INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

1014

COMPOSITE LOGNORMAL-PARETO MODEL WITH RANDOM THRESHOLD

PIGEON, M. and M. DENUIT

COMPOSITE LOGNORMAL-PARETO MODEL WITH RANDOM THRESHOLD

Mathieu Pigeon & Michel Denuit
Institut de statistique, biostatistique et sciences actuarielles
Université Catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium

March 10, 2010

Abstract

This paper further considers the composite Lognormal-Pareto model proposed by Cooray & Ananda (2005) and suitably modified by Scollnik (2007). This model is based on a Lognormal density up to an unknown threshold value and a Pareto density thereafter. Instead of using a single threshold value applying uniformly to the whole data set, the model proposed in the present paper allows for heterogeneity with respect to the threshold and let it vary among observations. Specifically, the threshold value for a particular observation is seen as the realization of a positive random variable and the mixed composite Lognormal-Pareto model is obtained by averaging over the population of interest. The performance of the composite Lognormal-Pareto model and of its mixed extension is compared using the well-known Danish fire losses data set.

Key words and phrases: Mixture, loss model, Danish fire losses, extreme value.

1 Introduction and Motivation

In nonlife insurance business, large losses sometimes occur: costs faced by insurance companies often originate from a mix of moderate and large claims. However, no standard parametric model seems to emerge as providing an acceptable fit to both small and large losses. Several distributions for modelling positive and right–skewed data arising in the insurance industry have been proposed by actuaries. For an extensive presentation, the reader is referred to Klugman, Panjer & Willmot (2004). Let us also mention the augmented mixture of exponentials distribution proposed in Klugman & Rioux (2004). When the main interest is in the tail of loss severity distributions, it is essential to have a good model for the largest claims. Distributions providing a good overall fit can be particularly bad at fitting the tails. Empirical actuarial analyses usually proceed in two steps: large losses are first isolated and then modelled separately.

Usually, being a large claim means exceeding some threshold, depending on the portfolio under study. Extreme Value Theory and Generalized Pareto distributions can be used to set the value of this threshold, as described in Cebrian, Denuit & Lambert (2003). Specifically, graphical tools including the Pareto index plot and the Gertensgarbe plot can be used to estimate the threshold defining the large losses. In the former case, the maximum likelihood estimator of the Pareto tail parameter is computed for increasing thresholds until it becomes approximately constant. The Gertensgarbe plot is based on the assumption that the optimal threshold can be found as a change point in the ordered series of claim costs and that the change point can be identified by mean of a sequential version of the Mann-Kendall test as the intersection point between a normalized progressive and retograde rank statistics. Once the threshold defining large claims has been selected, losses above this threshold are modelled using the Generalized Pareto distribution. Different models can be used to describe the behavior of the moderate claims (i.e. claims with an incurred cost less than the threshold), including Gamma, Inverse Gaussian and Lognormal distributions.

Another approach is proposed by Cooray & Ananda (2005) who combined a Lognormal probability density function together with a Pareto one. Specifically, these authors introduced a two-parameter smooth continuous composite Lognormal-Pareto model that is a two-parameter Lognormal density up to an unknown threshold value and a two-parameter Pareto density for the remainder. Continuity and differentiability are imposed at the unknown threshold to ensure that the resulting probability density function is smooth, reducing the number of parameters from 4 to 2. The resulting two-parameter probability density function is similar in shape to the Lognormal density, yet its upper tail is thicker than the Lognormal density (and accomodates for the large losses observed in liability insurance). This approach clearly outperforms the classical two-step strategy described above, in that all the parameters (including the threshold) are estimated in the same model. However, the proposal made by Cooray & Ananda (2005) has been amended by Scollnik (2007) who pointed out that this model fixes the proportion of large claims, which appears very restrictive. Let us also mention the work by Tancredi, Anderson & O'Hagan (2006) who modelled data with a distribution composed of a piecewize constant density from a low threshold up to an unknown end point and a Generalized Pareto distribution for the remaining tail part.

In the composite Lognormal-Pareto models proposed by Cooray & Ananda (2005) and

Scollnik (2007), a threshold parameter is estimated from the data and the exceedances over this threshold obey the Pareto distribution. Estimating the threshold together with the other parameters account for threshold uncertainty but assuming a unique threshold value applying to all the claims may appear quite unrealistic. In this paper, we allow for heterogeneity with respect to the threshold, and we treat it as a random variable.

The structure of this paper is as follows. In Section 2, we describe existing composite Lognormal–Pareto models of Cooray & Ananda (2005) and Scollnik (2007). In Section 3, we introduce a new mixed composite Lognormal–Pareto model which has a random threshold and give its basic properties. In Section 4, we compare the performance of our mixed composite model, existing composite models and classical distributions based on well–known Danish fire insurance loss data set. The final Section 5 concludes.

2 Composite Lognormal–Pareto models

2.1 Cooray–Ananda's model

Let

$$f_1(x) = \frac{1}{\sqrt{2\pi}x\sigma} \exp\left(-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right), \qquad x > 0 \tag{2.1}$$

be a two–parameter Lognormal density function and

$$f_2(x) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}}, \qquad x > \theta, \tag{2.2}$$

be a two–parameter Pareto density function. Let Φ denote the cumulative distribution function of the standard Normal distribution. The composite Lognormal–Pareto probability density function defined by Cooray & Ananda (2005) is given by

$$f(x) = \begin{cases} \psi\dfrac{1}{\Phi(k)}f_1(x), & 0 < x \leq \theta, \\ (1-\psi)f_2(x), & \theta < x < \infty, \end{cases} \tag{2.3}$$

with $k \approx 0.37224$, $\psi = \Phi(k)/(1+\Phi(k)) \approx 0.39215$ and

$$\frac{\ln(\theta)-\mu}{\sigma} = \alpha\sigma = k.$$

The probability density function (2.3) has a scale parameter or *threshold* ($\theta > 0$) and a shape parameter or *tail index* ($\alpha > 0$).

As noted by Scollnik (2007), this model is very restrictive because weights ψ and $1-\psi$ are fixed and *a priori* known. Whatever the data set under study, exactly 39.215% of the observations are expected to fall below θ. Moreover, parameters of Lognormal portion of the distribution are determined as function of the values of threshold and tail index, which can lead to poor adjustment to data. Finally, threshold is considered to be fixed for all observations which may not be realistic in an actuarial context.

2.2 Scollnik's models

Scollnik (2007) developed a second composite Lognormal–Pareto model in order to fix some problems identified in Cooray–Ananda's model (2.3). Let f_1 and f_2 be the Lognormal and Pareto density functions given by (2.1) and (2.2), respectively. The composite Lognormal–Pareto probability density function defined by Scollnik (2007) is given by

$$f(x) = \begin{cases} r\dfrac{1}{\Phi(\alpha\sigma)} f_1(x), & 0 < x \leq \theta, \\ (1-r) f_2(x), & \theta < x < \infty, \end{cases} \tag{2.4}$$

with

$$r = \frac{\sqrt{2\pi}\alpha\sigma\Phi(\alpha\sigma)\exp\left(\frac{1}{2}(\alpha\sigma)^2\right)}{\sqrt{2\pi}\alpha\sigma\Phi(\alpha\sigma)\exp\left(\frac{1}{2}(\alpha\sigma)^2\right)+1} \tag{2.5}$$

and

$$\frac{\ln(\theta)-\mu}{\sigma} = \alpha\sigma. \tag{2.6}$$

The probability density function (2.4) is defined by means of a threshold ($\theta > 0$), a tail index ($\alpha > 0$) and a *small loss parameter* ($\sigma > 0$). One can observe in (2.4) that, contrarily to ψ and $1-\psi$ in (2.3), the mixing weights r and $1-r$ are no more fixed and known values.

Scollnik (2007) also introduced an alternative composite Lognormal–Pareto model in which the generalized Pareto distribution (GPD) with density function

$$h(x) = \frac{\alpha(\lambda+\theta)^\alpha}{(\lambda+x)^{\alpha+1}}, \qquad x > \theta,\ \theta > 0,\ \alpha > 0 \text{ and } \lambda > -\theta, \tag{2.7}$$

is used above the threshold. Theoretical grounds supporting the use of the GPD can be found in Embrechts, Klüppelberg & Mikosch (1997). The resulting probability density function is then given by

$$f(x) = \begin{cases} r\dfrac{1}{\Phi(\nu)} f_1(x), & 0 < x \leq \theta, \\ (1-r) h(x), & \theta < x < \infty. \end{cases} \tag{2.8}$$

with

$$r = \frac{\sqrt{2\pi}\alpha\theta\sigma\Phi(\nu)\exp\left(\frac{1}{2}\nu^2\right)}{\sqrt{2\pi}\alpha\theta\sigma\Phi(\nu)\exp\left(\frac{1}{2}\nu^2\right)+\lambda+\theta}. \tag{2.9}$$

and

$$\frac{\ln(\theta)-\mu}{\sigma} = \left(\frac{\alpha\theta-\lambda}{\lambda+\theta}\right)\sigma = \nu. \tag{2.10}$$

For the sake of simplicity, we will develop our new model from definition (2.4). A similar analysis could be performed on (2.8).

3 Mixed composite Lognormal-Pareto model

3.1 Definition

The main aim of this paper is to introduce a new composite Lognormal–Pareto model based on Scollnik's model (2.4). Let $X_1, \ldots, X_n$ denote a random sample of size n. Now we assume that each observation may have its own threshold $\theta_1, \ldots, \theta_n$. In fact, we consider $\theta_1, \ldots, \theta_n$ as realizations of some non-negative random variable Θ with cumulative distribution function $G(\cdot)$. Fisrt, we develop main properties for a general random variable Θ then we give some examples.

Using (2.1), (2.2) and (2.5), the mixed composite Lognormal–Pareto density function is given by

$$f(x) = (1-r)\int_0^x f_2(x)\, dG(\theta) + r\int_x^\infty \left(\frac{1}{\Phi(\alpha\sigma)}\right) f_1(x)\, dG(\theta), \tag{3.1}$$

with

$$\frac{\ln(\theta)-\mu}{\sigma} = \alpha\sigma. \tag{3.2}$$

Let X be a random variable with probability density function given by (3.1). For $0 < k < \alpha$, the k^{th} raw moment of X is given by

$$\mathbb{E}[X^k] = \left((1-r)\left(\frac{\alpha}{\alpha-k}\right) + r\left(\frac{1}{\Phi(\alpha\sigma)}\right)(1-\Phi(\sigma(k-\alpha)))\exp\left(-\alpha\sigma^2 k + k^2\sigma^2/2\right)\right)\mathbb{E}[\Theta^k].$$

Since the upper tail distribution is especially useful for reinsurance purposes, it is interesting to derive the expresssion of the stop-loss transform

$$\pi_X(d) = \mathbb{E}[\max(X-d, 0)],$$

corresponding to the mixed composite Lognormal-Pareto model. The stop-loss transform π_X is a continuous convex function that is strictly decreasing in the retention d as long as $F_X(d) < 1$. Furthermore $\lim_{d\to+\infty}\pi_X(d) = 0$ as long as the expectation is finite. If X is non-negative then $\pi_X(0) = \mathbb{E}[X]$. Using model (3.1), we get

$$\begin{aligned}\pi_X(d) &= \int_d^\infty (1-F_X(x))\, dx \\ &= \mathbb{E}[X] - d(1-F_X(d)) - \int_0^d x f_X(x)\, dx.\end{aligned}$$

In the next two subsections, we present two examples of distributions for the threshold.

3.2 Gamma distributed threshold

We develop now basic properties of model (3.1) with $\Theta \sim \mathcal{G}am(\beta, \lambda)$, that is, the probability density function of Θ is given by

$$g(\theta; \beta, \lambda) = \frac{\lambda^\beta \theta^{\beta-1}\exp\big(-\lambda\theta\big)}{\Gamma(\beta)}, \qquad \theta > 0, \tag{3.3}$$

for $\beta > 0$ and $\lambda > 0$. The probability density function defined in equation (3.1) is

$$\begin{aligned} f(x) = & \frac{r}{\Phi(\alpha\sigma)y\sigma}\int_y^\infty \phi\left(\frac{\ln(y)-\ln(\theta)+\alpha\sigma^2}{\sigma}\right) g(\theta;\beta,\lambda))\,d\theta \\ & + (r+1)g(y;\beta,\lambda) \\ & - (1-r)\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)}\right)\left(\frac{-\alpha}{\lambda^\alpha y^{\alpha+1}}G(y;\alpha+\beta,\lambda)+\frac{1}{(\lambda y)^\alpha}g(y;\alpha+\beta,\lambda)\right), \end{aligned}$$

where $G(x;a,b)$ denote the Gamma cumulative distribution function with shape parameter a and scale parameter $1/b$ evaluated at x and $\phi(x)$ denote the probability density function of the standard Normal distribution evaluated at x. It can be shown that $f(x)$ is continuous and differentiable on the half-positive real line $(0,\infty)$.

Let us now illustrate the shape of the probability density function (3.1) compared to (2.4). To this end, let us display the graph of the probability density function (2.4) with parameter values $\alpha = 1.5$, $\sigma = 1$ and $\theta = 50$ together with the probability density function (3.1) first with parameter values $\alpha = 1.5$, $\sigma = 1$, $\beta = 50$ and $\lambda = 1$, and second with parameter values $\alpha = 1.5$, $\sigma = 1$, $\beta = 5$ and $\lambda = 0.1$. This is done in Figure 3.1. One can note that model (3.1) reduces to model (2.4) as $\mathbb{E}[\Theta] \to \theta$ and $\mathbb{V}[\Theta] \to 0$. Therefore, the probability density function (2.4) corresponds to $\mathbb{E}[\Theta] = 50$ and $\mathbb{V}[\Theta] = 0$ whereas for the probability density function (3.1), we still have $\mathbb{E}[\Theta] = 50$ but with a moderate variance $\mathbb{V}[\Theta] = 50$ in the former case and with a larger variance $\mathbb{V}[\Theta] = 500$ in the latter case. The other parameters remain identical in the three cases. Figure 3.1, thus, illustrates the effect of increasing the degree of heterogeneity in the thresholds within the sample data. Increasing $\mathbb{V}[\Theta]$ produces a higher peak and fatter tails, as expected. This clearly shows the difference between the probability density function (3.1) compared to (2.4).

For $0 < k < \alpha$, the k^{th} raw moment of X is given by

$$\begin{aligned} \mathbb{E}[X^k] = & (1-r)\left(\frac{\alpha}{\alpha-k}\right)\left(\frac{\Gamma(\beta+k)}{\Gamma(\beta)\lambda^k}\right) \\ & + r\left(\frac{1}{\Phi(\alpha\sigma)}\right)(1-\Phi(\sigma(k-\alpha)))\exp\left(-\alpha\sigma^2 k + k^2\sigma^2/2\right)\left(\frac{\Gamma(\beta+k)}{\Gamma(\beta)\lambda^k}\right), \qquad \alpha > k. \end{aligned}$$

In particular, for $k = 1$, we get the expected value

$$\begin{aligned} \mathbb{E}[X] = & (1-r)\left(\frac{\alpha}{\alpha-1}\right)\left(\frac{\beta}{\lambda}\right) \\ & + r\left(\frac{1}{\Phi(\alpha\sigma)}\right)(1-\Phi(\sigma(1-\alpha)))\exp\left(-\alpha\sigma^2 + \sigma^2/2\right)\left(\frac{\beta}{\lambda}\right), \qquad \alpha > 1. \end{aligned}$$

The variance is then derived from the second moment

$$\begin{aligned} \mathbb{E}[X^2] = & (1-r)\left(\frac{\alpha}{\alpha-2}\right)\left(\frac{\beta(\beta+1)}{\lambda^2}\right) \\ & + r\left(\frac{1}{\Phi(\alpha\sigma)}\right)(1-\Phi(\sigma(2-\alpha)))\exp\left(-2\alpha\sigma^2 + 2\sigma^2\right)\left(\frac{\beta(\beta+1)}{\lambda^2}\right), \qquad \alpha > 2. \end{aligned}$$

For the stop-loss transform, we get

$$\begin{aligned}\pi_X(d) &= \mathbb{E}[X] - d(1-F_X(d)) \\ &\quad - \left(\frac{\beta}{\lambda}\right) G(d;\beta+1,\lambda)\left(\frac{(1-r)\alpha}{\alpha-1} + \frac{r\Phi((\alpha-1)\sigma)\exp\left(-\alpha\sigma^2+\sigma^2/2\right)}{\Phi(\alpha\sigma)}\right) \\ &\quad - \frac{\alpha(1-r)\Gamma(\alpha+\beta)}{(\alpha-1)d^{\alpha-1}\Gamma(\beta)\lambda^{\alpha}} G(d;\alpha+\beta,\lambda),\end{aligned}$$

where $\alpha > 1$.

The shape of the stop-loss transform is illustated in Figure 3.2 for a moderate threshold heterogeneity ($\mathbb{V}[\Theta] = 50$) and a larger threshold heterogeneity ($\mathbb{V}[\Theta] = 500$). We clearly see that allowing for more dispersion in the threshold values increases the stop-loss transform.

3.3 Lognormal distributed threshold

In the same way as the previous subsection, we now use the lognormal distribution for the threshold in model (3.1). Let $\Theta \sim LN(\beta,\lambda)$, that is, the probability density function of Θ is given by

$$g(\theta;\beta,\lambda) = \frac{1}{\sqrt{2\pi}\theta\lambda}\exp\left(-\frac{1}{2}\left(\frac{\ln(\theta)-\beta}{\lambda}\right)^2\right), \qquad \theta > 0, \tag{3.4}$$

for $-\infty < \beta < \infty$ and $\lambda > 0$.

Using model (3.1) and equation (3.4), we get

$$\begin{aligned}f(x) &= \frac{(1-r)\alpha\exp\left(\frac{1}{2}\left(2\alpha\beta+\alpha^2\lambda^2\right)\right)\Phi\left(\frac{\ln(x)-(\beta+\alpha\lambda^2)}{\lambda}\right)}{x^{\alpha+1}} \\ &\quad + \frac{r\exp\left(-\frac{1}{2}\left(\frac{(\ln(x)-\beta)^2+\alpha\sigma^2(2\ln(x)+\alpha\sigma^2-2\beta)}{\lambda^2+\sigma^2}\right)\right)\left(1-\Phi\left(\frac{\sigma(\ln(x)-\alpha\lambda^2-\beta)}{\lambda\sqrt{\sigma^2+\lambda^2}}\right)\right)}{\sqrt{2\pi}\sigma\Phi(\alpha\sigma)\lambda x\sqrt{\frac{1}{\sigma^2}+\frac{1}{\lambda^2}}}.\end{aligned}$$

For $0 < k < \alpha$, the k^{th} raw moment of X is given by

$$\begin{aligned}\mathbb{E}[X^k] &= (1-r)\left(\frac{\alpha}{\alpha-k}\right)\exp\left(k\beta+\lambda^2k^2/2\right) \\ &\quad + r\left(\frac{1}{\Phi(\alpha\sigma)}\right)(1-\Phi(\sigma(k-\alpha)))\exp\left((\beta-\alpha\sigma^2)k+k^2(\sigma^2+\lambda^2)/2\right), \qquad \alpha > k.\end{aligned}$$

For $k = 1$, we get the expected value

$$\begin{aligned}\mathbb{E}[X] &= (1-r)\left(\frac{\alpha}{\alpha-1}\right)\exp\left(\beta+\lambda^2/2\right) \\ &\quad + r\left(\frac{1}{\Phi(\alpha\sigma)}\right)(1-\Phi(\sigma(1-\alpha)))\exp\left(\beta-\alpha\sigma^2+(\sigma^2+\lambda^2)/2\right), \qquad \alpha > 1.\end{aligned}$$

The variance is then derived from the second moment

$$\mathbb{E}[X^2] = (1-r)\left(\frac{\alpha}{\alpha-2}\right)\exp\left(2^(\beta+\lambda^2)\right) + r\left(\frac{1}{\Phi(\alpha\sigma)}\right)(1-\Phi(\sigma(2-\alpha)))\exp\left(2(\beta-\alpha\sigma^2)+2(\sigma^2+\lambda^2)\right), \qquad \alpha > 2.$$

In this case, the stop-loss transform formula is not particularly interesting.

4 Numerical illustration

4.1 Data set

In this example, we apply the mixed composite Lognormal–Pareto model (3.1) to a classical insurance data set. This allows us to compare our fit with the results obtained from previous models and several two-parameter distributions. Parameters are estimated by maximum likelihood (ML). In order to perform the comparison, goodness-of-fit is measured by means of the following criteria: (i) the value of the negative log-likelihood (NLL) at the values of the ML estimators (smaller values are good), and (ii) the Akaike information criterion (AIC equal to twice the NLL plus twice the number of parameters) evaluated at the ML estimators (higher values are good).

The data set comprises 2,492 Danish fire insurance losses and can be found in the R **SMPraticals** add–on package available from CRAN web page http://cran.r-project.org/. Losses are in millions of Danish Krone (DKK) from the years 1980 to 1990 inclusive and have been adjusted to reflect 1985 values. Among others McNeil (1997) and Resnick (1997) have analyzed upper portion of these data.

4.2 Maximum likelihood estimation of the parameters

Maximum likelihood estimators for the formulas of the Lognormal distribution, of the Pareto distribution, of the Gamma distribution, of the Weibull distribution, of the model (2.3), of the model (2.4), of the model (3.1) with Gamma distributed threshold, of the model (3.1) with Lognormal distributed threshold and of the model (2.8) are presented in Table 4.1. The maximum likelihood estimations were performed in R using **actuar** add–on package (see Dutang, Goulet & Pigeon (2008) for more information). The population mean is estimated to 3.598 with model (3.1) (Gamma) and to 3.637 with model (3.1) (Lognormal), to be compared to the sample mean $\bar{x} = 3.063$. Since $\hat{\alpha} < 2$, theoretical variance is infinite. The results for model (3.1) with Gamma distributed threshold and Lognormal distributed threshold are graphically similar. Moreover, estimated values for tail index α are quite similar (1.3580 and 1.3508) and we retain throughout the remainder of this example the first model (Gamma). The probability density function (3.1) satifies multiparameter Cramér–Rao conditions for asymptotic normality (Lehmann & Casella (1999)), so we provide in Table 4.2 confidence intervals at the 90% level for estimated parameters.

It might be interesting to compare the estimated value of tail index ($\hat{\alpha} = 1.3580$) obtained using model (3.1) with those calculated in McNeil (1997) using extreme value theory.

Distributions	Parameters			
Lognormal	$\hat{\mu} = 0.6718$	$\hat{\sigma} = 0.7323$	–	–
Pareto	$\hat{\theta} = 0.3134$	$\hat{\alpha} = 0.5460$	–	–
Gamma	$\hat{\lambda} = 0.4107$	$\hat{\alpha} = 1.2578$	–	–
Weibull	$\hat{\theta} = 2.9531$	$\hat{\tau} = 0.9476$	–	–
Model (2.3)	$\hat{\theta} = 1.3851$	$\hat{\alpha} = 1.4363$	–	–
Model (2.4)	$\hat{\theta} = 1.2075$	$\hat{\sigma} = 0.1965$	$\hat{\alpha} = 1.3282$	–
Model (3.1) (Gamma)	$\hat{\sigma} = 0.0005$	$\hat{\alpha} = 1.3580$	$\hat{\beta} = 42.8038$	$\hat{\lambda} = 45.0955$
Model (3.1) (Lognormal)	$\hat{\sigma} = 0.1653$	$\hat{\alpha} = 1.3508$	$\hat{\beta} = 0.1554$	$\hat{\lambda} = 0.0995$
Model (2.8)	$\hat{\theta} = 1.1447$	$\hat{\sigma} = 0.1823$	$\hat{\alpha} = 1.5631$	$\hat{\lambda} = 0.3633$

Table 4.1: Estimated values of fitted models for fire Danish loss data.

Parameters	Lower bounds	Estimated values	Upper bounds
α	1.305	1.358	1.412
σ	0.000	0.0005	0.127
λ	33.828	45.095	56.363
β	35.045	42.804	50.562

Table 4.2: Confidence intervals at level 90% for Danish fire insurance loss data.

Table 4.3 presents some estimated values of the tail index of generalized Pareto distribution for different thresholds u. Estimations are similar to those suggested in McNeil (1997).

4.3 Goodness-of-fit

We present in Figure 4.1 empirical histogram and fitted composite Lognormal–Pareto models. We provide in Table 4.4 the values of the NLL and AIC evaluated at the maximum likelihood estimators. The values of NLL and AIC show that the mixed composite Lognormal–Pareto model provides a better fit than classical distributions and models (2.3) and (2.4). Moreover, it presents a similar adjustment to model (2.8) and it is more intuitive.

A measure that provides useful information for insurers are the high quantiles of the distribution of the claim amounts. Usually quantiles can be estimated by their empirical counterparts but when we are interested in the very high quantiles, this approach is no longer valid since estimation based on a low number of large observations would be strongly

u	ξ	$\alpha = 1/\xi$
0	0.60	1.67
3	0.67	1.49
4	0.72	1.39
5	0.63	1.59
10	0.50	2.00
20	0.68	1.47

Table 4.3: Estimated values of tail index for different thresholds.

Distributions	NLL	AIC
Lognormal	4,434	8,872
Pareto	5,675	11,354
Gamma	5,243	10,490
Weibull	5,270	10,544
Model (2.3)	3,878	7,760
Model (2.4)	3,866	7,739
Model (3.1)	3,860	7,728
Model (2.8)	3,860	7,728

Table 4.4: Values of statistical criteria evaluated at the MLEs.

		Fitted Lognormal–Pareto models			
Quantiles	Empirical	Model (2.3)	Model (2.4)	Model (3.1)	Model (2.8)
0.90	5.086	4.866	5.282	5.191	5.164
0.95	8.459	7.884	8.901	8.648	8,249
0.99	24.870	24.177	29.901	28.288	23.750
0.999	146.010	120.121	169.123	154.158	104.808
0.9999	263.250	596.921	960.384	840.096	458.917

Table 4.5: Empirical and fitted models quantiles.

inaccurate. The QQ–plots against models (2.3), (2.4), (3.1) and (2.8) are presented in Figures 4.2 and 4.3. As usual, estimated quantiles are plotted on y–axis and ordered observations on x–axis, where $\hat{F}^{-1}(p)$ is the estimated p^{th} quantile and $p = k/(n+1)$ with $k = 1, \ldots, n$. According to these graphs, we can see that model (3.1) is a reasonable choice for the given data. Moreover, empirical and fitted models quantiles in the extreme portion of the tail are presented in Table 4.5. It should be noted that the largest observations in the Danish set is 263.250 and empirical quantiles were obtained by

$$p_k = \frac{k - 1/3}{n + 1/3}$$

as suggested in Hyndman & Fan (1996). One can observe than model (3.1) less dramatically overstates extreme quantiles than model (2.4). However, we must remain cautious in the conclusions drawn from Table 4.5 because sample size is only 2,492 and, for example, the 99.99% empirical quantiles (maximum of the data set) represents an event that occurs 1 in 10,000 times.

We can compare models (2.4) and (3.1) by determining whether the variance of the Gamma distribution is significantly different from 0. Defining the alternative parameters $\kappa = \beta/\lambda^2$ and $\tau = \alpha$ for the threshold distribution, the 90% confidence interval for κ is $(0.013, 0.029)$, so variance is significantly different from 0. This suggests that thresholds indeed change from one contract to another.

Finally, we can also examine the stop-loss transform. We present in Figure 4.4 stop-loss transform curve using estimated parameters and empirical stop-loss transform curve. As one can see, the model tends to overestimate the stop-loss transform which may be problematic

for reinsurance applications. One can see that using a Lognormal distributed threshold does not improve the model. Also, this problem was already present in models presented by Scollnik (2007) and Cooray & Ananda (2005). The authors are developing a solution to this problem.

4.4 Probable maximal loss

Finally, we can evaluate the probable maximum loss (PML) using the mixed composite Lognormal–Pareto model. Broadly speaking, PML is the worst loss likely to happen. Let N be a random variable with Poisson distribution with mean κ and let $X_1, \ldots, X_N$ be a random sample with common cumulative distribution function $F(x)$ corresponding to (3.1). We define $M_N = \max(X_1, \ldots, X_N)$ and we estimate κ by average annual frequency, $\hat{\kappa} = 2,492/11 = 226.5455$.

As in Cebrian, Denuit & Lambert (2003), we set the PML equal to the solution of equation $\Pr[M_N \leq \text{PML}] = q$, for some high q. This means that the PML is a high quantile of the maximum of a random sample of size N. Since the maximum M_N will exceed the so-defined PML only in $100(1-q)\%$ of the cases, it is very unlikely that an individual claim amount assumes a value larger than the PML. Now,

$$\Pr[M_N \leq y] = \mathbb{E}[(F(y))^N] = \exp\left(-\kappa(1-F(y))\right).$$

Using respectively $q = 0.05$ and $q = 0.01$, we get PML $= 460.2464$ and PML $= 1,528.432$. Recall that the sample maximum is 263.25.

5 Conclusion

In this paper, we proposed an extension of the composite Lognormal–Pareto model introduced by Scollnik (2007). This new model is obtained by allowing the threshold separating the Lognormal and Pareto mixture components to become random. Several theoretical features of the new model are discussed. The classical Danish fire insurance losses data set is then successfully fitted with the help of the mixed composite Lognormal–Pareto model. We also performed this analysis on a second data set which consists of $1,797$ Sweden third party insurance loss data for year 1977 (available from http://lib.stat.cmu.edu) and previously employed by Hallin & Ingenbleek (1983). In this case also, we obtained satisfactory results.

Since the threshold becomes random in the model proposed in the present paper, this allows the actuary to update its distribution using credibility mechanisms. A posteriori distributions can be used to track the changes in the threshold behavior over calendar time for each policy. This may be particularly interesting in industrial insurance, where the threshold separating standard losses from large ones can be influenced by many individual risk characteristics.

To end with, let us mention an alternative has been developed by Buch-Kromann (2006) based on Buch-Larsen, Nielsen, Guillen & Bolance (2005). This approach is based on a Champernowne distribution, corrected with a non-parametric estimator (that is

obtained by transforming the data set with the estimated modified Champernowne distribution function and then estimating the density of the transformed data set using the classical kernel density estimator). Based on the analysis of a Danish data set, Buch-Kromann (2006) concluded that the Generalized Pareto approach performs better than the Champernowne one in terms of goodness-of-fit, whereas both methods are comparable in terms of predicting future claims.

It might be interesting to develop a semiparametric model such as that presented in Buch-Larsen, Nielsen, Guillen & Bolance (2005). This method improves the quality of estimation, particularly for heavy-tailed data sets. Moreover, this approach seems to lead to interesting results in contexts dealing with severity in insurance.

Acknowledgements

Michel Denuit acknowledges the financial support of the *Communauté française de Belgique* under contract "Projet d'Actions de Recherche Concertées" ARC 04/09-320, of the *Banque Nationale de Belgique* under grant "Risk measures and Economic capital", and of the *Onderzoeksfonds K.U. Leuven* (GOA/07: Risk Modeling and Valuation of Insurance and Financial Cash Flows, with Applications to Pricing, Provisioning and Solvency).

References

Buch-Kromann, T. (2006). Estimation of large insurance losses: A case study. *Journal of Actuarial Practice* **13**, 191-211.

Buch-Larsen, T., Nielsen, J.P., Guillén, M., & Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics* **39**, 503-518.

Cebrian, A., Denuit, M., & Lambert, Ph. (2003). Generalized Pareto fit to the society of Actuaries' large claims database. *North American Actuarial Journal* **7**, 18-36.

Cooray, K., & Ananda, M.M.A. (2005). Modeling actuarial data with a composite Lognormal-Pareto model. *Scandinavian Actuarial Journal*, 321–334.

Dutang, C., Goulet, V., & Pigeon, M. (2008). actuar: An R package for actuarial science. *Journal of Statistical Software* **25**.

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

Hallin, M., & Ingenbleek, J.-F. (1983). The Swedish automobile portfolio in 1977. A statistical study. *Scandinavian Actuarial Journal*, 49–64.

Hyndman, R.J., & Fan, Y. (1996). Sample quantiles in statistical packages *American Statistician* **50**, 361–365.

Klugman, S., Panjer, H.H., & Willmot, G.E. (2004). Loss Models: From Data to Decisions. *Wiley, New York.*

Klugman, S., & Rioux, J. (2004). Toward a unified approach to fitting loss models. *North American Actuarial Journal* **10**, 63–83.

Lehmann, E.L., & Casella, G. (1999). Theory of Point Estimation, Second Edition. *Springer, New York*.

McNeil, A.J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin* **27**, 117–137.

Resnick, S.I. (1997). Discussion of the Danish data on large fire insurance losses. *ASTIN Bulletin* **27**, 139–151.

Scollnik, D.P.M. (2007). On composite Lognormal-Pareto models. *Scandinavian Actuarial Journal*, 20–33.

Tancredi, A., Anderson, C., & O'Hagan, A. (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes* **9**, 87-106.

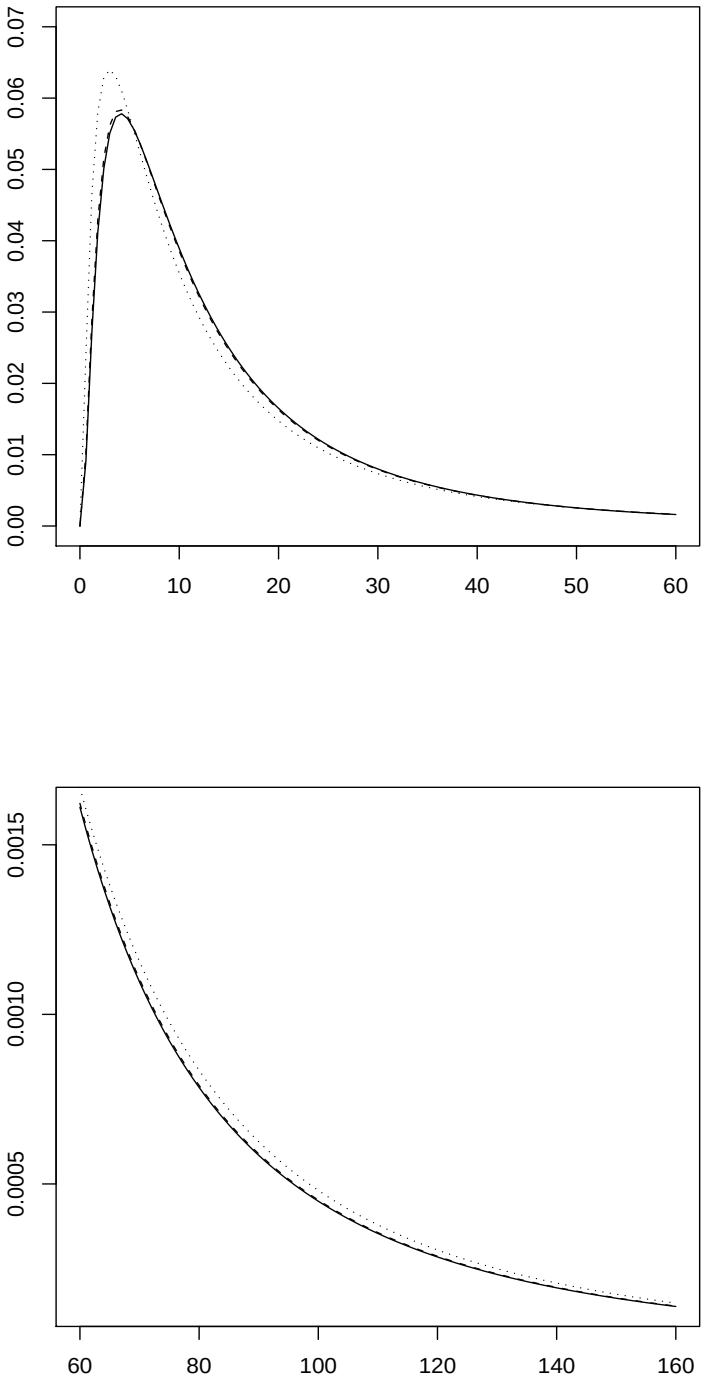


Figure 3.1: Probability density function (2.4) corresponding to the composite Lognormal–Pareto (solid line) together with the probability density function (3.1) corresponding to the mixed composite Lognormal–Pareto with $\mathbb{V}[\Theta] = 50$ (dashed line) and $\mathbb{V}[\Theta] = 500$ (dotted line).

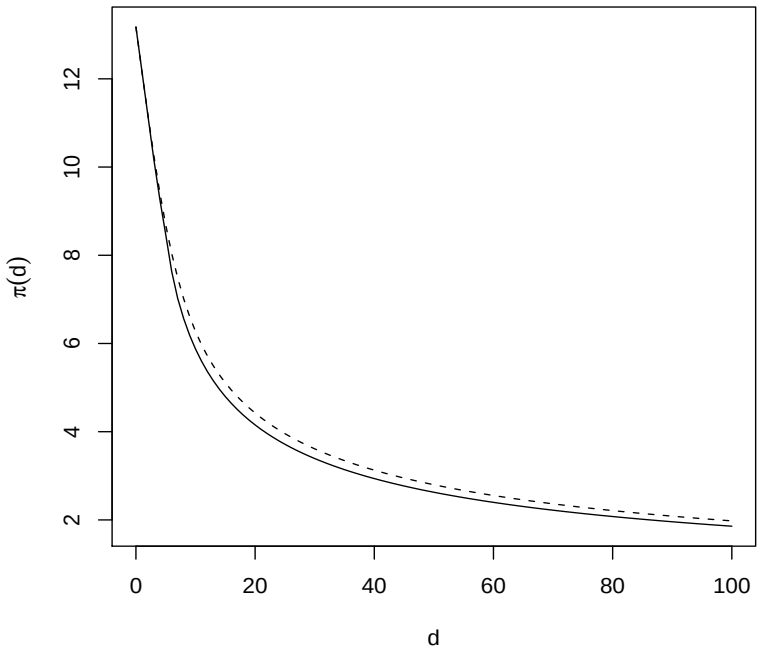


Figure 3.2: Stop-loss transform for $\alpha = 1.5$, $\sigma = 0.1$, $\beta = 50$ and $\lambda = 10$ in solid line and $\alpha = 1.5$, $\sigma = 0.1$, $\beta = 5$ and $\lambda = 1$ in dashed line.

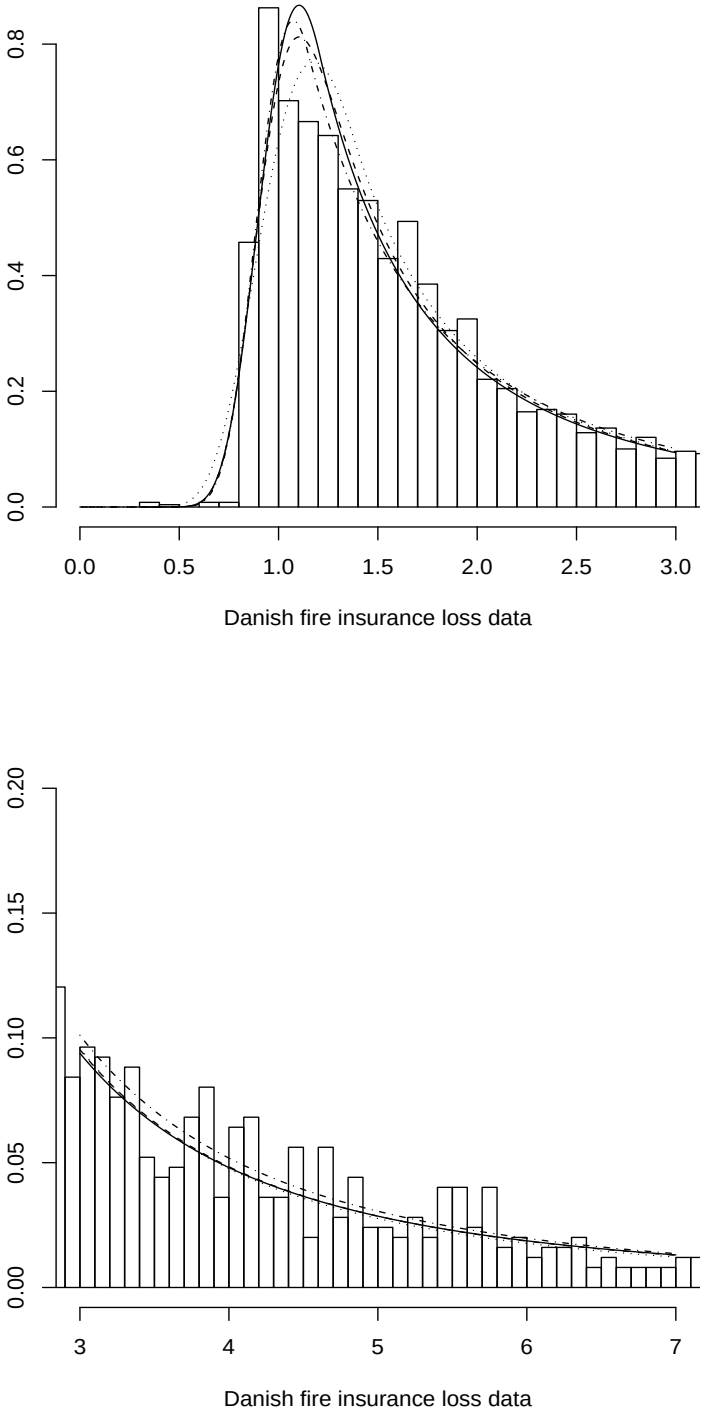
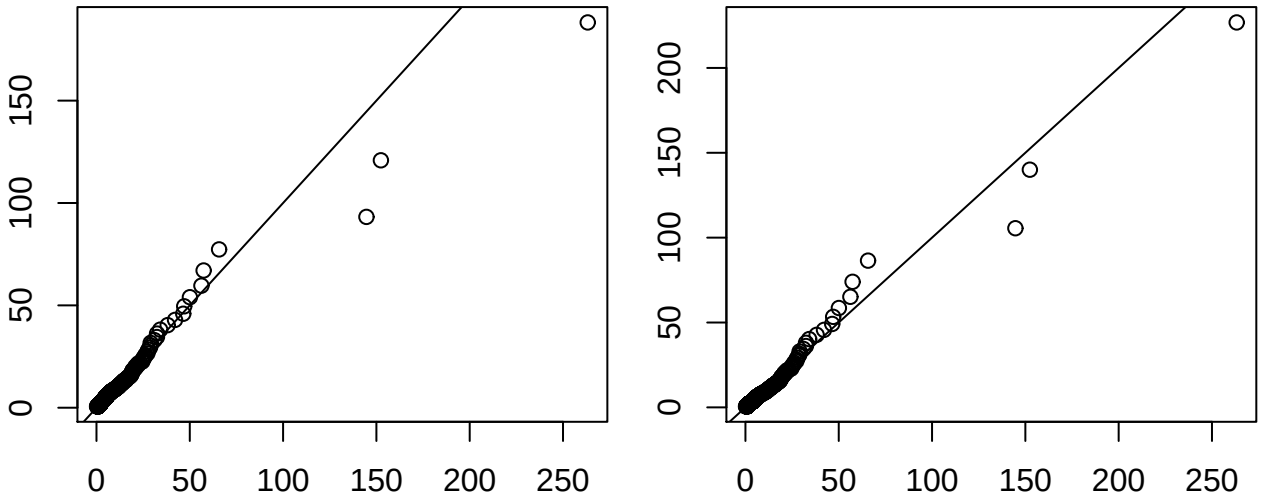


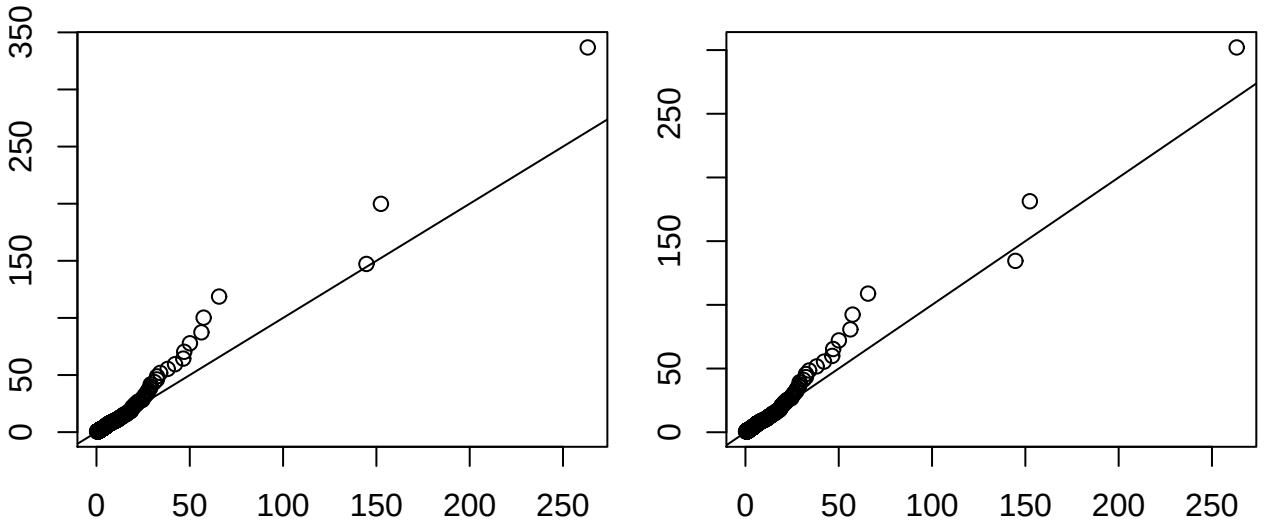
Figure 4.1: Comparison of empirical histogram of Danish fire insurance loss data, fitted model (2.3) in dotted line, fitted model (2.4) in solid line, fitted model (3.1) in dashed line and fitted model (2.8) in dotted–dashed line.



(a) Model (2.3)

(b) Model (2.4)

Figure 4.2: Q–Q plot for Danish fire insurance loss data.



(a) Model (3.1) (b) model (2.8)

Figure 4.3: Q–Q plot for Danish fire insurance loss data.

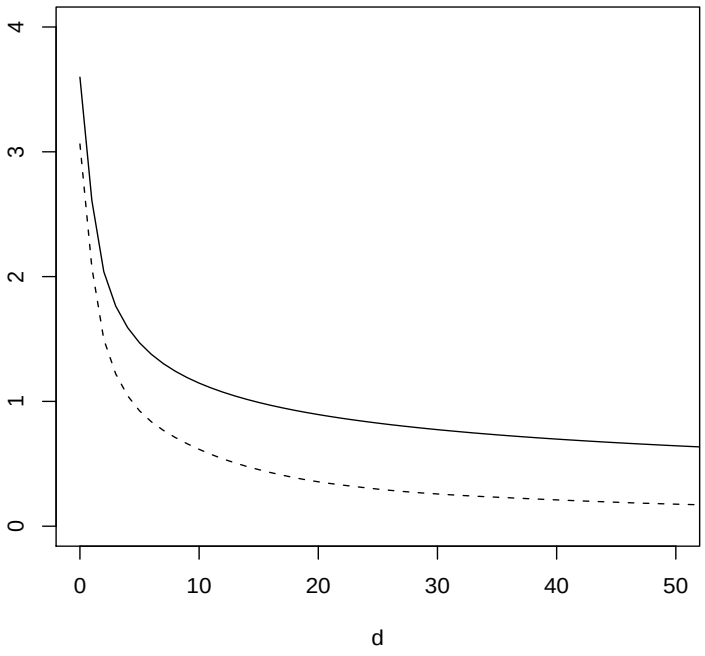


Figure 4.4: Stop-loss transform curve using model (3.1) and estimated parameters in solid line and empirical stop-loss transform curve in dashed line.

Recent Titles

0933. SIMAR, L. and P.W. WILSON, Inference by subsampling in nonparametric frontier models: Appendix

0934. PENG, L., QI, Y. and I. VAN KEILEGOM, Jackknife empirical likelihood method for copulas

0935. LAMBERT, Ph., Smooth semi- and nonparametric bayesian estimation of bivariate densities from bivariate histogram data

0936. HAFNER, C.M. and H. MANNER, Dynamic stochastic copula models: Estimation, inference and applications

0937. DETTE, H. and C. HEUCHENNE, Scale checks in censored regression

0938. MANNER, H. and J. SEGERS, Tails of correlation mixtures of elliptical copulas

0939. TIMMERMANS, C., VON SACHS, R. and V. DELOUILLE, Comparaison et classifications de séries temporelles via leur développement en ondelettes de Haar asymétriques. Actes des XVIe rencontres de la société francophone de classification, 2009.

0940. ÇETINYÜREK-YAVUZ, A. and Ph. LAMBERT, Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines

0941. ROUSSEAU, R., GOVAERTS, B. and M. VERLEYSEN, Combination of Independent Component Analysis and statistical modelling for the identification of metabonomic biomarkers in 1H-NMR spectroscopy

1001. BASRAK, B., KRIZMANIĆ, D. and J. SEGERS, A functional limit theorem for partial sums of dependent random variables with infinite variance

1002. MEINGUET, T. and J. SEGERS, Regularly varying time series in Banach spaces

1003. HUNT, J., A short note on continuous-time Markov and semi-Markov processes

1004. COLLÉE, A., LEGRAND, C., GOVAERTS, B., VAN DER VEKEN, P., DE BOODT, F. and E. DEGRAVE, Occupational exposure to noise and the prevalence of hearing loss in a Belgian military population: a cross-sectional study. Military Medicine, under revision.

1005. MEYER, N., LEGRAND, C. and G. GIACCONE, Samples sizes in oncology trials: a survey.To be submitted in the coming weeks (British Medical Journal).

1006. HAFNER, C.M. and O. REZNIKOVA, On the estimation of dynamic conditional correlation models

1007. HEUCHENNE, C. and I. VAN KEILEGOM, Estimation of a general parametric location in censored regression

1008. DAVYDOV, Y. and S. LIU, Transformations of multivariate regularly varying tail distributions

1009. CHRISTIANSEN, M. and M. DENUIT, First-order mortality rates and safe-side actuarial calculations in life insurance

1010. EECKHOUDT, L. and M. DENUIT, Stronger measures of higher-order risk attitudes

1011. DENUIT, M., HABERMAN, S. and A. RENSHAW, Comonotonic approximations to quantiles of life annuity conditional expected present values: extensions to general ARIMA models and comparison with the bootstrap

1012. DENUIT, M. and M. MESFIOUI, Generalized increasing convex and directionally convex orders

1013. DENUIT, M. and L. EECKHOUDT, A general index of absolute risk attitude

1014. PIGEON, M. and M. DENUIT, Composite Lognormal-Pareto model with random threshold

See also http://www.stat.ucl.ac.be/ISpub.html