads to a generalized version of the traditional quantile regression. If we consider the response variable y, the regressors $X = (x_1, \ldots, x_p)$, and the linear model $y = X\beta + \epsilon$, with $\epsilon \sim \text{ACDTG}(\alpha, \tau_1, \tau_2)$, then to find $\hat{\beta}$ using a sample of size n, one may solve the following problem:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} T\left((y - X\beta)_i\right) \tag{19}$$

where T is the loss function in (16). **As in traditional quantile regression, the loss function in (16) is applied to the error; thus, the resulting regression model is robust to outliers but not necessarily robust to leverage points.**

The sparse penalized model

As it was already remarked, nowadays high dimensional problems are very common. This may lead to problems, especially when the number of variables is much larger than the number of observations (i.e. ill-posed problems). A way of dealing with these problems is penalizing the loss function in a way that leads to a sparse model vector $\hat{\beta}$. Lasso of [21], which penalized the LS loss function with the ℓ_1-norm of the parameters vector is the most famous solution of this type. Here we may consider the ℓ_1-penalized GQR. It is defined as follows:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} T_{\alpha,\tau_1,\tau_2}\left((y - X\beta)_i\right) + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{20}$$

where $\lambda \geq 0$ is the penalization parameter (which should also be determined). A larger λ leads to a more sparse $\hat{\beta}$ and $\lambda = 0$ is the non-penalized problem.

There are different approaches for selecting the appropriate penalization parameter λ for a penalized model. One may use Akaike's information criterion (AIC) [1], the Bayesian information criterion (BIC) [19], Mallows's C_p, [14], or its outlier robust version, [17], cross validation, [15], and its generalized version [10]. An extensive overview of most of the existing model selection techniques is given in [5].

Solution method

In formula (20) one has a convex differentiable loss function with a convex (non-differentiable) penalty function. In such setting one can use techniques of convex optimization. In [2] an efficient algorithm (so called fast iterative soft threshold algorithm: FISTA) was introduces to solve just such a problem. FISTA consists of the following simple steps:

- Input: $\tilde{L}$, a Lipschitz constant of $\nabla\left(\sum_i T_{\alpha,\tau_1,\tau_2}\left((X\beta - y)_i\right)\right)$
- Step 0: take $\omega^{(1)} = \beta^{(0)} \in \mathbb{R}^p$, $t^{(1)} = 1$
- Step k ($k \geq 1$): Compute:

$$\begin{aligned} \beta^{(k)} &= p_{\tilde{L}}(\omega^{(k)}), \qquad t^{(k+1)} = \frac{1 + \sqrt{1 + 4{t^{(k)}}^2}}{2} \\ \omega^{(k+1)} &= \beta^{(k)} + \left(\frac{t^{(k)} - 1}{t^{(k+1)}}\right)(\beta^{(k)} - \beta^{(k-1)}). \end{aligned} \tag{21}$$

In this algorithm $p_{\tilde{L}}(\omega)$ is defined as

$$p_{\tilde{L}}(\omega) = S_{\lambda/\tilde{L}}\left(y - \frac{1}{\tilde{L}}\nabla T(\omega)\right) \tag{22}$$

where S_σ is the (non linear) soft thresholding operator:

$$S_\sigma(\beta) = \begin{cases} \beta - \sigma & \text{if} \quad \beta > \sigma \\ 0 & \text{if} \quad |\beta| \leq \sigma \\ \beta + \sigma & \text{if} \quad \beta < -\sigma \end{cases} \tag{23}$$

for $\sigma \geq 0$, see e.g. [6]. If L is the Lipschitz constant for ∇T, then $\tilde{L} = \|X'X\|_2 L$, where $\|X'X\|_2$ is the spectral norm of $X'X$, i.e., its largest eigenvalue. As we have computed the derivative of T (∇T) and also the Lipschitz constant for it, applying FISTA to our problem is straightforward.

The FISTA algorithm was implemented in Matlab. It is a simple algorithm and it may be implemented in any other programming environment such as e.g. R.

Simulation study

In order to study the performance of FISTA for ℓ_1-penalized GQR, a simulation study (the same as the one for one variable model) is done. Here we consider an 85%-sparse parameters vector with 150 components. Its non-zero components are generated from a Gaussian distribution with zero mean and standard deviation 7. They are randomly allocated to 15% nonzeros of the parameters vector. Two sample sizes are taken: $n = 20$ and $n = 100$. Thus, we always have more variables

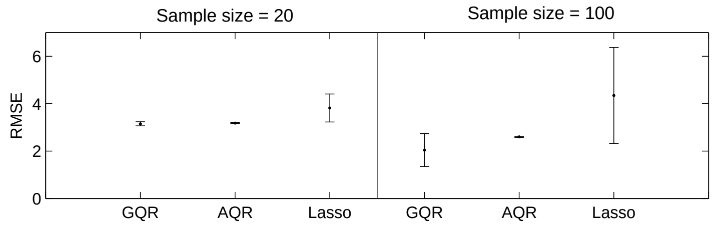

Figure 3. Simulation results for the penalized regression estimates comparing the Lasso, the generalized quantile regression GQR (for the best α) and the approximate quantile regression AQR (with fixed $\alpha = 0.0001$)

than observations, once with a large sample size and once with a very small sample size.

Here we consider a more sophisticated error than for the simple regression model of Section 4.1. Several error distributions are used: Gaussian, Laplace and Student's-t with 3 degrees of freedom. Numerical experiments are performed where the contamination percentage is equal to 5%, 10%, and 15%. The location parameter of the contaminated part is chosen randomly in the intervals $[0, 100]$ or $[100, 200]$, and its scale parameter is set as 10% or 30% of the standard deviation of y. Studying all these possibilities separately would produce 36 different cases; the presentation of the results would require a lot of space, whether in the form of tables of graphs. Therefore, instead of considering all the combinations separately and performing 100 replications for each of them, 5000 replications were done in total with each replication one of the 36 cases (chosen at random). This way the performance of the proposed model is examined for many different error types, while the results are kept simple. The high number of replications (5000) ensures that all of the 36 cases will occur many times.

The Lasso and GQR (for the best α and for $\alpha = 0.0001$) models are considered. The error ϵ on the data y is considered to be known. The penalty parameter $\lambda > 0$ is chosen each time such that $T(\epsilon) \approx T(e_\lambda)$, where T is the loss function in (16) and $e_\lambda = y - X\beta$ is the model residual for the chosen λ. The number of iterations in the FISTA algorithm is equal to 1000. Figure 3 presents the results.

As one may see in Figure 3, for both sample sizes GQR shows better performance than Lasso. But when the sample size is small, approximated quantile regression (i.e. GQR with α close to zero) and GQR for the best α give almost the same results. By increasing the sample size, one may easily see the better performance of GQR when the α is chosen, as compared to a near zero α. This result also can be seen in the estimated α for small and large sample sizes. For small sample size, the mean, median and standard deviation of α are obtained as $(0.2109, 0.0001, 0.8863)$, respectively, while for large sample size they are $(1.7182, 0.1001, 2.0910)$.

Real data

The simulation study of the previous subsection showed the better performance of the proposed method. In this subsection we analyze a real data set and examine the performance of the method. The Current Population Survey (CPS) is a survey of households conducted by the Bureau of Census for the Bureau of Labor Statistics in United States. The data we consider is taken from [4] and consist of a random sample

of 534 persons from the CPS-1985. Table 2 lists the variables in this data set.

Table 2. Variables in wage data

Variable	Description
Education	number of years of education
South	1 = person lives in the South, 0 = person lives elsewhere
Sex	1 = female, 0 = male
Experience	number of years of work experience
Union	1 = union member, 0 = not a union member
Age	age in years
Race	$(0, 0)$ = other, $(1, 0)$ = black, $(0, 1)$ = white
Occupation	$(0, 0, 0, 0, 0)$ = other, $(1, 0, 0, 0, 0)$ = management, $(0, 1, 0, 0, 0)$ = sales, $(0, 0, 1, 0, 0)$ = clerical, $(0, 0, 0, 1, 0)$ = service, $(0, 0, 0, 0, 1)$ = professional,
Sector	$(0, 0)$= other, $(1, 0)$= manufacturing, $(0, 1)$= construction
Marriage	0 = unmarried, 1 = married
Wage	wage in dollars per hour (response variable)

For analyzing the effective factors on wage it is interesting to analyze the effect of different factors on the whole conditional distribution of wage (lower, average and higher wages) and not only its center. While it's not possible to do such analysis using Lasso, with GQR or quantile regression we have the possibility of using an asymmetric loss function by choosing different τ's. We have therefore studied conditional quartiles of the wage data ($\tau = 0.25, 0.5, 0.75$), and have performed (non penalized) GQR and quantile regression [13] with wage as the response variable. The shape parameter α is found in the same manner as before. In order to study the prediction precision, three-quarter of the data were used to train the model, and the rest were used to test the prediction precision. The RMSE's for the GQR model are obtained as 5.2033, 4.5495, and 4.7274 for $\tau = 0.25$, $\tau = 0.5$, and $\tau = 0.75$, respectively. While the QR model gives RMSE's equal to 5.4369, 4.7794, and 5.0985 for these quartiles. As one may see in this example, for all τ's GQR gives a more precise prediction than quantile regression. As GQR includes QR as an special case for $\alpha = 0$, GQR would always give results at least as good as QR.

5. Conclusions

In this paper a new class of probability distributions, the so-called connected double truncated gamma distributions have been introduced. Many properties of both its symmetric an asymmetric versions have been studied.

Using it as the error distribution in a linear model will give a generalized quantile regression which combines desirable properties of LS and QR, i.e. it has a differentiable convex loss function and it is robust to the outliers and to the asymmetric error. An efficient fast algorithm is adapted to solve the penalized version of the linear model, considering the SCDTG and ACDTG error. The immediate use of such a model is approximating a QR in a fast and efficient manner.

One may extends the PDF in (15) to a density with location and scale parameter as well:

$$f_{\alpha,\tau_1,\tau_2,\mu,\sigma}(x) = \frac{1}{2\Gamma(\alpha+1,\alpha/\sigma)}\frac{2\tau_1\tau_2}{\tau_1+\tau_2}\begin{cases}(\alpha-\tau_1\dfrac{x-\mu}{\sigma})^{\alpha}e^{-(\alpha-\tau_1\frac{x-\mu}{\sigma})} & \text{if} \quad x<\mu, \\ (\alpha+\tau_2\dfrac{x-\mu}{\sigma})^{\alpha}e^{-(\alpha+\tau_2\frac{x-\mu}{\sigma})} & \text{if} \quad x\geq\mu\end{cases}$$

where $\sigma > 0$ and $\mu \in \mathbb{R}$ are scale and location parameters, respectively. Therefore, ACDTG can be considered as a class of log-concave densities with shape, scale, location and skewness parameters. Such density would be very flexible to model different types of data. A later study would be concerned with the distributional properties of this class with parameters other than only α. Also the idea of connected double truncated distributions can be extended to any other distribution with its support on $\mathbb{R}^+$. A direct example would be the Chi-square distribution. As is well-known, the Chi-square distribution is connected with the gamma distribution, and one may derive the SCDT-Chi-square(α) with ν degrees of freedom as follows:

$$f_{\alpha,\nu}(x) = \frac{1}{4\Gamma(\nu/2,(\nu-2)/4)}\left(\nu/2-1+\frac{|x|}{2}^{\nu/2-1}e^{-\left(\nu/2-1+\frac{|x|}{2}\right)}\right)$$

for $x \in \mathbb{R}$.

Acknowledgements(s)

I.L. is a research associate of the F.R.S.-FNRS (Belgium). This research was supported by VUB GOA-062 and by the FWO-Vlaanderen grant G.0564.09N.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974. System identification and time-series analysis.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-threshold algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, 2009.

[3] R. Bellman. *Dynamic programming.* Princeton University Press, Princeton, N. J., 1957.

[4] E. Berndt. *The practice of econometrics: classic and contemporary.* Addison Wesley Longman Publish, New York, 1991.

[5] G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging.* Cambridge University Press, Cambridge, 2008.

[6] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[7] J. Fan and R. Li. Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *International Congress of Mathematicians. Vol. III*, pages 595–622. Eur. Math. Soc., Zürich, 2006.

[8] R. W. Farebrother. Studies in the history of probability and statistics. XLII. Further details of contacts between Boscovich and Simpson in June 1760. *Biometrika*, 77(2):397–400, 1990.

[9] I. A. Ibragimov. On the composition of unimodal distributions. *Teor. Veroyatnost. i Primenen.*, 1:283–288, 1956.

[10] M. Jansen, M. Malfait, and A. Bultheel. Generalized cross validation for wavelet thresholding. *Signal Processing*, 56:33–44, 1997.

[11] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions. Vol. 1.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1994. A Wiley-Interscience Publication.

[12] U. Küchler and S. Tappe. On the shapes of bilateral Gamma densities. *Statistics & Probability Letters*, 78(15):2478–2484, 2008.

[13] R. Koenker and G. Bassett. Regression quantiles. *Econometrics*, 46:33–50, 1978.

[14] C. L. Mallows. Some comments on c_p. *Technometrics*, (15):661–675, 1973.

[15] G. P. Nason. Wavelet shrinkage using cross-validation. *Journal of The Royal Statistical Society, Series B*, 58:463–479, 1996.

[16] J. P. Nolan. *Stable Distributions - Models for Heavy Tailed Data*. Birkhauser, Boston, 2012. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

[17] E. Ronchetti and R. G. Staudte. A robust version of Mallows' C_P. *J. Amer. Statist. Assoc.*, 89(426):550–559, 1994.

[18] G. Samorodnitsky and M. S. Taqqu. *Stable non-Gaussian random processes. Stochastic models with infinite variance.* Stochastic Modeling. Chapman & Hall, New York, 1994.

[19] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[20] S. M. Stigler. Studies in the history of probability and statistics. XXXIV. Napoleonic statistics: the work of Laplace. *Biometrika*, 62(2):503–517, 1975.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.