Robust and Efficient Fitting of Loss Models: Diagnostic Tools and Insights

Vytaras Brazauskas[1]
University of Wisconsin-Milwaukee

Abstract

We consider robust and efficient fitting of claim severity models whose parameters are estimated using the method of trimmed moments, which was recently introduced by Brazauskas, Jones, and Zitikis (2009). In this article, we take the 'next' step by going beyond the theory and simulations, and address some important issues that arise in practical application of the method. Specifically, we introduce two graphical diagnostic tools that can be used to choose the trimming proportions, and hence help one to decide on the appropriate trade-off between robustness and efficiency. What is equally important, such tools are useful in model selection, for assessing the overall goodness of model fit, and for identification of outliers. Some insights about the choice between a 'good' fit and an 'even better' fit and its impact on risk evaluations are provided. Data analysis and illustrations are performed using real data which represents the total damage done by 827 fires in Norway for the year 1988.

[1]Department of Mathematical Sciences, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, Wisconsin 53201, U.S.A. E-mail address: vytaras@uwm.edu

1 Introduction

Fitting of loss models is a necessary first step in many insurance applications such as premium calculations, risk evaluations and determination of required reserves. A variety of methodologies is used for model-fitting and include empirical nonparametric, parametric, robust parametric, and other techniques. Each approach has its own advantages and disadvantages. Empirical methods, for example, are relatively simple and are based on very weak assumptions, but, by using such methods, one cannot make reliable inference beyond the range of the observed data. Parametric techniques, on the other hand, allow one to extrapolate beyond the range of the actual data, yield more efficient estimators but all of it comes at the expense of strong assumptions which can be of questionable validity in practice. The robust parametric methodology maintains all the advantages of the parametric approach but, at the same time, alleviates its main weakness (i.e., strong assumptions) by fitting models via methods that are relatively insensitive to the underlying assumptions. The latter approach seems indeed as a sensible compromise between the other two, and thus it is no surprise that robust models have seen a fair share of success in various fields of application. For a sample of recent examples, see: Dell'Aquila and Embrechts (2006) and Dupuis and Victoria-Feser (2006), for extreme-value applications; Cowell and Victoria-Feser (2006, 2007), for modeling of income inequality in economics; Marceau and Rioux (2001), for risk theory applications.

For a statistical procedure to be successful and widely accepted in practice, it should (*i*) be transparent with its actions on the data, (*ii*) have analytically tractable properties, small-sample or asymptotic, and (*iii*) be computationally straightforward. These three conditions are ambitious, yet not unreasonable, and they present a big challenge for any statistical procedure. This is also the case for robust procedures. An attempt to resolve the aforementioned issues was made by Brazauskas, Jones, and Zitikis (2009) who introduced a general method, called the method of trimmed moments (MTM), for robust fitting of distributions. This method essentially works like the standard method-of-moments and thus is easy-to-understand and is fairly simple analytically and computationally; however, it has its 'gray zones' too. More specifically, typical robust methods maintain their resistance against outliers and other non-representative data by trading off some efficiency at the assumed model. Such trade-offs

are controlled through the so-called tuning constants. The MTM procedures are no exception, they also have such built-in constants which are called trimming proportions. Thus, in this setting, a new question arises: How to choose tuning constants (trimming proportions) for a robust procedure? Usually the answer is given along these lines: "It depends on how much efficiency one is willing to sacrifice at the assumed model." While such answer is perhaps sufficient for further theoretical investigations, in most practical situations it is not entirely satisfactory as more clarity is desirable.

In search of better solutions for the problem just described, the importance of well-designed graphical tools can hardly be overstated. This certainly is not a new idea. Various diagrams and data displays have always existed in one form or another in science and engineering. But with the appearance of Tukey's pioneering book "Exploratory Data Analysis" (Tukey, 1977), graphics became far more concrete and effective. Also, advances in computing power over the last 30 years have changed how we carry out visualization. Finally, there are scores of books, some of them quite good, written on graphics. Cleveland (1993), for example, is an excellent read on data visualization, its underlying principles and relationships to classical statistical methods.

The power of graphical displays seems to stem primarily from the human eye's ability to notice patterns; for example, to detect deviations from linearity. Keeping this in mind, in this article we introduce two graphical diagnostic tools: a quantile-quantile-percentile plot and a percentile-residual plot. As will be demonstrated later, such plots can be used to choose the trimming proportions of an MTM estimator, and hence help one to decide on the appropriate trade-off between robustness and efficiency. What is equally important, they also are useful in model selection, for assessing the overall goodness of model fit, and for identification of outliers.

The rest of the article is organized as follows. Section 2 presents the definition of and some key facts about the MTM estimators. It also contains specialized results for the MTM estimators of lognormal and log-t distribution parameters; these two families will be used to fit the Norwegian fire claims data. In Section 3, the new diagnostic tools are introduced and employed for fitting and analyzing the Norwegian fire claims data (for the year 1988). Section 4 provides some insights about robust and efficient fitting, the choice between a 'good' fit and an 'even better' fit and its impact on risk evaluations. A summarizing discussion is presented in Section 5.

2 MTM Estimation

Consider a sample of n independent and identically distributed random variables, $X_1, \ldots, X_n$, whose distribution function F depends on k unknown parameters, $\theta_1, \ldots, \theta_k$; the number k is fixed. Let us denote $X_{1:n} \leq \cdots \leq X_{n:n}$ the order statistics of $X_1, \ldots, X_n$. The MTM estimators of $\theta_1, \ldots, \theta_k$ are found by the following three step procedure:

1. Compute k sample trimmed moments

$$\widehat{\mu}_j = \frac{1}{n - m_n - m_n^*} \sum_{i=m_n+1}^{n-m_n^*} h_j(X_{i:n}), \qquad j = 1, \ldots, k, \tag{2.1}$$

   where h_j is a real-valued specially chosen function, and m_n and m_n^* are integers such that $0 \leq m_n < n - m_n^* \leq n$ and $m_n/n \to a$, $m_n^*/n \to b$ as $n \to \infty$, where the proportions a and b are chosen by the researcher.

2. Derive the corresponding population trimmed moments

$$\mu_j := \mu_j(\theta_1, \ldots, \theta_k) = \frac{1}{1-a-b} \int_a^{1-b} h_j(F^{-1}(u))\, \mathrm{d}u, \qquad j = 1, \ldots, k, \tag{2.2}$$

   where F^{-1} denotes the quantile function of F. Note that when $a = b = 0$, then $\mu_j = \mathbf{E}[h_j(X)]$.

3. Match the population and sample trimmed moments and solve the following system of equations with respect to $\theta_1, \ldots, \theta_k$:

$$\begin{cases} \mu_1(\theta_1, \ldots, \theta_k) &= \widehat{\mu}_1, \\ & \vdots \\ \mu_k(\theta_1, \ldots, \theta_k) &= \widehat{\mu}_k. \end{cases} \tag{2.3}$$

The above obtained solutions, which we denote by

$$\widehat{\theta}_1 = g_1(\widehat{\mu}_1, \ldots, \widehat{\mu}_k), \ldots\ldots, \widehat{\theta}_k = g_k(\widehat{\mu}_1, \ldots, \widehat{\mu}_k),$$

are, by definition, the MTM estimators of the parameters $\theta_1, \ldots, \theta_k$. Note that the functions g_j are such that $g_j(\mu_1, \ldots, \mu_k) = \theta_j$.

Next, the vector of MTM estimators $(\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ is asymptotically normal with the mean vector $(\theta_1, \ldots, \theta_k)$ and the covariance matrix $n^{-1}\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'$, where $\mathbf{D} = [d_{ij}]_{i,j=1}^k$ is the Jacobian of the transformations $g_1, \ldots, g_k$ evaluated at the vector $(\mu_1, \ldots, \mu_k)$, that is, $d_{ij} = \partial g_i / \partial \widehat{\mu}_j|_{(\mu_1, \ldots, \mu_k)}$. Also, the

matrix $\mathbf{\Sigma} := \big[\sigma_{ij}\big]_{i,j=1}^{k}$ whose entries are

$$\sigma_{ij} = \frac{1}{(1-a-b)(1-a-b)} \int_a^{1-b} \int_a^{1-b} \big(\min\{u,v\} - uv\big)\, \mathrm{d}h_j\big(F^{-1}(v)\big)\, \mathrm{d}h_i\big(F^{-1}(u)\big), \tag{2.4}$$

represents the asymptotic covariance matrix of the sample trimmed moments (2.1). Adopting the notation introduced by Serfling (1980), this result can be neatly summarized as follows:

$$\big(\widehat{\theta}_1, \ldots, \widehat{\theta}_k\big) \quad \text{is} \quad \mathcal{AN}\big((\theta_1, \ldots, \theta_k),\, n^{-1}\, \mathbf{D\Sigma D'}\big), \tag{2.5}$$

where $\mathcal{AN}$ stands for 'asymptotically normal'.

Remark 1: The procedure (2.1)–(2.3) is presented for general k and in theory it should always work. In practice, however, it can happen that the system of equations (2.3) does not have a solution (e.g., when k is large). The good news for actuaries is that majority of loss models fall within the $k \leq 3$ case (see Klugman *et al.*, 2004, Appendix A). Moreover, not infrequently the claim severity distributions belong to a general class of location-scale families or their variants, which implies that $k \leq 2$. In the latter instances, the procedure is indeed straightforward. □

In view of Remark 1, we will first fit the lognormal distribution to the Norwegian fire claims data and later a log-t distribution with known degrees of freedom. Since both of these distributions are log-location-scale families, their cdfs have the following form:

$$F(x) = F_0\left(\frac{\log(x) - \theta}{\sigma}\right), \quad x > 0,$$

defined for $-\infty < \theta < \infty$ and $\sigma > 0$, and where F_0 is the standard (i.e., with $\theta = 0$ and $\sigma = 1$) cdf of the underlying location-scale family. (In all our examples, F_0 will denote either the standard normal or standard t, with 8 degrees of freedom, cdf.) The corresponding quantile function satisfies the following relationship:

$$\log\big[F^{-1}(t)\big] = \theta + \sigma F_0^{-1}(t), \quad 0 < t < 1. \tag{2.6}$$

A straightforward application of the procedure (2.1)–(2.3), with $h_1(t) = \log t$ and $h_2(t) = (\log t)^2$, yields that the MTM estimators of θ and σ are:

$$\begin{cases} \widehat{\theta}_{\text{MTM}} &= \widehat{\mu}_1 - c_1\, \widehat{\sigma}_{\text{MTM}}, \\ \widehat{\sigma}_{\text{MTM}} &= \sqrt{(\widehat{\mu}_2 - \widehat{\mu}_1^2)/(c_2 - c_1^2)}, \end{cases} \tag{2.7}$$

where $\widehat{\mu}_j = (n - m_n - m_n^*)^{-1} \sum_{i=m_n+1}^{n-m_n^*} \big[\log X_{i:n}\big]^j$ are computed from the sample, and constants $c_j \equiv c_j(F_0,\, a,\, b) := \frac{1}{1-a-b} \int_a^{1-b} \big[F_0^{-1}(u)\big]^j \, \mathrm{d}u$ can be evaluated using, for example, the Trapezoidal Rule. That is, $\int_a^{1-b} \big[F_0^{-1}(u)\big]^j \, \mathrm{d}u \approx h\Big[0.5\big[F_0^{-1}(a)\big]^j + \sum_{k=1}^{m-1} \big[F_0^{-1}(a+kh)\big]^j + 0.5\big[F_0^{-1}(a+mh)\big]^j\Big]$, where $h = ((1-b)-a)/m$ and $m \geq 1000$.

Further, the estimators given by (2.7) are asymptotically normal, namely,

$$\big(\widehat{\theta}_{\text{MTM}},\, \widehat{\sigma}_{\text{MTM}}\big) \quad \text{is} \quad \mathcal{AN}\left(\big(\theta,\, \sigma\big),\, \frac{\sigma^2}{n}\, \mathbf{S}\right) \tag{2.8}$$

with the parameter-free matrix

$$\mathbf{S} = \frac{1}{(c_2 - c_1^2)^2} \left[\begin{array}{cc} c_1^* c_2^2 - 2c_1 c_2 c_2^* + c_1^2 c_3^* & -c_1^* c_1 c_2 + c_2 c_2^* + c_1^2 c_2^* - c_1 c_3^* \\ -c_1^* c_1 c_2 + c_2 c_2^* + c_1^2 c_2^* - c_1 c_3^* & c_1^* c_1^2 - 2c_1 c_2^* + c_3^* \end{array} \right] =: \left[\begin{array}{cc} s_{11} & s_{12} \\ s_{21} & s_{22} \end{array} \right],$$

where constants $c_j^* \equiv c_j^*(F_0,\, a,\, b)$ are parameter-free components of (2.4) and are given by

$$\begin{aligned} c_1^* \;=\; & \frac{1}{(1-a-b)^2}\Bigg\{ a(1-a)\big[F_0^{-1}(a)\big]^2 + b(1-b)\big[F_0^{-1}(1-b)\big]^2 - 2ab\, F_0^{-1}(a) F_0^{-1}(1-b) \\ & - 2(1-a-b)\Big[aF_0^{-1}(a) + bF_0^{-1}(1-b)\Big]\, c_1 - (1-a-b)^2\, c_1^2 + (1-a-b)\, c_2 \Bigg\}, \\ c_2^* \;=\; & \frac{1}{2(1-a-b)^2}\Bigg\{ a(1-a)\big[F_0^{-1}(a)\big]^3 + b(1-b)\big[F_0^{-1}(1-b)\big]^3 \\ & - ab\, F_0^{-1}(a) F_0^{-1}(1-b)\Big[F_0^{-1}(a) + F_0^{-1}(1-b)\Big] - (1-a-b)\Big[a\big[F_0^{-1}(a)\big]^2 + b\big[F_0^{-1}(1-b)\big]^2\Big]\, c_1 \\ & - (1-a-b)\Big[aF_0^{-1}(a) + bF_0^{-1}(1-b)\Big]\, c_2 - (1-a-b)^2\, c_1\, c_2 + (1-a-b)\, c_3 \Bigg\}, \\ c_3^* \;=\; & \frac{1}{4(1-a-b)^2}\Bigg\{ a(1-a)\big[F_0^{-1}(a)\big]^4 + b(1-b)\big[F_0^{-1}(1-b)\big]^4 - 2ab\big[F_0^{-1}(a)\big]^2\big[F_0^{-1}(1-b)\big]^2 \\ & - 2(1-a-b)\Big[a\big[F_0^{-1}(a)\big]^2 + b\big[F_0^{-1}(1-b)\big]^2\Big]\, c_2 - (1-a-b)^2\, c_2^2 + (1-a-b)\, c_4 \Bigg\}. \end{aligned}$$

For derivations and numerical illustrations, see Brazauskas, Jones, and Zitikis (2009).

Remark 2: As is well-known (cf., e.g., Serfling, 2002), the MLE of lognormal $(\theta,\, \sigma)$ is given by

$$\left\{ \begin{array}{rcl} \widehat{\theta}_{\text{MLE}} & = & n^{-1} \sum_{i=1}^n \log X_i, \\ \widehat{\sigma}_{\text{MLE}} & = & \sqrt{n^{-1} \sum_{i=1}^n \big(\log X_i - \widehat{\theta}_{\text{MLE}}\big)^2}, \end{array} \right.$$

and

$$\big(\widehat{\theta}_{\text{MLE}},\, \widehat{\sigma}_{\text{MLE}}\big) \quad \text{is} \quad \mathcal{AN}\left(\big(\theta, \sigma\big),\, \frac{\sigma^2}{n}\, \mathbf{S_0}\right) \quad \text{with} \quad \mathbf{S_0} = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1/2 \end{array} \right].$$

When F_0 is the standard normal cdf, the MTM estimators (2.7) become $(\widehat{\theta}_{\text{MLE}}, \widehat{\sigma}_{\text{MLE}})$ for $m_n = m_n^* = 0$; also, since $\mathbf{S} \to \mathbf{S_0}$ when $a = b \to 0$, the MLE's asymptotic distribution follows from statement (2.8). Further, the MLE of log-t (θ, σ), with a *known* degrees of freedom ν, has no closed-form solution and is found by numerically maximizing the log-likelihood. Its asymptotic distribution, however, is straightforward to establish. In particular,

$$\left(\widehat{\theta}_{\text{MLE}}, \widehat{\sigma}_{\text{MLE}}\right) \quad \text{is} \quad \mathcal{AN}\left((\theta, \sigma), \frac{\sigma^2}{n}\mathbf{S}_*\right) \quad \text{with} \quad \mathbf{S}_* = \begin{bmatrix} \frac{\nu+3}{\nu+1} & 0 \\ 0 & \frac{\nu+3}{2\nu} \end{bmatrix}.$$

□

3 Diagnostics and Model Fitting

In this section, we introduce two graphical diagnostic tools: a quantile-quantile plot supplemented with a vertical axis that represents empirical percentile levels, which we call a *quantile-quantile-percentile* (QQP) plot, and a plot of empirical percentile levels versus standardized residuals, which we call a *percentile-residual* (PR) plot. These tools should be employed in the model fitting/parameter calibration process after preliminary diagnostics (e.g., histogram and QQ-plot) has been completed and necessary distributional assumptions have been made. For all numerical and graphical illustrations, we use the Norwegian fire claims data which is taken from Beirlant, Teugels, and Vynckier (1996). The data set has been studied in the actuarial literature, and it represents the total damage done by $n = 827$ fires in Norway for the year 1988, which exceed 500 thousand Norwegian krones.

For this data set, the histogram of the raw observations is not very informative since about 90% of the losses are between 500 and 3,000 and the two largest claims (150,597 and 465,365) are much larger than the others. That is, one claim visually suppresses 750 claims into about 5% of the scale on a graph. Therefore, we first take the logarithmic transformation of the data and then perform preliminary diagnostics. As one can see from the left panel of Figure 1, the histogram of the transformed data looks *approximately* bell-shaped, which implies that the original, i.e., not transformed, losses can be assumed as (roughly) lognormally distributed. The lognormal QQ-plot in Figure 1 (right panel) supports our initial conclusion and exposes 'bad' segments of the lognormal distribution—significant part of the lower tail and several upper tail claims—where this model will not fit the data well.

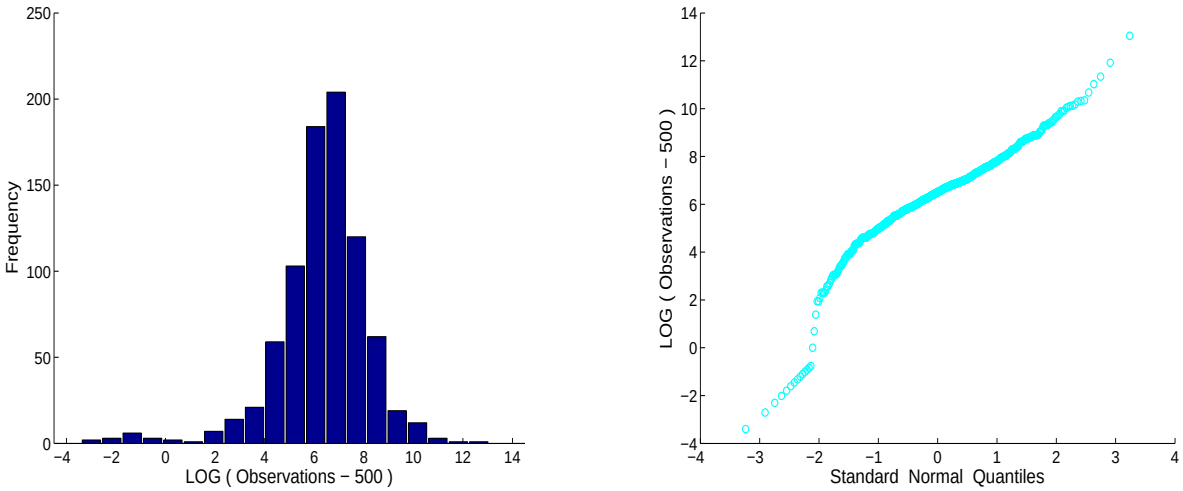

FIGURE 1: Preliminary diagnostics for the Norwegian fire claims (1988) data.

3.1 Quantile-Quantile-Percentile Plot

The data set we analyze here is fairly large (more than 800 observations) and because of that individual data points in the 'middle' of the QQ-plot cannot be seen. It is hence difficult to judge what fraction of the data follows the linear pattern. As a consequence, we propose to equip the QQ-plot with one more vertical axis that would show the percentile levels of empirical quantiles. Besides revealing empirical quantile's relative position within the sample, such plots will provide guidance about the minimal trimming requirements for MTM estimators. Notice also that the percentile levels are key inputs for some important risk measures (e.g., value-at-risk and conditional tail expectation).

Next, we fit the lognormal model using the MTM approach and several pairs of trimming proportions. For comparison, we also fit the model using the MLE approach which, as mentioned in Remark 2, corresponds to the MTM approach with no trimming. The resulting fits are illustrated in Figure 2, where the fitted models are labeled T1, T2, T3 and MLE. The goodness of model fit is measured by a *trimmed* mean absolute deviation between the fitted lognormal quantiles and the logarithm of the observed data (measured in thousands):

$$\Delta_\delta = \frac{1}{n-[n\delta]} \sum_{j=1}^{n-[n\delta]} \big| \log(\widehat{F}^{-1}(j/n)) - \log(X_{j:n} - 500) \big|,$$

where $\log(\widehat{F}^{-1}(t))$ is computed using equation (2.6) with F_0^{-1} denoting the quantile function of the

standard normal distribution. We use the following values of δ: 0, 0.05, 0.10, 0.25, 0.50. The choice $\delta = 0.25$, for example, indicates how close, on the average, are the 75% closest observations to a fitted line. Our empirical investigations suggest that the use of several goodness-of-fit measures provides a more objective picture about the model fit than an approach based on a single global criterion. The resulting parameter estimates and goodness-of-fit measurements appear in Table 1.

TABLE 1: Parameter estimates and goodness-of-fit measurements of the lognormal model.

Fitting Procedure	Parameter Estimates		Model Fit (Δ_δ)				
	θ	σ	$\delta = 0.50$	$\delta = 0.25$	$\delta = 0.10$	$\delta = 0.05$	$\delta = 0$
MLE	6.341	1.851	0.19	0.26	0.29	0.31	0.37
T1 ($a = 0.45$, $b = 0.45$)	6.452	8.334	2.29	3.60	4.55	4.91	5.38
T2 ($a = 0.10$, $b = 0.10$)	6.454	1.370	0.03	0.04	0.07	0.10	0.22
T3 ($a = 0.10$, $b = 0.01$)	6.461	1.449	0.04	0.06	0.07	0.09	0.21

The left panel of Figure 2 shows two lognormal fits: T1 and MLE. It is known (cf., Brazauskas, Jones, and Zitikis, 2009) that the T1 estimator is highly robust, i.e., it has high trimming proportions (see Table 1), but also very inefficient, its efficiency (relative to MLE) is only 18%. This translates into completely meaningless fit. The MLE procedure being most efficient, its efficiency is 100%, but non-robust yields a fair overall fit, especially when compared to that of T1. However, a carefull examination of the QQP-plot also reveals that data moves away from the linear pattern around the 10th and 99th percentile levels. Therefore, we can expect that trimming approximately 10% of lower observations and 1% of upper observations should improve the fit of MTM procedures. Indeed, as one can see from the right panel of Figure 2, the T2 and T3 fits are: reasonable overall, very accurate in the 'middle', and virtually identical. Moreover, it seems that the values of the trimming proportions do not have to be highly precise, as is demonstrated by the two procedures. For T2, we symmetrically trimmed 10% of lower and upper observations, and, for T3, $a = 0.10$ and $b = 0.01$ were chosen. We finally note that these two MTM procedures are quite efficient with the efficiencies of 77% (for T2) and 87% (for T3).

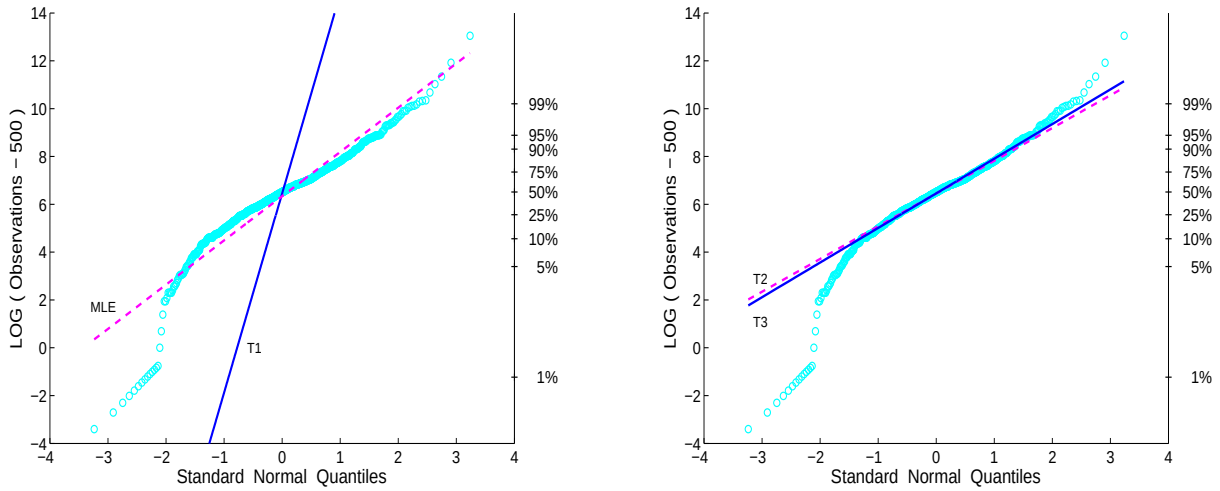

FIGURE 2: Lognormal QQP-plots and lognormal models fitted by MLE and three MTM methods with $a = b = 0.45$ (T1), $a = b = 0.10$ (T2; dashed line), $a = 0.10$, $b = 0.01$ (T3; solid line). (The right vertical axis represents empirical percentile levels.)

Although visual assessment may suggest that T2 and T3 fits yield minor improvements over the MLE fit, numerical evaluations of the fits reveal substantial differences (see Table 1). Indeed, for the overall fit ($\delta = 0$), MLE's mean absolute deviation is 68%–76% worse than that of T2, T3. And for the restricted ranges, the differences are even more dramatic: 210%–244% ($\delta = 0.05$), 314% ($\delta = 0.10$), 333%–550% ($\delta = 0.25$), 375%–533% ($\delta = 0.50$). Notice also that neither the numerical nor visual comparisons of the T2 and T3 fits allow to identify a clear winner between the two procedures. Thus, a 'practical' solution in this case would be to choose the method with better efficiency, i.e., T3.

3.2 Percentile-Residual Plot

The QQP-plot that we presented in the previous section works well for location-scale families or their variants, e.g., log-location-scale families. However, one has to be very careful when using such plots for other distributions, for which equation (2.6) does not hold, because they can be misleading. For instance, in order to assess the quality of a gamma distribution fit, the unknown shape parameter needs to be estimated first and then plugged in the formula of gamma quantiles. Subsequently, these quantile *estimates* are used to define the numerical scale on the horizontal axis of the QQP-plot, which means that the scale of the axis will crucially depend on the method of estimation (of parameters and quantiles). To see why that is dangerous, let us revisit the models of Figure 2, where we know which

fits are good and which are bad. The left panel of Figure 3, shows the lognormal QQP-plot where the

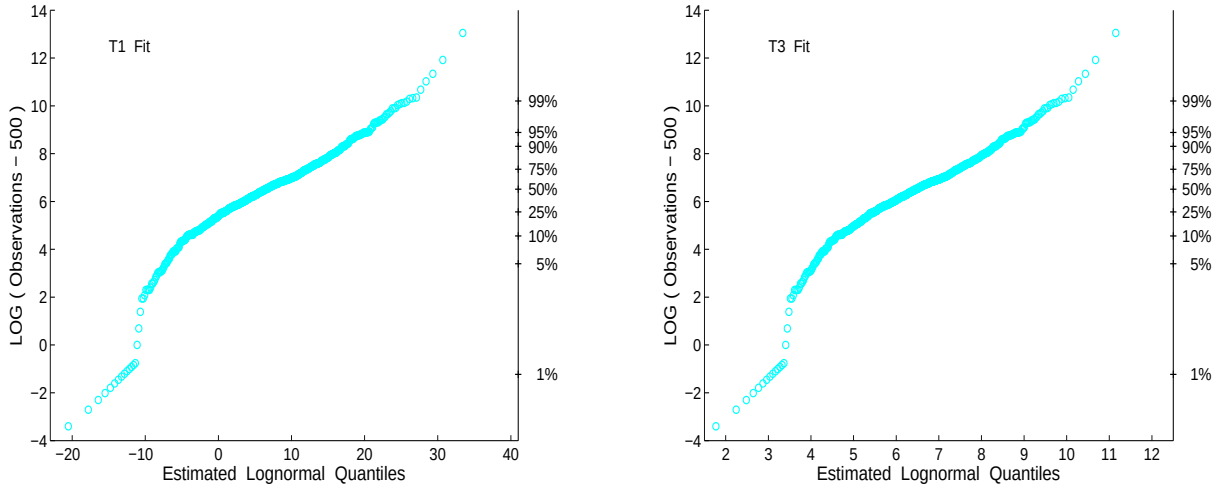

FIGURE 3: Lognormal QQP-plots with the model quantiles estimated by T1 and T3 methods. (The right vertical axis represents empirical percentile levels.)

lognormal quantiles on the horizontal axis are computed using the estimates from the T1 fit, the bad fit. The right panel represents the same plot based on the T3 fit, the good fit. Looking at the plots, the first impression we get is that both fits are of similar quality. Only after noticing significantly different scales on the horizontal axis do we realize that the fits are rather dissimilar. In conclusion, this example demonstrates that visualization tools should be carefully constructed, and plots based on the variables that are free of unknown parameters should be preferred.

In addition to the QQP-plot's ineffectiveness beyond the location-scale situations, there is another aspect of such plots (and of graphical diagnostics in general) that we have not addressed yet: the variability of the fitted line. Let us revisit Figure 2. There, one would like to know whether the central quantiles of the MLE line are significantly far away from the corresponding sample quantiles; they appear fairly close on the graph. Also, since the T1 method is very inefficient, its variance is much larger than that of MLE and thus confidence intervals (for fixed quantiles) constructed using this method will be much wider. Thus, the question of interest is whether those deviations of the estimated quantiles are still within the T1's confidence bounds.

To resolve the issues discussed above, we first need to find the asymptotic distribution of the point estimator of the quantile $\log F^{-1}(t)$, $0 < t < 1$. According to (2.6), the point estimator is given by

$$\widehat{\log F^{-1}}(t) = \widehat{\theta} + \widehat{\sigma} F_0^{-1}(t), \tag{3.1}$$

where $\widehat{\theta}$ and $\widehat{\sigma}$ denote either MLE or MTM estimators of θ and σ, respectively. As follows from (2.8) and Remark 2, both types of estimators—MLE and MTM—of θ and σ are asymptotically normal. Hence, direct application of the delta method (cf., e.g., Serfling, 1980, Section 3.3) to function (3.1) yields that the estimator $\widehat{\log F^{-1}}(t)$ is also asymptotically normal; specifically,

$$\widehat{\log F^{-1}}(t) \quad \text{is} \quad \mathcal{AN}\left(\log F^{-1}(t),\ \frac{\sigma^2}{n}\left[s_{11} + 2F_0^{-1}(t)s_{12} + \left[F_0^{-1}(t)\right]^2 s_{22}\right]\right), \tag{3.2}$$

where s_{11}, s_{12} and s_{22} are computed as in (2.8) and Remark 2.

Next, we define the *standardized residual* as

$$R_{j,n} \ = \ \frac{\log(X_{j:n} - 500) - \widehat{\log F^{-1}}(j/n)}{\left(\widehat{\sigma}/\sqrt{n}\right)\sqrt{s_{11} + 2F_0^{-1}(j/n)s_{12} + \left[F_0^{-1}(j/n)\right]^2 s_{22}}} \tag{3.3}$$

Due to the asymptotic normality result (3.2), at each fixed point j/n, $1 < j < n$, the standardized residual $R_{j,n}$ approximately behaves like the standard normal random variable. (Strictly speaking, the quality of this approximation depends on how extreme the ratio j/n is. For the cases where j/n is near 0 or 1, convergence rates to the standard normal random variable are slower than those for j/n around 1/2. In this paper, however, such technical distinctions do not concern us.) Plotting of the empirical percentile levels, $(j/n)100\%$, versus $R_{j,n}$ will yield a PR-plot. To decide on which residuals are statistically large and which are not, we will follow Hubert, Rousseeuw, and Van Aelst (2004) and use the cut-off points of ± 2.5. (The standard normal variable exceeds 2.5 with probability 0.0062.) These authors used similar standardized residuals plots to identify outliers in regression models.

In Figure 4, we present PR-plots for the lognormal model fitted by MLE and T1, T2, T3 methods. In each plot, the horizontal line at 0 corresponds to the respective fitted line in Figure 2, and the ± 2.5 lines are the tolerance limits. A good fit would be the one for which the majority of points (ideally, all points) are scattered between the tolerance limits. The conclusions that emerge from these PR-plots are consistent with what we have observed in the QQP-plots of Figure 2.

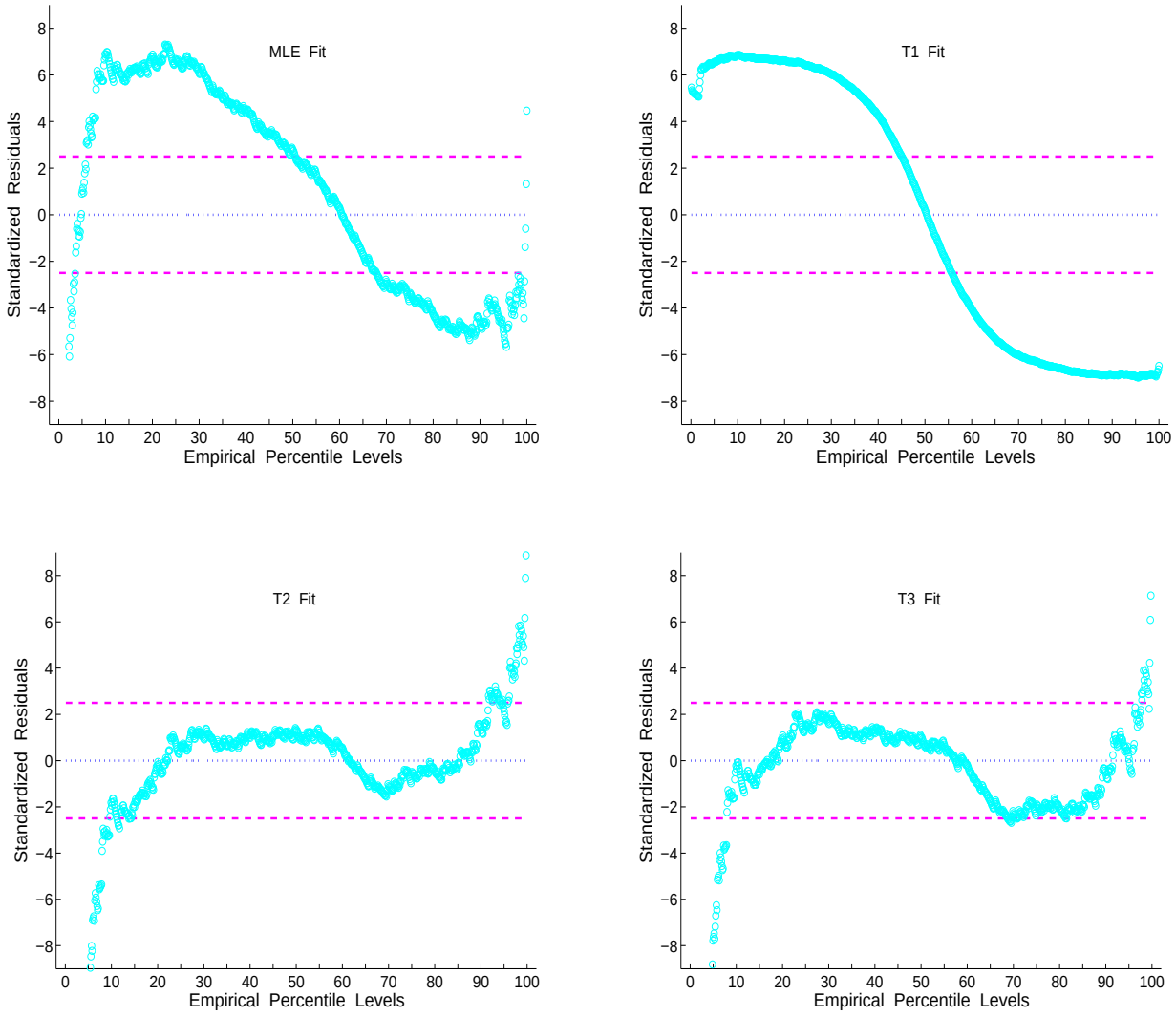

FIGURE 4: PR-plots for the lognormal model fitted by MLE and three MTM methods with $a = b = 0.45$ (T1), $a = b = 0.10$ (T2), $a = 0.10$, $b = 0.01$ (T3).

First of all, there are a few spots—around 5th and 99th percentiles, and between 50th and 65th percentiles—where the fit between the MLE-calibrated lognormal model and the data set is good. But overall MLE does not fit the main body of data well because it attempts to fit *all* data points. Secondly, as expected, the T1-based model is inappropriate for the data at hand but it provides an acceptable fit for the 'center', i.e., between approximately 45th and 55th percentiles. Thirdly, the T2- and T3-calibrated models accomodate about 80% (between 10th and 90th percentiles) and 90% (between 7th and 97th percentiles) of the data, respectively.

4 Insights

We start this section by summarizing our observations about the model-fitting exercise of Section 3, and by elaborating on few noteworthy features of robust and efficient fitting. Then, we identify an even better fitting model to this data and discuss whether that is worthwhile pursuing. After that, we investigate the implications of a model fit on risk evaluations.

4.1 Diagnostics and Robust-Efficient Fitting

As was seen in Section 3, the QQP-plot and PR-plot enable us to visually assess the disparity between the observed data and the fitted model. Moreover, for location-scale families, the QQP-plot can be used to diagnose whether an assumed distribution is acceptable for the data at hand without even fitting it. For non-location-scale distributions, this is not the case as the model has to be fitted first and then the PR-plot diagnostics employed. Let us discuss in more detail some observations about the plots and robust-efficient fitting.

First of all, it seems that well-designed graphical devices carry more information, thus lead to better informed decisions, than regular goodness-of-fit testing. To see this, let us use two popular normality tests—Lilliefors' and Jarque-Bera—and test whether the transformed Norwegian fire claims data, $\log(X_1 - 500), \ldots, \log(X_{827} - 500)$, is normally distributed. (Lilliefors' test is based on the usual one-sample Kolmogorov-Smirnov statistic but does not require that the null distribution be completely specified. The Jarque-Bera test is also suitable for situations where a fully-specified null distribution is not known. It is specifically designed for alternatives in the Pearson family of distributions. Both tests use relevant sample moments, i.e., method-of-moments (MM) estimators, to estimate the unknown normal distribution moments.) The tests strongly reject the normality assumption with both p-values being less than 0.001. The test statistic values are: 0.0962, for Lilliefors', and 1136.9, for Jarque-Bera. What we learn from these calculations is that, although the decision about the quality of MM fit is easy to make, we have no idea what caused it. Does it imply that the distributional assumption was completely inappropriate? Was it wrong to choose MM as a fitting method? Well, answers to these questions are not provided by the tests. On the other hand, the QQP-plot suggests that about 90% out of $n = 827$ data points are in tune with the normality assumption. Hence, the only thing one needs

is an estimation method that yields a close fit for those 90% of losses, declaring the rest as outliers.

The ability of these plots to expose outliers is another useful feature which can be utilized as a guide (but not a clear-cut answer) on minimum trimming requirements. What is important here, this guidance must always be supplemented with the knowledge of robustness-efficiency trade-offs the chosen MTM procedure can offer. Why? Let us review Figure 2 one more time. The QQP-plots indicate that trimming less than 10% of lower observations and less than 1% of upper observations may affect/rotate the line. This is exactly what happens to the MLE line, which is the most efficient but non-robust method. At the other extreme, severe over-trimming (method T1) may rotate the line even more. The latter fact suggests that highly trimmed but inefficient estimators may capture spurious patterns in the sample, and thus should be used with caution. Note also that usually more trimming implies less efficiency but there are exceptions to this rule (e.g., Weibull and Gumbell distributions).

Finally, it is quite clear by now that the choice of trimming proportions should be prescribed by the data; in other words, it should be adaptive. However, although it is reasonable to expect that the underlying probability distribution for the insurance loss will not change under similar circumstances, its realization (e.g., next year's sample) will. And the fraction of outliers in future samples is likely to be different (hopefully, not too much) from that in the current one. Therefore, a prudent actuary should always trim a bit more than the data indicates. Such an approach is safety oriented, and safety comes at a price. Fortunately, the price is not steep. For instance, for the Norwegian data, it seems that we could trim only 7% of lower observations and 1% of upper observations, thus accomodating even more data with the lognormal model. (Data accomodation, not removal, is the main objective of model-fitting!) However, the end result is pretty much the same: QQP- and PR-plots are visually indistinguishable from those of T3, and the efficiency of the new method is 90%; it is 87% for T3.

4.2 'Good' Fit versus 'Even Better' Fit

As we have seen so far, the lognormal model fitted by the T3 method seems to be optimal and not much else can be squeezed out of the lognormal assumption. But can we do better by changing the distributional assumption? Is it worthwhile to search for anything better? The answer to the first question is easy: yes, we can. In particular, notice that several upper claims are above the T3 line,

which means that the observed (log) claims have heavier upper tail than that of the standard normal. Since the histogram of the transformed data is roughly bell-shaped (see Figure 1), we should explore

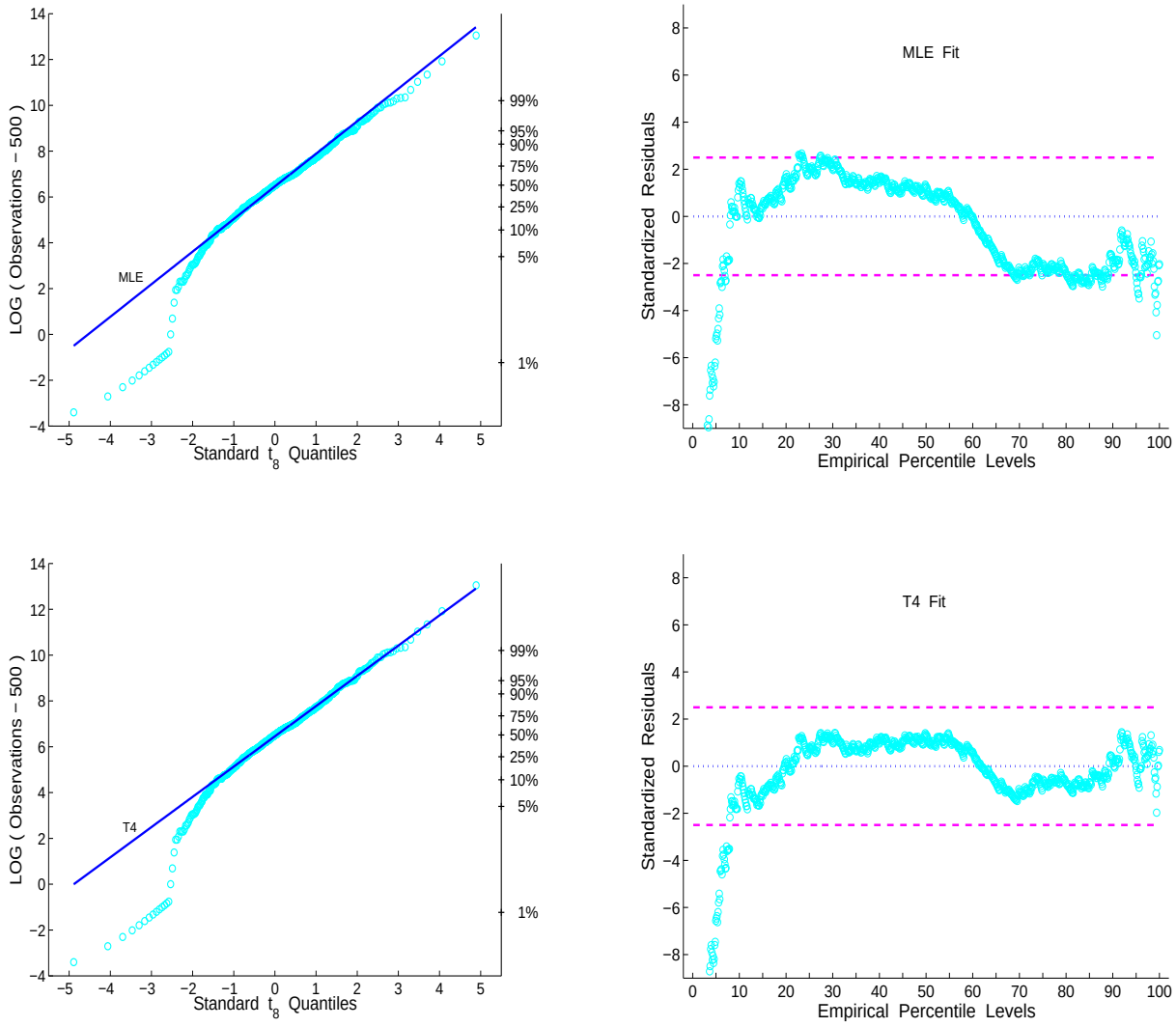

FIGURE 5: QQP- and PR-plots for the log-t_8 model fitted by MLE and the MTM method with $a = 0.10$, $b = 0.01$ (T4). Goodness-of-fit measurements Δ_δ ($\delta = 0, 0.05, 0.10, 0.25, 0.50$): $0.18, 0.10, 0.09, 0.07, 0.05$ (MLE) and $0.16, 0.06, 0.05, 0.04, 0.03$ (T4). Parameter estimates: $(\widehat{\theta}_{\text{MLE}}, \widehat{\sigma}_{\text{MLE}}) = (6.457,\ 1.422)$ and $(\widehat{\theta}_{\text{MTM}}, \widehat{\sigma}_{\text{MTM}}) = (6.452,\ 1.320)$.

the t distribution alternative. Following similar steps as in Section 3, we find that the log-t_8 distribution fitted using the MTM approach, with $a = 0.10$, $b = 0.01$, yields an even better fit and accomodates about 93% of the data. For comparison, we also include the MLE fit which is better than in the

lognormal case, yet it is still uniformly improved by the MTM fit. See Figure 5, for QQP- and PR-plots of this model, and for the goodness-of-fit measurements Δ_δ.

To answer the second question, we have to ask ourselves what the fitted model is going to be used for. For example, if we are interested in pricing of the top 10% of risks, then the log-t_8 fit is excellent. Or, if the middle 50% of the loss distribution is of interest, then both models—robustly fitted lognormal and log-t_8—are sufficiently good. But if the riskiness of the 10% lowest claims needs to be evaluated, then none of the considered models is adequate. In most insurance applications, however, small claims, those that have high probability but relatively low economic impact, are not a primary concern. Usually, a good fit in the upper tail, i.e., for the low-probability-but-high-consequence events, is needed. Thus, it seems safe to say that the log-t_8 model would be appropriate for most applications based on the Norwegian fire claims data.

4.3 Quantitative Risk Management

To see how the quality of the model fit affects insurance risk evaluations, we will construct confidence intervals for several value-at-risk (VaR) measures. Mathematically, this measure is the $(1-\beta)$-level quantile of the distribution function F, that is, $\mathrm{VaR}(\beta) = F^{-1}(1-\beta)$. For empirical estimation, we replace F with the empirical cdf $\widehat{F}_n$. For parametric (MLE) and robust parametric (MTM) estimation, $\widehat{F}$ is found by replacing F's parameters with their respective MLE and MTM estimates. In particular, as presented by Kaiser and Brazauskas (2006), the empirical point estimator and the $100(1-\alpha)\%$ *distribution-free* confidence interval of $\mathrm{VaR}(\beta) = F^{-1}(1-\beta)$ are given by

$$\widehat{\mathrm{VaR}}_{\mathrm{EMP}}(\beta) = X_{n:n-[n\beta]} \qquad \text{and} \qquad \big(X_{n:k_1}\,,\; X_{n:k_2}\big),$$

where $k_1 = \Big[n\Big((1-\beta) - z_{\alpha/2}\sqrt{\beta(1-\beta)/n}\Big)\Big]$ and $k_2 = \Big[n\Big((1-\beta) + z_{\alpha/2}\sqrt{\beta(1-\beta)/n}\Big)\Big]$. Here $[\cdot]$ denotes "greatest integer part" and $z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution.

Formulas for the robust point estimators of $\mathrm{VaR}(\beta)$ can be found by applying the exponential transformation to (3.1). The corresponding $100(1-\alpha)\%$ confidence intervals are then derived by applying the delta method to (3.2). These two steps lead to:

$$\widehat{\mathrm{VaR}}_{\mathrm{MTM}}(\beta) = 500 + \exp\Big\{\widehat{\theta}_{\mathrm{MTM}} + \widehat{\sigma}_{\mathrm{MTM}} F_0^{-1}(1-\beta)\Big\}$$

and

$$\widehat{\mathrm{VaR}}_{\mathrm{MTM}}(\beta) \;\times\; \left(1 \;\pm\; z_{\alpha/2}\big(\widehat{\sigma}_{\mathrm{MTM}}/\sqrt{n}\big)\sqrt{s_{11} + 2F_0^{-1}(1-\beta)s_{12} + \big[F_0^{-1}(1-\beta)\big]^2 s_{22}}\right),$$

where F_0 denotes either the standard normal or standard t_8 cdf, and $z_{\alpha/2}$ is again the $(1-\alpha/2)$th quantile of the standard normal distribution. The MLE point and interval estimators are constructed by following similar steps which lead to:

$$\widehat{\mathrm{VaR}}_{\mathrm{MLE}}(\beta) = 500 + \exp\left\{\widehat{\theta}_{\mathrm{MLE}} + \widehat{\sigma}_{\mathrm{MLE}}F_0^{-1}(1-\beta)\right\}$$

and

$$\widehat{\mathrm{VaR}}_{\mathrm{MLE}}(\beta) \;\times\; \left(1 \;\pm\; z_{\alpha/2}\big(\widehat{\sigma}_{\mathrm{MLE}}/\sqrt{n}\big)\sqrt{\frac{\nu+3}{\nu+1} + \frac{\nu+3}{2\nu}\left[F_0^{-1}(1-\beta)\right]^2}\right),$$

where F_0 denotes either the standard normal or standard t_8 cdf. The respective values for the degrees of freedom ν are: $\nu \to \infty$ and $\nu = 8$. Table 2 presents empirical, parametric, and robust parametric point estimates and 95% interval estimates of $\mathrm{VaR}(\beta)$ for several levels of β.

TABLE 2: Point estimates and 95% confidence intervals of various value-at-risk measures computed by employing empirical, parametric (MLE), and robust parametric (T3,T4) methodologies.

Risk	Estimation Methodology				
Measure	EMPIRICAL	LOGNORMAL		LOG-t_8	
VaR(β)		MLE	T3	MLE	T4
$\beta = 0.25$	2,058	2,480	2,203	2,242	2,112
	(1,830; 2,268)	(2,133; 2,827)	(1,960; 2,446)	(1,970; 2,514)	(1,867; 2,357)
$\beta = 0.10$	4,555	6,595	4,607	5,155	4,512
	(3,758; 5,974)	(5,472; 7,719)	(3,973; 5,242)	(4,355; 5,955)	(3,821; 5,203)
$\beta = 0.05$	7,731	12,372	7,422	9,437	7,850
	(6,905; 11,339)	(9,979; 14,764)	(6,244; 8,601)	(7,703; 11,171)	(6,410; 9,290)
$\beta = 0.01$	26,791	41,753	18,856	38,693	28,788
	(20,800; 84,464)	(31,646; 51,860)	(15,025; 22,686)	(28,829; 48,558)	(21,360; 36,217)

A number of conclusions emerge from the table. First, for risk evaluations of moderate significance levels ($\beta = 0.25,\ 0.10$), where both parametric models are in close agreement with the data, there is essentially no difference between the empirical and robust parametric methodologies as all the point and interval estimates are similar. Second, for more extreme significance levels ($\beta = 0.05,\ 0.01$), the lognormal model does not fit the data well, which translates into its underestimation of the empirical

risk. Third, the robust point estimates of the risk based on the log-t_8 model and the empirical approach are very close for all levels of β because the MTM fit of the log-t_8 distribution is excellent (see Figure 5). Fourth, the MLE fits are mediocre for both models and therefore its risk evaluations are off the empirical target. Fifth, aside from the robust parametric model's stretchability beyond the range of the observed data, the main advantage of such methodology over the empirical one is that it produces substantially shorter confidence intervals.

Remark 3: As is clear from the results of Table 2 and from the accompanying discussion, one should always search for a model that fits the data at hand well, and then apply it for further actuarial modeling. If in particular situation one cannot identify a 'good' model for the data, but there is enough data, the tail-risk measures should be computed using the empirical approach and wider confidence intervals would have to be tolerated. If there is no good model nor sufficient amount of data, then extreme-value theory should be consulted. In the latter case, however, robust fitting procedures supplemented with graphical diagnostics would also be useful. □

5 Conclusions

In this article, we have introduced two graphical tools, QQP-plot and PR-plot, and have shown that they can be useful for choosing the trimming proportions of an MTM estimator, for assessing the overall goodness of model fit, for model selection, and for identification of outliers. Incorporation of such tools into the robust-efficient model-fitting process (of the Norwegian fire claims data) have yielded several conclusions/insights: (*i*) robustness and efficiency must be considered simultaneously, for robust but inefficient procedures may capture spurious patterns in the sample, (*ii*) while parsimony is always relevant, it is worthwhile to pursue *reasonably* complex models that can accomodate more data; to paraphrase Einstein, a model should be as simple as possible but not simpler, (*iii*) thinking outside-the-model is also important; that is, knowledge about further modeling objectives and practical usage of the model helps and should play a role in the fitting procedure.

Further, we have demonstrated that the quality of the model fit is very important for actuarial modeling. Moreover, graphical tools can be used to project the fitted model effects on subsequent analysis. For example, without performing actual calculations, we can get a hint from the QQP- and

PR-plots on how the model will act on a risk measure. Specifically, if, for some segment of the risk, the fitted model under- or over-estimates corresponding empirical quantiles, then the risk evaluations for that segment will be below or above the observed (empirical) risk. The choice of risk measure in this case is not essential but the choice of its estimation method is.

Finally, the plots we have presented here are exclusively designed for the MTM estimators, though similar graphs can be constructed using other robust estimators (e.g., M-estimators) as well. One would need to figure out a one-to-one relationship between the empirical percentile levels and tuning constants of the estimator, and then modify the plots accordingly. However, no matter what statistical tools we select, one thing is certain, namely, a carefully constructed plot can enrich any model-fitting procedure. In summary, while statistical inference (e.g., parameter estimation and hypothesis testing) alone is designed to *extract* and *reduce* information from the sample, the inference supplemented with visualization can *summarize* and *retain* it.

Acknowledgment

The author is very appreciative of valuable insights and comments provided by three anonymous referees, leading to many improvements in the paper. Also, the support provided by a grant from the Actuarial Foundation, the Casualty Actuarial Society, and the Society of Actuaries is gratefully acknowledged.

References

[1] Beirlant, J., Teugels, J.L., and Vynckier, P. (1996). *Practical Analysis of Extreme Values*. Leuven University Press, Leuven, Belgium.

[2] Brazauskas, V., Jones, B., and Zitikis, R. (2009). Robust fitting of claim severity distributions and the method of trimmed moments. *Journal of Statistical Planning and Inference*, **139**(6), 2028–2043.

[3] Cleveland, W.S. (1993). *Visualizing Data*. Hobart Press, Summit, New Jersey.

[4] Cowell, F.A. and Victoria-Feser, M.P. (2006). Distributional dominance with trimmed data. *Journal of Business and Economic Statistics*, **24**, 291–300.

[5] Cowell, F.A. and Victoria-Feser, M.P. (2007). Robust stochastic dominance: a semiparametric approach. *Journal of Economic Inequality*, **7**, 21–37.

[6] Dell'Aquila, R. and Embrechts, P. (2006). Extremes and robustness: a contradiction? *Financial Markets and Portfolio Management*, **20**(1), 103–118.

[7] Dupuis, D. and Victoria-Feser, M.P. (2006). Robust prediction error criterion for Pareto modeling of upper tails. *Canadian Journal of Statistics*, **34**(4), 639–658.

[8] Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2004). Robustness. In *Encyclopedia of Actuarial Science* (B. Sundt and J. Teugels, eds.), volume **3**, 1515–1529; Wiley, London.

[9] Kaiser, T. and Brazauskas, V. (2006). Interval estimation of actuarial risk measures. *North American Actuarial Journal*, **10**(4), 249–268.

[10] Klugman, S.A., Panjer, H.H., and Willmot, G.E. (2004). *Loss Models: From Data to Decisions*, 2nd edition. Wiley, New York.

[11] Marceau, E. and Rioux, J. (2001). On robustness in risk theory. *Insurance: Mathematics and Economics*, **29**, 167–185.

[12] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

[13] Serfling, R. (2002). Efficient and robust fitting of lognormal distributions (with discussion). *North American Actuarial Journal*, **6**(4), 95–109. Discussion: **7**(3), 112–116. Reply: **7**(3), 116.

[14] Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.