The Double Pareto-Lognormal Distribution – A New Parametric Model for Size Distributions.

William J. Reed*
Department of Mathematics and Statistics,
University of Victoria,
PO Box 3045, Victoria, B.C.,
Canada V8W 3P4
(e-mail:reed@math.uvic.ca).
and
Murray Jorgensen
Department of Statistics
University of Waikato
Private Bag 3105, Hamilton
New Zealand.
(e-mail:maj@waikato.ac.nz).

July, 2000. Revised January, October, 2003

*Research supported by NSERC grant OGP 7252 and originated at the Department of Mathematics and Statistics, University of Melbourne, whose support and hospitality are gratefully acknowledeged.

Abstract

A family of probability densities, which has proved useful in modelling the size distributions of various phenomena, including incomes and earnings, human settlement sizes, oil-field volumes and particle sizes, is introduced. The distribution, named herein as the *double Pareto-lognormal* or *dPlN* distribution, arises as that of the state of a geometric Brownian motion (GBM), with lognormally distributed initial state, after an exponentially distributed length of time (or equivalently as the distribution of the killed state of such a GBM with constant killing rate). A number of phenomena can be viewed as resulting from such a process (e.g. incomes, settlement sizes), which explains the good fit. Properties of the distribution are derived and estimation methods discussed. The distribution exhibits Paretian (power-law) behaviour in both tails, and when plotted on logarithmic axes, its density exhibits hyperbolic-type behaviour.

Keywords: size distribution; Pareto law; power-law distribution; fat tails; EM algorithm; WWW file size; financial returns.

1 Introduction.

The purpose of this paper is to describe a new distribution which has proved to be very useful in modelling the size distributions of various phenomena arising in a wide range of areas of inquiry. These include economics (distributions of incomes and earnings); finance (stock price returns); geography (populations of human settlements); physical sciences (particle sizes) and geology (oil-field volumes). A glance ahead to the figures in Sec. 5 will indicate how well the distribution fits such data.

The distribution, which has four parameters, is somewhat similar in form to the log-hyperbolic distribution (Barndorff-Nielsen, 1977) and like that distribution exhibits power-law (Paretian) behaviour asymptotically in both tails. Also like the log-hyperbolic, it can be derived as a mixture of lognormal distributions. However the form of the mixing is in a sense more natural, arising from the distribution of the final state of a geometric Brownian motion (GBM), killed or stopped with a constant killing rate (or equivalently as the state of a GBM after an exponentially distributed period of evolution)[1]. This appears to be the reason behind the excellent fits obtained for much empirical size data. Also the new distribution is somewhat simpler to handle analytically than the log-hyperbolic.

In the next two sections the distribution is defined and its properties, and those of its close relative, the Normal-Laplace distribution, are presented. Es-

[1] Generalized hyperbolic distributions can also arise as killed GBMs, albeit with more complicated killing rate functions. See Eberlein (2001) for a discussion of the genesis of hyperbolic distributions by subordination.

timation is discusssed in Section 4 and the paper concludes in Section 5 with some examples illustrating the excellent fit of the model to a variety of different empirical size distributions.

2 Genesis and definitions.

Consider a geometric Brownian motion (GBM) defined by the Itô stochastic differential equation

$$dX = \mu X dt + \sigma X dw \tag{1}$$

with initial state $X(0) = X_0$ distributed lognormally, $\log X_0 \sim N(\nu, \tau^2)$. After T time units the state $X(T)$ will also be distributed lognormally with

$$\log X(T) \sim N(\nu + (\mu - \sigma^2/2)T, \tau^2 + \sigma^2 T). \tag{2}$$

Suppose now that the time T at which the process is observed is an exponentially distributed random variable with density$f_T(t) = \lambda e^{-\lambda t}, \quad t > 0$ (or equivalently that the process is 'killed' (*e.g.* Karlin & Taylor, 1981) with constant killing rate $k(X) \equiv \lambda$ and the final 'killed' state observed). The distribution of the state $\hat{X}$, say, at the time of observation or killing is a mixture of lognormal random variables (2) with mixing parameter T.

To find the distribution of $\hat{X}$ it is easiest to work in the logarithmic scale. Thus let $\hat{Y} = \log \hat{X}$ (so that $\hat{Y}$ is the state of an ordinary Brownian motion after an exponentially distributed time). The distribution of $\hat{Y}$ can be shown (see Appendix in Reed, 2003) to be that of the sum of independent random variables W and Z say, where Z follows an $N(\nu, \tau^2)$ distribution; and W follows a a skewed Laplace distribution (see *e.g.* Kotz *et al.*, 2001) with probability density function (pdf)

$$f_W(w) = \begin{cases} \frac{\alpha\beta}{\alpha+\beta}\, e^{\beta w}, & \text{for } w \leq 0 \\ \frac{\alpha\beta}{\alpha+\beta}\, e^{-\alpha w}, & \text{for } w > 0 \end{cases} \tag{3}$$

where α and $-\beta$ $(\alpha, \beta > 0)$ are the roots of the characteristic equation

$$\frac{\sigma^2}{2} z^2 + (\mu - \frac{\sigma^2}{2}) z - \lambda = 0. \tag{4}$$

The distribution of $\hat{Y}$ can be obtained by the convolution of the Laplace and normal densities. It is most conveniently expressed in terms of the *Mills' ratio* of the complementary cumulative distribution function (cdf) to the pdf of a standard normal distribution:

$$R(z) = \frac{\Phi^c(z)}{\phi(z)}.$$

With some algebra, the pdf of $\hat{Y}$ can be shown to be

$$g(y) = \frac{\alpha\beta}{\alpha+\beta}\, \phi\left(\frac{y-\nu}{\tau}\right) \; \left[R\left(\alpha\tau - (y-\nu)/\tau\right) + R\left(\beta\tau + (y-\nu)/\tau\right)\right]. \tag{5}$$

We shall refer to this as the *Normal-Laplace distribution* and write $Y \sim NL(\alpha, \beta, \nu, \tau^2)$ to indicate that Y follows this distribution.

Note that since a Laplace random variable can be expressed as the difference between two exponentially distributed variates (Kotz *et al.* 2001), a $NL(\alpha, \beta, \nu, \tau^2)$, random variable, Y, can be expressed as

$$Y \stackrel{d}{=} \nu + \tau Z + E_1/\alpha - E_2/\beta \tag{6}$$

where E_1, E_2 are independent standard exponential deviates and Z is a standard normal deviate independent of E_1 and E_2. This is the easiest way to generate pseudo-random numbers from the NL distribution. The pdf of $\hat{X}$ is easily found from (5). It can be expressed in terms of the Mills' ratio as

$$f(x) = \frac{1}{x}\, g(\log x) \tag{7}$$

or alternatively in terms of the cdf and complementary cdf Φ and Φ^c of $N(0,1)$, as

$$\begin{aligned} f(x) = \tfrac{\alpha\beta}{\alpha+\beta}\Big[&A(\alpha,\nu,\tau)x^{-\alpha-1}\Phi\left(\tfrac{\log x-\nu-\alpha\tau^2}{\tau}\right) + \\ &x^{\beta-1}A(-\beta,\nu,\tau)\Phi^c\left(\tfrac{\log x-\nu+\beta\tau^2}{\tau}\right)\Big] \end{aligned} \tag{8}$$

where

$$A(\theta,\nu,\tau) = \exp(\theta\nu + \alpha^2\tau^2/2). \tag{9}$$

We shall refer to this distribution as the *double Pareto-lognormal distribution* and write

$$X \sim dPlN(\alpha,\beta,\nu,\tau^2)$$

to indicate that a random variable X follows this distribution. Clearly (from (6)) a $dPlN(\alpha,\beta,\nu,\tau^2)$ random variable can be represented as

$$X \overset{d}{=} UV_1/V_2 \tag{10}$$

where U, V_1 and V_2 are independent, with U lognormally distributed ($\log U \sim N(\nu,\tau^2)$) and with V_1 and V_2 following *Pareto* distributions with parameters α and β respectively *i.e.* with pdf

$$f(v) = \theta v^{-\theta-1}, \quad v > 1$$

with $\theta = \alpha$ and $\theta = \beta$ respectively. Alternatively we can write

$$X \overset{d}{=} UQ, \tag{11}$$

where Q is the ratio of the above Pareto random variables, so that Q has pdf

$$f(q) = \begin{cases} \frac{\alpha\beta}{\alpha+\beta}\, q^{\beta-1}, & \text{for } 0 < q \leq 1 \\ \frac{\alpha\beta}{\alpha+\beta}\, q^{-\alpha-1}, & \text{for } q > 1 \end{cases} \tag{12}$$

We shall refer to the distribution (12) as the *double Pareto distribution* – hence the name double Pareto lognormal distribution for the distribution of $\hat{X}$ (since such a distribution results from the product of independent double Pareto and lognormal components).

To generate pseudo-random deviates from the $dPlN(\alpha, \beta, \nu, \tau^2)$ distribution, one can exponentiate pseudo-random deviates from $NL(\alpha, \beta, \nu, \tau^2)$ generated using (6).

3 Some properties.

Since most of the results concerning the dPlN distribution are most easily derived using the Normal-Laplace we present results for that distribution first.

3.1 The Normal-Laplace distribution.

Two special cases of the Normal-Laplace distribution are of interest, corresponding to $\alpha = \infty$ and $\beta = \infty$. In the latter case the NL distribution is that of the sum of independent normal and exponentially distributed components and exhibits extra normal variation (*i.e.* has a fatter tail than the normal) only in the upper tail. The pdf (5) in this case reduces to

$$g_1(y) = \alpha\phi\left(\frac{y-\nu}{\tau}\right) R\left(\alpha\tau - (y-\nu)/\tau\right) \tag{13}$$

Similarly with $\alpha = \infty$ the NL distribution is that of the difference between independent normal and exponential components, exhibiting extra normal variation only in the lower tail, and its pdf is

$$g_2(y) = \alpha\phi\left(\frac{y-\nu}{\tau}\right) R\left(\beta\tau + (y-\nu)/\tau\right). \tag{14}$$

We shall refer to these distributions respectively as the right-handed and left-handed normal-exponential distributions and use the notation $Y \sim NE_r(\alpha,\nu,\tau^2)$ to indicate that Y has the pdf $g_1(y)$; and $Y \sim NE_l(\beta,\nu,\tau^2)$ when Y has pdf $g_2(y)$.

We now give some properties of the general $NL(\alpha,\beta,\nu,\tau^2)$ distribution.

- *Cumulative distribution function.* A closed-form expression for the cdf of $NL(\alpha,\beta,\nu,\tau^2)$ can be obtained. It is

$$G(y) = \Phi\left(\frac{y-\nu}{\tau}\right) - \phi\left(\frac{y-\nu}{\tau}\right)\frac{\beta R(\alpha\tau - (y-\nu)/\tau) - \alpha R(\beta\tau + (y-\nu)/\tau)}{\alpha+\beta} \tag{15}$$

This expression is useful for calculating cell probabilities when fitting the model to grouped data.

- *Moment generating function (mgf).* From the representation (6) it follows that the mgf of $NL(\alpha,\beta,\nu,\tau^2)$ is the product of the mgfs of its normal and Laplace components. Precisely it is

$$M_Y(s) = \frac{\alpha\beta\exp(\nu s + \tau^2 s^2/2)}{(\alpha-s)(\beta+s)}. \tag{16}$$

- *Mean and variance.* Expanding the cumulant generating function, $K_Y(s) = \log M_Y(s)$, yields

$$E(Y) = \nu + 1/\alpha - 1/\beta; \qquad \mathrm{var}(Y) = \tau^2 + 1/\alpha^2 + 1/\beta^2 \tag{17}$$

The third and fourth order cumulants are

$$\kappa_3 = 2/\alpha^3 - 2/\beta^3; \qquad \kappa_4 = 6/\alpha^4 + 6/\beta^4. \tag{18}$$

- *Representation as a mixture.* The $NL(\alpha, \beta, \nu, \tau^2)$ can be represented as a mixture of mixture of right-handed and left handed normal-exponential distributions:

$$g(y) = \frac{\beta}{\alpha+\beta} g_1(y) + \frac{\alpha}{\alpha+\beta} g_2(y). \tag{19}$$

where g_1 and g_2 are the pdfs of $NE_r(\alpha, \nu, \tau^2)$ and $NE_l(\beta, \nu, \tau^2)$ respectively.

- *Closure under linear transformation.* The NL distribution is closed under linear transformation. Precisely if $Y \sim NL(\alpha, \beta, \nu, \tau^2)$ and a and b are any constants, then $aY + b \sim NL(\alpha/a, \beta/a, a\nu + b, a^2\tau^2)$.

- *Infinite divisibility.* The NL distribution is infinitely divisible. This follows from writing its mgf as

$$M_Y(s) = \left[\exp(\frac{\nu}{n} s + \frac{\tau^2}{2n} s^2) \left(\frac{\alpha}{\alpha - s} \right)^{1/n} \left(\frac{\beta}{\beta + s} \right)^{1/n} \right]^n$$

for any integer $n > 0$ and noting that the term in square brackets is the mgf of a random variable formed as $Z + G_1 - G_2$, where Z, G_1 and G_2 are independent and $Z \sim N(\frac{\nu}{n}, \frac{\tau^2}{n})$ and G_1 and G_2 have gamma distributions with parameters $1/n$ and α and $1/n$ and β respectively. The infinite divisibility implies that it is possible to construct a Lévy process with increments following the NL distribution. Such a process could be used to model the logarithmic returns of financial instruments (stock prices, foreign currency prices *etc.*) reflecting the fact that observed logarithmic returns for high frequency data have fatter tails than those of the normal distribution (see Sec. 5.5).

3.2 The double Pareto-lognormal distribution.

Corresponding to right-handed and left-handed normal-exponential distributions arising as the two limiting cases of the normal-Laplace distribution are the right-handed and left-handed Pareto-lognormal distributions with pdfs

$$f_1(x) = \alpha x^{-\alpha-1} A(\alpha,\nu,\tau)\Phi\left(\frac{\log x - \nu - \alpha\tau^2}{\tau}\right). \tag{20}$$

and

$$f_2(x) = \beta x^{\beta-1} A(-\beta,\nu,\tau)\Phi^c\left(\frac{\log x - \nu + \beta\tau^2}{\tau}\right). \tag{21}$$

which are the limiting forms (as $\beta \to \infty$ and $\alpha \to \infty$) of the $dPlN(\alpha,\beta,\nu,\tau^2)$ distribution. Colombi (1990) considered the distribution (20), which he called the Pareto-lognormal, as a model for income distibrutions.

- *Representation as a mixture.* From (19) it follows that the $dPlN(\alpha,\beta,\nu,\tau^2)$ distribution can be represented as a mixture as

$$f(x) = \frac{\beta}{\alpha+\beta} f_1(x) + \frac{\alpha}{\alpha+\beta} f_2(x). \tag{22}$$

- *Cumulative distribution function.* The cdf of $dPLN(\alpha,\beta,\nu,\tau^2)$ can be written either as $F(x) = G(e^x)$ where G is given by (15); or as

$$\begin{aligned} F(x) = \quad & \Phi\left(\tfrac{\log x - \nu}{\tau}\right) - \tfrac{1}{\alpha+\beta}\Big[\beta x^{-\alpha} A(\alpha,\nu,\tau)\Phi\left(\tfrac{\log x - \nu - \alpha\tau^2}{\tau}\right) + \\ & \alpha x^{\beta} A(-\beta,\nu,\tau)\Phi^c\left(\tfrac{\log x - \nu + \beta\tau^2}{\tau}\right)\Big] \end{aligned} \tag{23}$$

- *Power-law tail behaviour.* The $dPlN(\alpha,\beta,\nu,\tau^2)$ distribution exhibits power-law (or Paretian) behaviour in both tails in the sense that

$$f(x) \sim k_1\ x^{-\alpha-1}\ \ (x \to \infty); \qquad f(x) \sim k_2\ x^{\beta-1}\ \ (x \to 0)$$

where $k_1 = \alpha A(\alpha, \nu, \tau)$ and $k_2 = \beta A(-\beta, \nu, \tau)$. The cdf $F(x)$ and complementary cdf $S(x) = 1 - F(x)$ also exhibit power-law tail behaviour with

$$S(x) \sim A(\alpha, \nu, \tau)\ x^{-\alpha} \quad (x \to \infty); \qquad F(x) \sim A(-\beta, \nu, \tau)\ x^{\beta} \quad (x \to 0).$$

The limiting (Pareto-lognormal) distribution ($\beta = \infty$) with pdf $f_1(x)$ exhibits only upper-tail power-law behaviour; while the other limiting (Pareto-lognormal) distribution ($\alpha = \infty$) with pdf $f_2(x)$ exhibits only lower-tail power-law behaviour.

• *Shape of distribution.* The $dPlN(\alpha, \beta, \nu, \tau^2)$ pdf is unimodal if $\beta > 1$ and is monotonically decreasing when $0 < \beta < 1$ (see Fig. 1, top row). Like the log-hyperbolic pdf, when plotted on logarithmic axes, the $dPlN(\alpha, \beta, \nu, \tau^2)$ pdf has a shape similar to a hyperbola, with asymptotes of slope $-(\alpha+1)$ and $\beta - 1$. In the case $0 < \beta < 1$, both arms have negative slope (see Fig. 1, bottom row). If the dPlN distribution arises as the final state of a killed GBM, $\beta < 1$ if and only if $\lambda < \sigma^2 - \mu$.

• *Moments.* The moment generating function does not exist. However lower-order moments about zero are easy to obtain. They are

$$\mu'_r = \mathrm{E}(X^r) = \frac{\alpha\beta}{(\alpha - r)(\beta + r)} \exp\left(r\nu + r^2\tau^2/2\right) \tag{24}$$

for $r < \alpha$. As with the Pareto distribution μ'_r does not exist for $r \geq \alpha$. The mean (for $\alpha > 1$) is

$$\mathrm{E}(X) = \frac{\alpha\beta}{(\alpha - 1)(\beta + 1)} e^{\nu + \tau^2/2} \tag{25}$$

while the variance and coefficient of variation (for $\alpha > 2$) are

$$\mathrm{var}(X) = \frac{\alpha\beta e^{2\nu + \tau^2}}{(\alpha - 1)^2(\beta + 1)^2} \left[\frac{(\alpha - 1)^2(\beta + 1)^2}{(\alpha - 2)(\beta + 2)} e^{\tau^2} - \alpha\beta\right] \tag{26}$$

and

$$\text{CV} = \left[\frac{(\alpha-1)^2(\beta+1)^2)}{\alpha\beta(\alpha-2)(\beta+2)} e^{\tau^2} - 1 \right]^{1/2}$$

Clearly the CV is independent of ν, increases with τ^2 and decreases with α and β.

• *Closure under power-law transformations.* The dPlN family of distributions is closed under power-law transformation. *i.e.* if $X \sim dPlN(\alpha, \beta, \nu, \tau^2)$, then for constants $a, b > 0$, $W = aX^b$ will also follow a dPlN distribution. Precisely

$$W = aX^b \sim dPlN(\alpha/b, \beta/b, b\nu + \log a, b^2\tau^2). \tag{27}$$

4 Estimation.

4.1 Method of Moments.

Given data assumed to be from the dPlN distribution one could, in principle, obtain method of moments estimates (MMEs) of α, β, ν and τ^2 using the first four moments of either the dPlN distribution or, having first log-transformed the data, of the NL distribution. The estimates are not the same. Use of the dPlN moments (with untransformed data) however is not recommended, since population moments of order α or greater do not exist. Thus there is in effect a lower bound (of 4) on the MME of α.

To find MMEs of α and β using the NL distribution, one needs only solve (18), with κ_3 and κ_4 set to their sample equivalents. Estimates of ν and τ can then be obtained from (17). Experience with simulated data has shown that occasionally (18) has no real solution. This is of no serious consequence since estimates can be obtained by maximum likelihood and we recommend the use of the method of moments only for finding starting values for iterative procedures for finding maximum likelihood estimates. In their absence trial and error can be used.

4.2 Maximum Likelihood.

Unlike MMEs maximum likelihood estimates (MLEs) are the same whether one fits the dPlN to data $x_1, x_2, \ldots, x_n$ or fits the NL to $y_1 = \log x_1, \ldots, y_n = \log x_n$. The log likelihood function is

$$\ell = n \log \alpha + n \log \beta - n \log(\alpha+\beta) + \sum_{i=1}^{n} \phi\left(\frac{y_i - \nu}{\tau}\right) + \sum_{i=1}^{n} \log\left[R(p_i) + R(q_i)\right] \quad (28)$$

where

$$p_i = \alpha\tau - (y_i - \nu)/\tau \qquad q_i = \beta\tau + (y_i - \nu)/\tau \quad (29)$$

This can be maximized analytically over ν to yield

$$\hat{\nu} = \bar{y} - 1/\alpha + 1/\beta \quad (30)$$

and a concentrated (profile) likelihood

$$\begin{aligned} \hat{\ell}(\alpha, \beta, \tau) \quad = \quad & n \log \alpha + n \log \beta - n \log(\alpha+\beta) + \textstyle\sum_{i=1}^{n} \phi\left(\frac{y_i - \bar{y} + 1/\alpha - 1/\beta}{\tau}\right) + \\ & \textstyle\sum_{i=1}^{n} \log\left[R(\alpha\tau - \frac{y_i - \bar{y} + 1/\alpha - 1/\beta}{\tau}) + R(\beta\tau + \frac{y_i - \bar{y} + 1/\alpha - 1/\beta}{\tau})\right]. \end{aligned} \quad (31)$$

This can be maximized numerically using the MMEs as starting values. Note that it is not difficult to code the score function, nor the elements of the Hessian matrix for obtaining the observed information matrix.

For grouped data the log-likelihood is of the form

$$\ell(\alpha, \beta, \nu, \tau^2) = \sum_{j=1}^{N} f_j \log(G(y^{(j)}) - G(y^{(j-1)})) \quad (32)$$

where G is the cdf (15), and $-\infty = y^{(0)} < y^{(1)} < \ldots, y^{(N-1)} < y^{(N)} = \infty$ separate the cells $1, 2, \ldots, N$.

Both when using grouped and ungrouped data, one can fit either of the limiting (Pareto-lognormal) pdfs $f_1(x)$ or $f_2(x)$ in a similar fashion, maximizing over only three parameters. Experience has shown that it is worthwhile examining the data for evidence of power-law behaviour in the tails. If for example there is only power-law behaviour in one tail, attempts to fit the dPlN may result in the non-convergence of the optimization algorithm, while fitting the appropriate Pareto-lognormal pdf (f_1 or f_2) will be satisfactory. One can examine the data for power-law tail behaviour, both by plotting (on a logarithmic scale) the frequencies in a histogram for the logged data; and by plotting the empirical cdf or complementary cdf on logarithmic axes[2]. In both cases linearity in the plots suggests power-law behaviour.

4.3 EM Algorithm.

The representation in Section 2 of a normal-Laplace (NL) variable Y as the sum of independent random variables W and Z, (where Z follows an $N(\nu, \tau^2)$ distribution and W follows a a skewed Laplace distribution) suggests that an approach to ML estimation via the EM algorithm (Dempster, Laird, and Rubin (1977); McLachlan and Krishnan (1997); Jorgensen (2002)) may prove to be effective. The EM algorithm uses a likelihood function based on augmented data as a stepping stone towards the maximization of the likelihood based on

[2] This is equivalent to producing *rank-size* plots, in which the observations are plotted (on a logarithmic scale) against their ascending or descending rank (again on logarithmic scale). *Zipf's law*, which claims upper-tail power-law behaviour in the size distribution of cities is known in the urban geography literature as the *rank-size property.* - see Sec. 5.2.

the observed data.

The algorithm alternates between two phases. In the first phase, known as the E-step, the log-likelihood function based on the augmented data is made to depend on the original data alone by taking the expectation with respect to the conditional distribution of the augmented data given the original data. In the case where the augmented data log-likelihood has exponential family form, the E-step may be accomplished by taking expectations of the sufficient statistics.

In the M-step an improved set of parameter estimates is constructed so as to maximize the expected augmented-data log-likelihood. It is shown in the references given that the new set of parameter estimates cannot have lower (original data) log-likelihood than the previous set.

Suppose then that we have a sample $y_1, y_2, \ldots, y_n$ from the normal-Laplace distribution. We take the augmented data to be $z_1, z_2, \ldots, z_n$ and $w_1, w_2, \ldots, w_n$, where $z_i + w_i = y_i$ for $i = 1, \ldots, n$, with $z_1, z_2, \ldots, z_n$ a sample from $N(\nu, \tau^2)$, and $w_1, w_2, \ldots, w_n$ a sample from the skewed Laplace distribution (3).

The E-step of the can be carried out as follows. Define v_i to be 1 if $w_i \geq 0$, and 0 otherwise. It is easy to see that the augmented data log-likelihood depends on the augmented data as a linear function of

$$z_i, \quad z_i^2, \quad w_i, \quad \text{and} \quad w_i v_i.$$

We therefore need to be able to take expectations of quantities of those four forms with respect to the joint density of (z_i, w_i) conditional on y_i. As $z_i + w_i = y_i$ we need only the conditional density of w_i given y_i which more algebra shows to be

$$h(w_i|y_i) = \frac{\tau^{-1}\exp\left(t_i w_i/\tau - w_i^2/(2\tau^2)\right)}{R(q_i)+R(p_i)}$$

where

$$t_i = q_i(v_i-1) - p_i v_i = \left\{ \begin{array}{rcll} q_i & = & \beta\tau + \frac{y_i-\nu}{\tau} & \text{if } w_i < 0 \\ -p_i & = & -\alpha\tau + \frac{y_i-\nu}{\tau} & \text{if } w_i \geq 0 \end{array} \right.$$

from which we may obtain

$$\begin{aligned} E[w_i|y_i] &= \tau\frac{q_i R(q_i) - p_i R(p_i)}{R(q_i)+R(p_i)} = \hat{w}_i \text{ , say} \\ E[w_i v_i|y_i] &= \tau\frac{1 - p_i R(p_i)}{R(q_i)+R(p_i)} = \widehat{w_i v_i} \\ E[w_i^2|y_i] &= \tau^2\frac{(1+q_i^2)R(q_i) + (1+p_i^2)R(p_i) - \tau(\alpha+\beta)}{R(q_i)+R(p_i)} \ . \end{aligned}$$

The conditional expectations of z_i and z_i^2 given y_i follow easily from these, completing the E-step.

The M-step may be carried out almost as easily as if we had available independent random samples $z_1, z_2, \ldots, z_n$ from $N(\nu, \tau^2)$ and $w_1, w_2, \ldots, w_n$ from the skewed Laplace distribution with parameters α and β, because in the augmented-data log-likelihood ℓ_c the two distributions are effectively decoupled:

$$\ell_c = -n\log\tau - \frac{1}{\tau^2}\sum_{i=1}^{n}(z_i-\nu)^2 +$$

$$n\log\alpha + n\log\beta - n\log(\alpha+\beta) + \beta\sum_{i=1}^{n} w_i(1-v_i) - \alpha\sum_{i=1}^{n} w_i v_i \ .$$

Firstly the updated estimates of ν and τ are obtained from conditional expectations of $\sum z_i$ and $\sum z_i^2$ in the usual way. Then define $A = \sum_i \widehat{w_i v_i}/n$ and $B = A - \sum_i \hat{w}_i/n$. The updated estimates of α and β are then

$$\frac{1}{A+\sqrt{AB}} \quad \text{and} \quad \frac{1}{B+\sqrt{AB}}$$

respectively. The E-step and the M-step are then repeated until convergence is reached.

The same caveat regarding whether to fit a dPlN distribution as opposed to a Pareto-lognormal with power-law behaviour in only one tail (as discussed at the end of Sec. 4.2) is appropriate here also. If there is power-law behaviour in only one tail, EM algorithm will be unlikely to converge – it will proceed making one or other of the parameters α or β progressively larger.

In terms of computing time the EM procedure appears to be more efficient than the numerical maximization of the 3-parameter concentrated log-likelihood (Sec. 4.2), probabaly due to the fact that the M-step in the EM procedure is accomplished analytically. However programming the EM procedure is somewhat more time consuming than simply programming the use of a numerical optimization routine.

5 Some Applications.

In this section examples of the fit of the dPlN to various size-distribution datasets are discussed. In two cases the data are grouped (incomes, particle sizes) while in the others, the actual size observations are available (settlement sizes, oil-fields, stock price returns). Maximum likelihood estimates of parameters (see Table 1) were obtained under the assumption that the observations constituted a simple random sample. When the observations do not represent a true random sample, this method will not of course provide true maximum likelihood estimates. However it can be justified as providing *maximum likeness* estimates (Barndorff-Nielsen, 1977) which minimize the discrimination information between the fitted distribution and the data.

5.1 Earnings and income distributions.

Examples of the fit of the dPlN (with plots) to various earnings/income distributions are presented in Reed (2003). It was to explain such distributions (and Pareto's law of Incomes) that the dPLN (as the state of a killed GBM) was developed. The explanation revolves around the assumption that an individual's earnings (or family's income) follows GBM (an assumption based on *Gibrat's law of proportional effects* (Gibrat, 1931), common in the income distribution literature *e.g.* Champernowne, 1953) and that the population of individuals (or families) is approximately growing at a fixed rate. Starting incomes are assumed to be lognormally distributed and evolving as GBM. The assumption of a growing population implies that the time that an individual has been earning (or a family been in existence) is approximately exponentially distributed, and thus that current earnings or income follow close to that of a GBM killed with a constant killing rate.

5.2 Human settlements size.

It has long been recognized that the distribution of size (human population) of cities within a particular country or jurisdiction frequently exhibits Paretian behaviour in the upper tail. This phenomenon is known as the *rank size property* or in the case when the Pareto exponent is unity as *Zipf's law.* There have been many attempts to explain this phenomenon, two of the more recent being by

Gabaix (1999) and Brakman *et al.* (1999). However the fact that there can also be Paretian behaviour in the lower tail of the distribution of human settlement size appears to have escaped notice. Such behaviour in both tails is manifest in the dPlN distribution which turns out to fit settlement size data very well (see Reed, 2002, for examples). The good fit can be explained in a similar way to that of incomes/earning data. If it is assumed that the growth in size of settlements follows GBM (*e.g.* Gabaix, 1999), then provided that the time since foundation follows an exponential distribution, and that at foundation sizes are lognormally distributed, then the current sizes should follow the dPlN distribution. The approximate exponential form of the distribution of the time since foundation follows if the foundation of settlements occur in a *Yule process* (Yule, 1924) (i.e. homogeneous pure birth process) over a long time period. Such an assumption seems reasonable, corresponding to the situation in which existing settlements create satellites at a fixed probabilistic rate.

5.3 Particle size.

Fig. 2 shows the fit of the dPlN to grouped data on aeolian sand particle size and diamond particle size presented in Barndorff-Nielsen (1977) who fitted the log-hyperbolic distribution to these data. Visual inspection of Fig. 4 and of the graphs in Barndorff-Nielsen (1977) suggest very similar fits for the two models. It may be possible to view the size of a particle as the outcome of a killed multiplicative (geometric) stochastic process (*i.e.* as the result of a random number of random fractures), although perhaps not precisely as the result of GBM with a constant killing rate. In spite of the fact of no explanatory model for why particle sizes should exactly follow dPlN, the empirical fit is nonetheless extremely good.

5.4 Oil-field size.

Fig. 3 shows the dPlN distribution fitted to the volumes of 634 oil fields in the West Siberian Basin[3] (the world's largest oil province). The fit is very good except possibly in the extreme upper tail. An oil-field can be thought of as a percolation cluster (Stauffer and Aharony, 1992) and thus as a killed stochastic process, although not of course exactly as a killed GBM.

5.5 Stock price returns.

It has been recognized for some time (see *e.g* Rydberg, 2000 for a discussion) that the logarithmic returns $\log[P(t+1)/P(t)]$, of a stock whose price P_t is observed at discrete times $t = 1, 2, \ldots$ follow a distibution with fatter tails than that of the normal distribution predicted by the standard GBM model of stock price movement. Furthermore the departures from normality increase as the reporting interval shortens. Various alternatives have been proposed including the asymmetric Laplace distribution (*e.g.* Madan and Milne, 1991; Kozubowski and Podgorski, 2001); and the generalized hyperbolic (Eberlein, 2001). For both of these alternatives option pricing formulas (á la Black-Scholes) have been developed. The crucial fact that enables this is that in both cases the distributions are infinitely divisible, so that a Lévy process can be constructed for which the increments (representing logarithmic returns) have the given distribution.

[3] Data from **http://energy.cr.usgs.gov:8080/energy/WorldEnergy/OF97-463** a website of*U.S Dept. of Interior Geological Survey.*

The dPlN distribution is similar in form to the log-hyperbolic (and the NL similar to the hyperbolic) and provides another candidate for stock-price returns. Fig 4 shows the fit of the NL to the logarithmic daily returns using the closing price of IBM ordinary stock from Jan 1, 1999 to Sept. 18, 2003 (929 observations). The NL distribution is infinitely divisible and so it possible to construct a Lévy process for the movement of stock prices, based on this distribution. Option prices for this model can be evaluated using the characteristic function approach (*e.g.* Schoutens, 2003, p. 20).

5.6 Size of WWW sites and computer files – a potential application.

Huberman and Adamic (1999) have shown that the size distribution (number of pages) of World-Wide Web sites follows power-law behaviour in the upper-tail, while Mitzenmacher (2001) showed that both upper- and lower-tail power-law behaviour occurs in the distribution of file sizes. Both papers offerred an explanation analogous to the derivation of the double Pareto distribution described in Sec. 2, *viz.* that it results from a multiplicative process observed after a geometrically distributed number of steps. No attempt is made to fit a theoretical distribution to the data, in either paper. However Mitzenmacher points out that file size distributions 'have a lognormal body and Pareto tail'. This suggests the dPlN as an obvious candidate for such data.

REFERENCES

Barndorff-Nielsen, O. Exponentially decreasing distributions for the logarithm of particle size. Proc. R. Soc. Lond. A. **1977**, *353*, 401-419

Brakman, S., H. Garretsen, C. Van Marrewijk and M. van den Berg (1999) The return of Zipf: towards a further understanding of the rank-size distribution, J. Reg. Sci,, **1999**, *29*, 183-213.

Colombi, R. A new model of income distribution: ThePareto lognormal distribution. In *Income and wealth distribution, inequality and poverty*, C. Dagum and M. Zenga, (eds.), Springer, Berlin, 1990.

Champernowne, D. A model of income distribution, Econ. J., **1953**, *63*, 318-351.

Dempster, A.P., Laird, N.M., D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. **1977**, *B39*, 1-38.

Eberlein E. Applications of generalized hyperbolic Lévy motions to finance. *Lévy Processes: Theory and Applications.* O. Barndorff-Nielsen, T. Mikorsky and S. Resnick (eds.). Birkhäuser, Boston, 2001.

Gabaix, X. Zipf's Law for cities: an explanation, Quart. J. Econ., **1999**, *114*, 739-767.

Gibrat, R.*Les inégalités économiques*, Librairie du Recueil Sirey, Paris, 1931.

Huberman, B. and L. A. Adamic. Internet: Growth dynamics of the World-Wide web. Nature, **1999**, *401*, 130-131.

Jorgensen, M. Expectation-maximization algorithm, *Encyclopedia of Environmetrics*, A. H. El-Shaarawi and W. W. Piegorsch (eds). J. Wiley and Sons, New York, 2001.

Karlin, S. and H. M. Taylor. *An Introduction to Stochastic Processes*, Vol II, New York, Academic Press, 1981.

Kotz, S., T. J. Kozubowski and K. Podgórski. *The Laplace Distribution and Generalizations.* Birkhäuser, Boston, 2001.

T. J. Kozubowski and K. Podgórski. Asymmetric Laplace laws and modeling financial data, Math. Comput. Model., **2001**, *34*, 1003-1021.

Madan, D.B.and F. Milne. Option pricing with VG martingale components. Math. Fin., **1991**, *1* 39-55.

McLachlan, G.J. and T. Krishnan. *The EM algorithm and Extensions.* J. Wiley and Sons, New York, 1997.

Mitzenmacher, M. Dynamic models for file sizes and double Pareto distributions. **2001**, Draft manuscript avilable at

http://www.eecs.harvard.edu/ michaelm/NEWWORK/papers/.

Reed, W. J. The Pareto law of incomes - an explanation and an extension. Physica A, **2003**, *319*, 579-597.

Reed, W. J. On the rank-size distribution for human settlements. J. Reg. Sci., **2002**, *42* 1-17.

Rydberg, T. H. (2000). Realistic statistical modelling of financial data. Inter. Stat. Rev., **2000**, *68*, 233-258.

Schoutens, W. *Levy Processes in Finance*, J. Wiley and Sons, Chichester, 2003.

Stauffer, D and A. Aharony (1992). *Introduction to Percolation Theory*, London, Taylor and Francis, 1992.

Yule, G., (1924) A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F.R.S., Philos. Trans. B, Roy. Soc. London, **1924**, *213* 21-87.

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\nu}$	$\hat{\tau}^2$
Aeolian sand	9.22	3.87	-0.643	0.136
Diamonds	1.79	3.89	-0.339	0.606
Oil fields	1.13	2.27	3.25	2.375
IBM price returns	60.89	64.00	-1.065×10^{-3}	1.028×10^{-4}

Table 1: Maximum likelihood estimates of parameters for the examples in Sec.5

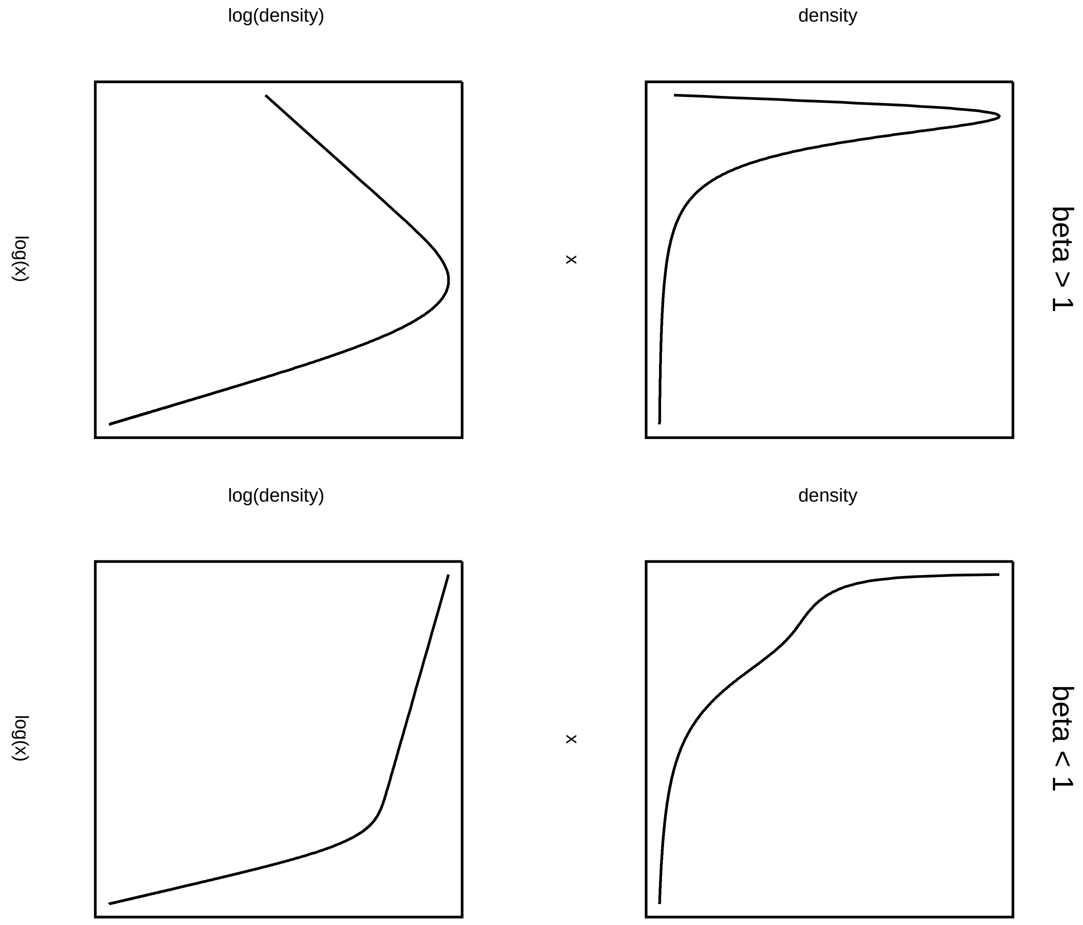
Figure captions.

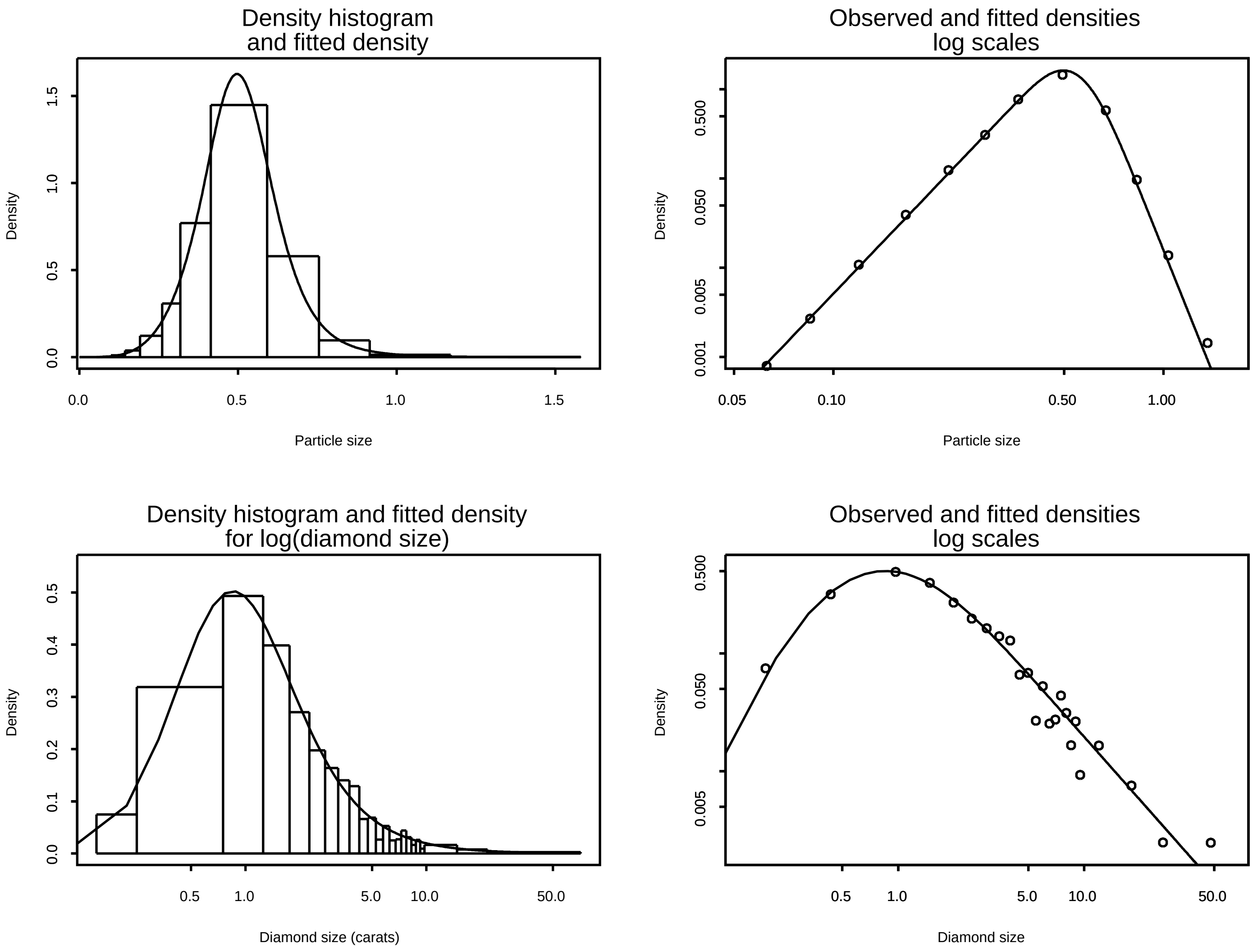
Fig.1 Two forms of the double Pareto-lognormal density, in the natural scale (top row) and logarithmic scales (bottom row). The panels in the left-hand column show the case $\beta > 1$ and those in the right-hand column the case $\beta < 1$.

Fig.2 Empirical and fitted dPlN distributions for aeolian sand particle size in millimetres (top row) and diamond size in carats from a South West African diamond mine (bottom row) using data in Barndorff-Nielsen (1977). In each case the left-hand panel shows a density histogram and the dPlN fitted density (size in logarithmic scale) while the right-hand panels shows the empirical density (calculated from histogram) and the fitted dPlN density (density and size both in logarithmic scales).

Fig.3 Empirical and fitted dPlN distributions for the volume (mmb) of 634 oil fields in the West Siberian basin. The left-hand panel shows a density histogram and the fitted dPlN density (volume in logarithmic scale). The centre panel shows the empirical density (calculated from histogram) and fitted dPlN distribution (density and volume both on logarithmic scales) and the right-hand panel shows a quantile-quantile plot of the empirical and fitted dPlN distributions (both in logarithmic scales).

Fig.4 Empirical and fitted dPlN distributions for the daily price returns on IBM common stock. The left-hand panel shows a density histogram and the fitted dPlN density. The centre panel shows the empirical density (calculated from histogram) and fitted dPlN distribution (density and price return both on logarithmic scales) and the right-hand panel shows a quantile-quantile plot of the empirical and fitted dPlN distributions.


log(density)
density
log(x)
x
beta > 1
log(density)
density
log(x)
x
beta < 1

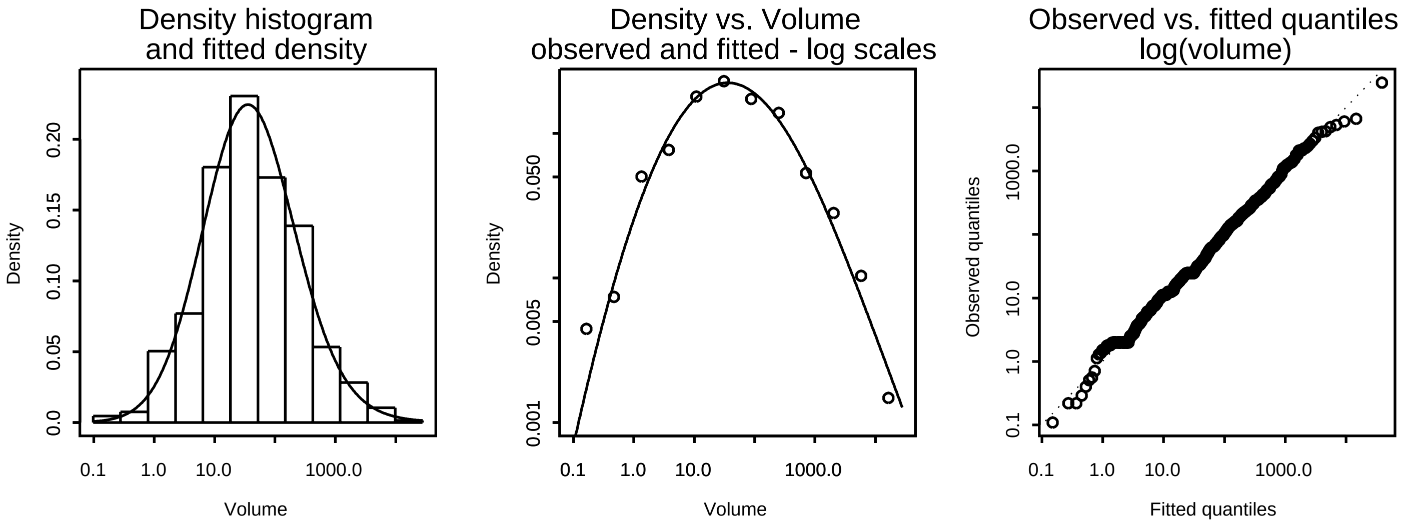


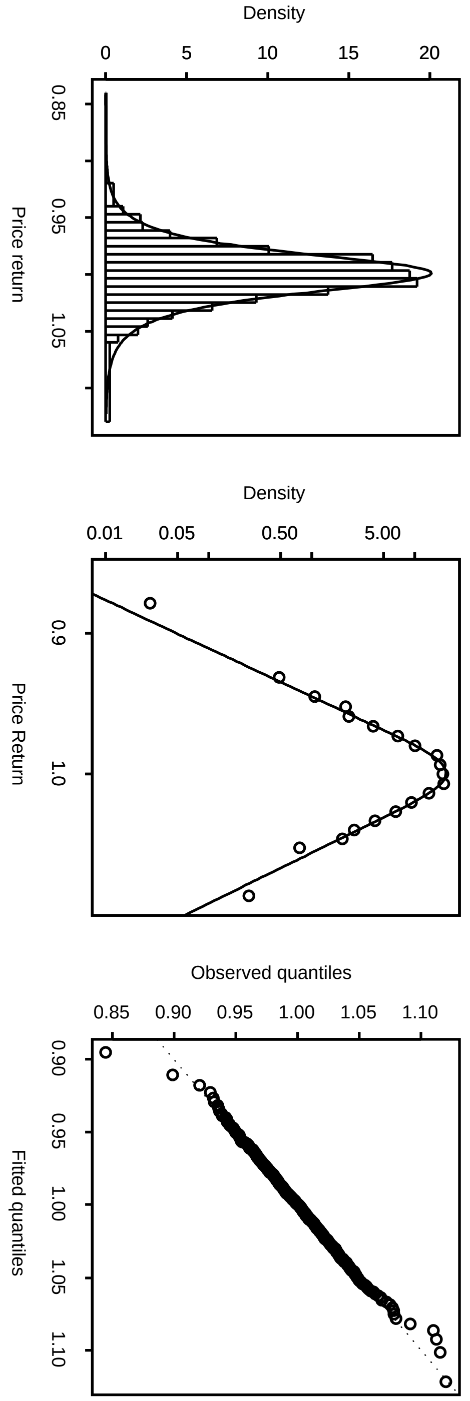
Density histogram and fitted density

Observed and fitted densities log scales

Density histogram and fitted density for log(diamond size)

Observed and fitted densities log scales


Density histogram
and fitted density
Density
0.0
0.05
0.10
0.15
0.20
Volume
0.1
1.0
10.0
1000.0
Density vs. Volume
observed and fitted - log scales
Density
0.001
0.005
0.050
Volume
0.1
1.0
10.0
1000.0
Observed vs. fitted quantiles
log(volume)
Observed quantiles
0.1
1.0
10.0
1000.0
Fitted quantiles
0.1
1.0
10.0
1000.0


Density
0
5
10
15
20
0.85
0.95
1.05
Price return
Density
0.01
0.05
0.50
5.00
0.9
1.0
Price Return
Observed quantiles
0.85
0.90
0.95
1.00
1.05
1.10
0.90
0.95
1.00
1.05
1.10
Fitted quantiles