Applications and Extensions of Three Statistical Models

A thesis presented

by

David Nathaniel Esch

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

May, 2003

Abstract

1. We present a new method for statistically deconvolving a Point Spread Function from a source image. This method includes some regularization, so it is ideally suited to extended source images. The regularization parameters are fit to the statistical model, so little user intervention is required. The output includes error information for the results, so confidence statements can be constructed from the results. A survey of popular deconvolution techniques is first presented. Then a review of the necessary statistical theory to develop our procedure is presented. Our model is then presented, starting with the likelihood, and then describing the regularization priors. The algorithmic details of fitting the model are then explained, and several examples are presented in detail.

2. We develop Generalized Linear models using Morris' NEF_6 distribution, the exponential family generated by the generalized hyperbolic secant distribution. First we review some basic properties of these distributions on $(-\infty, \infty)$. Then we specify the model and its features: quadratic variance function with no real roots, and tangent canonical link function. The doubly bounded parameter space for the natural exponential parameter presents some issues peculiar to this family of GLMs. This paper describes and develops a procedure for fitting the regression coefficients and the convolution parameter simultaneously via maximum likelihood. Some simulations from the model are used to assess model fits, and to make frequentist evaluations of our fitting procedure and confidence intervals based on Gaussian approximations. Finally, an analysis of stock return data shows the steps needed to improve model fits, illustrating several key features of this unusual model.

3. We present some basic properties of the Pearson type IV distribution, also known as the Skew-t distribution, as well as algorithms for computation of the distribution function and quantiles. This distribution arises as the conjugate prior for Morris' NEF-GHS distribution, and has many possible applications, heretofore largely unexplored, perhaps due to the lack of computational tools. Here we discuss basic properties of the distribution including its characteristic function and schemes for computation of the distribution, and review the existing literature.

Contents

Acknowledgements

The author would like to thank the primary advisors, Carl N. Morris, Ph. D. and David van Dyk, Ph. D., all three readers, as well as the primary research collaborator for part 1, Margarita Karovska, Ph. D., for their contributions to this work.

The author would also like to Donald Rubin, Ph. D., for his valuable comments regarding part 2, Robin Esch for his comments and ideas for part 3, and the rest of the faculty of the Statistics Department for their encouragement.

The research described in Part 1 was funded in part by NSF grant DMS 01-04129 and in part by NASA contract NAS8-39073 (CXC).

Part I
Image Deconvolution with Error Estimates: Hierarchical Multiscale Models for Poisson Images

1 Introduction

The process of discovering information in astronomical image data involves several technical problems best solved through statistical methods. At the center of these problems is the estimation of a true image which is observed imperfectly in several ways. Firstly, recorded events from the source of interest may be mixed with photons from other sources or background, and these extraneous sources must be separated or removed from the true image. Secondly, the observed events' spatial locations and possibly recorded energies can be shifted by the telescope's focusing and detecting apparatus, and so the pre-shifted positions and energies of the photons must be estimated. Thirdly, faint images are recorded with very few photons, and thus are imperfectly recorded as Poisson counts with intensity rates equal to the true image. Low-count observations will typically record one photon in an area with uniform low intensity, and thus some smoothing must be introduced to match our knowledge of the qualitative nature of the true image. Observations may also not be made for long enough times to adequately represent faint sources, which often are observed serendipitously on the side of an observation intended for another object, Thus the issues confronted are mainly source classification and background removal, statistical deconvolution, and smoothing or regularization of the image. Statistical modeling and estimation provide the perfect framework for solving these problems, especially in low-count imaging situations.

2 Currently used methods

Several methods are currently used in the astrophysics community for deconvolving Point Spread Functions (PSFs) from observations. These include the Richardson-Lucy technique (Richardson, 1972) and (Lucy, 1974), an implementation of EM for Poisson maximum likelihood with no smoothing, as well as adaptive smoothing methods, implemented in the CIAO software (Chandra Interactive Analysis of Observations) procedure ASMOOTH, and as a standalone IDL widget CSMOOTH. (Ebeling *et al.*, 2000) These two methods stand at opposite ends of a spectrum of methods designed to work well for point sources (Richardson-Lucy) or extended sources (Adaptive smoothing methods. Many other methods lie along this continuum, and they generally work by optimizing some criterion. The main issue with these methods is that they usually work well for only one type of source, point or extended, and the user must decide how to calibrate the algorithm. The Richardson-Lucy algorithm lacks any smoothing or regularization, and this is usually compensated for by stopping the algorithm before convergence. Many of the methods also do not produce uncertainty information about their results, and thus the user cannot evaluate the strength of the claims made based on the deconvolution. The *EMC2* procedure does produce error information, as well as some smoothing, with the added ability to effectively reconstruct point sources, and sharper features in the image.

We offer two improvements over the methods described above. Firstly, we use a wavelet-like multiscale representation of the true image to achieve smoothing at all scales of resolution simultaneously, thus capturing detailed features in the image at the same time as larger-scale extended features. Secondly, we use a principled fully Bayesian model-based analysis, which produces extensive information about the uncertainty of our conclusions, allowing the scientist to assess the relative strength of the inferential conclusions made based on the method. The fully Bayesian formulation allows a consistent top-to-bottom approach to both model and error

estimation and should produce better results than methods which process the data multiple times to produce their results.

In this paper we first provide the reader in Section 3 with an overview of the standard statistical tools that we employ including Bayesian modeling and inference; then in Section 4 we describe the statistical model used in our procedure, describing the likelihood and the prior distributions. Next in Section 5 we describe our techniques for fitting the model, and in Section 6 how the output from the procedure should be interpreted. After that we will show some examples in Section 7, and comment on the knowledge gained through the use of our method, and end in Section 9 with a discussion and conclusions.

3 Bayesian Modeling and Inference

Statistical analysis of an astronomical image composed of pixel counts begins with a *statistical model* that is a summary of both a *source* or *astrophysical model* for the image and the stochastic processes involved in data collection and recording. The source model describes the spatial structure in the relative intensities of photon emission in the source; this model is of principle scientific interest. In addition to the source model, the statistical model describes such instrumental effects as the point spread function (PSF) and other processes that effect the observed data, such as variability in the observed source counts and background contamination of the counts.

The statistical model is expressed in terms of a set of unknown parameters, which are of scientific interest in that they directly describe the image. The variability in the observed data is represented by a probability distribution, $p(X \mid \Pi)$, where X represents the counts and Π represents the astrophysical model or image; statisticians refer to Π as a *model parameter.* Consider, for example, an image composed of sixteen pixels in a four by four grid. The middle grid in Figure 1 represents $\Pi = \{\pi_i, i = 1, \ldots, 16\}$, where π_i is the expected proportion of the counts

recorded in pixel i; here $\sum_{i=1}^{16} \pi_i = 1$. We begin by assuming an ideal data set that is not subject to background contamination or blurred by a PSF. The third grid in Figure 1 represents such an ideal data set, $X = \{X_i, i = 1, \ldots, 16\}$; the first grid in the Figure will be discussed shortly.

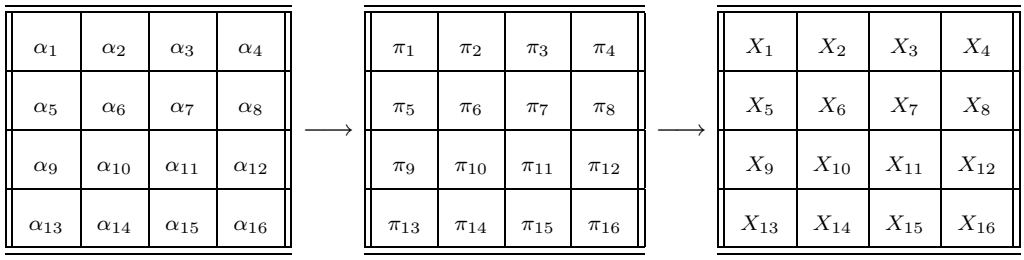


Figure 1: A sixteen pixel image. The first grid represents the sixteen Dirichlet prior parameters, the second grid represents the Π parameterization of the image, and the last grid represents the grid counts. Each π_i is the expected proportion of counts in pixel i for a perfect instrument that is not subject to a PSF or background contamination. Since no such perfect instrument exists, these counts are necessarily "missing data".

A standard mode of statistical inference "reconstructs the image" by computing the value Π that maximizes $p(X \mid \Pi)$ as a function of Π with X fixed at the values of observed counts. This estimate of Π is known as the maximum likelihood estimate (MLE) and

$$L(\Pi \mid X) = p(X \mid \Pi) \tag{3.1}$$

is called the *likelihood function.* (Although the likelihood function and $p(X \mid \Pi)$ are equal, the likelihood is viewed as a function of the parameter given the observed data while $p(X \mid \Pi)$ is viewed a function of the data for a given value of the parameter.) Notice that the MLE is the value of Π that maximizes $L(\Pi|X)$. Returning to the example illustrated in Figure 1, each photon that arrives at the detector has probability π_i of being counted in pixel i, and, the photons are counted independently of one another. Statistically this model is known as a *multinomial distribution*, specifically,

$$p(X|\Pi) = L(\Pi|X) \propto \prod_{i=1}^{16} \pi_i^{X_i}. \tag{3.2}$$

Under the constraint $\sum_{i=1}^{16} \pi_i = 1$, it is easy to verify that the MLE of Π is simply the set of count proportions, $\widehat{\Pi}_{\text{MLE}} = X/\sum_{i=1}^{16} X_i$.

A Bayesian statistical analysis is based not only on likelihood function, but is also based on a second component known as the *prior distribution*. The prior distribution quantifies information or knowledge about the likely values of the parameter or structures in the parameter that are known *prior* to data collection. Such prior information may include an expectation that the true image is smooth, i.e., the image intensities in nearby or adjacent regions of the image should be similar. This information is quantified through a (prior) probability function, that we denote $p(\Pi)$. In the methods that we describe in Sections 4 and 5 $p(\Pi)$ primarily quantifies smoothness constraints rather than information as to the actual parameter values; we illustrate such prior distributions below and full details in Section 4.2.

Bayes' Theorem can be used to combine the information in the prior distribution with that in the data. In particular, we can compute the *posterior distribution* via,

$$p(\Pi \mid X) = \frac{p(X \mid \Pi)p(\Pi)}{p(X)}, \tag{3.3}$$

where $p(\Pi \mid X)$ represents the posterior distribution, $p(X \mid \Pi)$ is equal to the likelihood function, and $p(X)$ can be viewed as a normalizing constant since it does not involve Π. From a Bayesian perspective $p(\Pi \mid X)$ is a complete summary of the information known about the Π. Since this is a probability function, its mean and its mode are both natural summaries and candidates for the reconstructed image. The mode is called the *maximum a posteriori* (MAP) estimate and denoted $\widehat{\Pi}_{\text{MAP}}$; the posterior mean is denoted $E(\Pi \mid X)$. The variability of this distribution represents uncertainty in the reconstructed image that can be quantified through error bars. A common computational strategy is to simulate a Monte Carlo sample from $p(\Pi \mid X)$ that can be used for Monte Carlo integration to compute means and error bars. Returning again to the example illustrated in Figure 1, a standard prior distribution used in conjunction with a multinomial likelihood is the Dirichlet

prior distribution,

$$p(\Pi) \propto \prod_{i=1}^{16} \pi_i^{\alpha_i - 1}, \tag{3.4}$$

where $A = \{\alpha_i, i = 1, \ldots, 16\}$ is a set of tuning parameters or *hyper parameters*. To motivate the use of the Dirichlet prior distribution and the choice of tuning parameters, consider the resulting posterior distribution,

$$p(\Pi \mid X) \propto p(X \mid \Pi) p(\Pi) \propto \prod_{i=1}^{16} \pi_i^{X_i + \alpha_i - 1}. \tag{3.5}$$

The MAP estimate of Π, the value of Π that maximizes $p(\Pi|X)$, is

$$\widehat{\Pi}_{\text{MAP}} = \frac{(X + A - \mathbf{1})}{\sum_{i=1}^{16}(X_i + \alpha_i - 1)}, \tag{3.6}$$

where $\mathbf{1}$ is a vector of ones. Notice that if each $\alpha_i = 1$, the Dirichlet prior distribution is flat and $\widehat{\Pi}_{\text{MLE}} = \widehat{\Pi}_{\text{MAP}}$. On the other hand, if α_i is relatively large, the MAP estimate of π_i will be larger than the MLE. The tuning parameters influence the computation of the MAP estimate in the same way that the counts influence the computation of the MLE; for this reason A may be referred to as the *prior counts*, see the first grid in Figure 1. The tuning parameters can be used to smooth the estimate of Π. In particular, if all of the tuning parameters are equal, $A = \alpha\mathbf{1}$ for some constant $\alpha > 1$, $\widehat{\Pi}_{\text{MAP}}$ will be smoother than $\widehat{\Pi}_{\text{MLE}}$. To see this, consider the case when α is much larger than the total number of counts, $\sum_{i=1}^{16} X_i$. If X is negligible relative to A, $\widehat{\Pi}_{MAP} \approx A/\sum_{i=1}^{16} \alpha_i = \mathbf{1}/16$, a perfectly smooth image. In real image reconstruction problems, we use moderate values of α to obtain a reasonable level of smoothing. As discussed in Section 4.3, we can even fit the tuning parameters to the data, thus eliminating the need for this used intervention.

We now turn to the details of our proposed statistical model for image reconstruction.

4 Modeling the Image

The statistical model used by the *EMC2* procedure is comprised of a likelihood function and a prior distribution, each of which is discussed in turn in the following two sections. The prior distribution is itself parameterized by a number of tuning parameters. In Section 4.3 we describe a hyper-prior distribution for these tuning parameters that allows us to fit the tuning parameters to the observed data; the computational methods are described in Section 5.

4.1 The Likelihood

To model the relative intensity of photon emission across the image, we begin by overlaying a grid of pixels on the source image. The photon counts originating from each of these pixels cannot actually be observed because of instrumental effects such as the point spread function and other data degradation effects such as background contamination. In the statistics literature, such unobservable quantities are called *missing data*. We emphasize that the term "missing data" does not imply that we expected to observe these data or that they were somehow lost, but rather the missing data are an elegant statistical construct that helps clarify model assumption and simplify statistical computation. Once we recognize that "true values" of quantities recorded with measurement error can be treated as "missing data" a large set of statistical tools designed to handle missing data can be employed. (van Dyk *et al.*, 2001)

We denote the missing pixel counts as X_i where i indexes the set of pixels, $\mathcal{I}$; we emphasize that X_i are not contaminated by background or blurred by the PSF and are therefore unobservable missing data. We model the ideal data, $X =$

$\{X_i, i \in \mathcal{I}\}$, as independent Poisson[1] variables,

$$X_i \overset{\mathrm{d}}{\sim} \text{Poisson}(\Lambda_i), \tag{4.1}$$

where Λ_i is the expected count in pixel i and $\Lambda = \{\Lambda_i, i \in \mathcal{I}\}$ describes the relative intensity of photon emission across the image; Λ is the quantity of primary interest. In Section 3 we parameterized the image in terms of the expected proportion of counts in each pixel rather than the expected total count per pixel; the two parameterizations are equivalent from a substantive point of view, $\Pi = \Lambda / \sum_{i \in \mathcal{I}} \Lambda_i$.

Unfortunately, the true counts are contaminated by background and blurred by the PSF. In particular, a photon originating in a region of the source corresponding to pixel i has probability P_{ij} of being counted in pixel j. We use the Multinomial[2] $(X_i, P_{i\cdot})$ distribution to model the observed location of an event, given its true location. Here, $\sum_j P_{ij} \leq 1$, with equality only if photons originating in this region are sure to be recorded somewhere on the detector. Background contamination occurs when a subset of the recorded counts are not due to photons arriving from the source of interest. Statistically, we may write

$$\mathbf{Y} \mid \mathbf{X} \overset{\mathrm{d}}{\sim} \sum_i \text{Multinomial}(X_i, P_i) + \Lambda^{\mathrm{B}} \tag{4.2}$$

where $Y = \{Y_j, j \in \mathcal{I}\}$ is the set of observed pixel counts, $P_i = \{P_{ij}, j \in \mathcal{I}\}$ is the vector of probabilities, and Λ^{B} is the set of pixel-wise background intensities. By including background contamination in the model in this way, we avoid the

[1]Recall, a random variable X is said to follow a Poisson distribution with parameter or intensity Λ if $\Pr(X = x) = e^{-\Lambda}\Lambda^x/x!$. In this case $\mathrm{E(X)} = \Lambda$ and we often write $X \overset{\mathrm{d}}{\sim} \text{Poisson}(\Lambda)$ (read as X is distributed as Poisson with intensity Λ). This representation conditions on the intensity parameter, Λ, which in turn may vary.

[2]The multinomial distribution is a random vector allocating a fixed number of counts independently to a series of bins, each with assigned probabilities. Thus a Multinomial$(n, \mathbf{p})$ will be a vector of the same length as $\mathbf{p}$ of nonnegative integers whose sum is n.

embarrassment of simply subtracting the background and ending up with possible negative intensities.

We may combine Equation 4.2 with Equation 4.1 via
$p(\mathbf{Y} \mid \mathbf{\Lambda}, \mathbf{\Lambda}^{\mathrm{B}})i = \sum_{\mathbf{X}} p(\mathbf{Y} \mid \mathbf{X}, \mathbf{\Lambda}, \mathbf{\Lambda}^{\mathrm{B}}) p(\mathbf{X} \mid \mathbf{\Lambda})$ to get

$$Y_j | \mathbf{\Lambda}, \mathbf{\Lambda}^{\mathrm{B}} \overset{\mathrm{d}}{\sim} \text{Poisson}\left(\sum_{i \in \mathcal{I}} P_{ij} \lambda_i + \Lambda_j^{\mathrm{B}} \right), \tag{4.3}$$

where Λ_j^{B} is the component of $\mathbf{\Lambda}^{\mathrm{B}}$ corresponding to pixel j. The model given in Equations 4.2 and 4.1 is exactly the same as that given in Equation 4.3, but with an intermediate stage added, involving the ideal data, $\mathbf{X}$. This additional layer in the model facilitates fitting $\mathbf{\Lambda}$, and will be discussed further below.

Equation 4.3 represents the distribution of the observed data given the model parameters, and, thus, is the building block of the likelihood function. In particular, since the observed counts $\mathbf{Y}$ are independent, we can write the likelihood function

$$L(\mathbf{\Lambda} \mid \mathbf{Y}) \propto \prod_j p(Y_j \mid \Lambda), \tag{4.4}$$

where $p(Y_j \mid \Lambda)$, is the Poisson distribution given in Equation 4.3. In Equation 4.4 we suppress Λ^{B} because we assume the background intensities are known.

In some cases, we may wish to model Λ in terms of an extended source and several point sources. In order to include K point sources, for example, we represent $\mathbf{\Lambda}$ as a sum,

$$\Lambda_i = {\Lambda_i}^{\mathrm{ES}} + \sum_{k=1}^{K} \Lambda_k^{\mathrm{PS}} P_{ik}^{\mathrm{PS}}, \tag{4.5}$$

where ${\Lambda_i}^{\mathrm{ES}}$ is the component of $\mathbf{\Lambda}^{\mathrm{ES}}$ corresponding to pixel i, Λ_k^{PS} is the total expected counts corresponding to point source k, and P_{ik}^{PS} is the proportion of these counts in pixel i. If point source k is represented by a delta function, P_{ik}^{PS} is one if the point source in pixel i and zero otherwise. Bivariate Gaussian or Lorentzian distributions may also be used to represent point sources, in which case P_{ik}^{PS} corresponds to a Gaussian or a Lorentzian probability. We generally assume $\{P_{ik}^{\mathrm{PS}}, i \in \mathcal{I}, \| = \infty, \ldots, \mathcal{K}\}$ are known and $\Lambda_1, \ldots, \Lambda_K$ are fit.

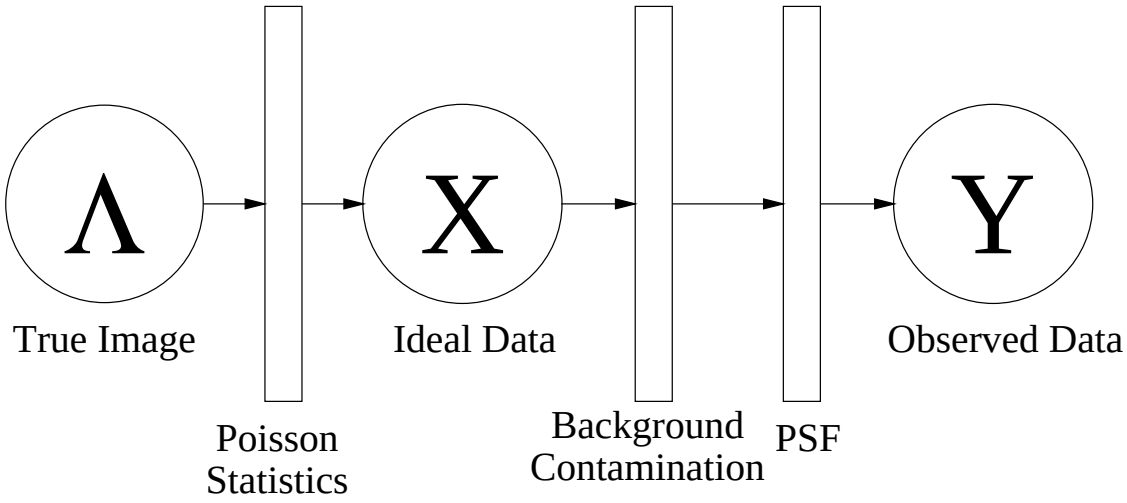


Figure 2: The data generation mechanism. The true image generates the ideal Poisson pixel counts, which are further degraded by background contamination and the PSF.

The Richardson-Lucy deconvolution technique, (Richardson, 1972) and (Lucy, 1974), is designed to compute the MLE under Equation 4.4 in the absence of added point sources, i.e., $K = 0$ and provides a fit for the model, with no error information.

The advantage of the Richardson-Lucy algorithm is that it is relatively fast and easy to implement, and performs well for sources which are concentrated in a few points. However, for observations with few counts and extended low-intensity source, the algorithm does not smooth the image at all, and typically concentrates intensity into points rather than allowing smooth low-intensity regions in the image. In practice, users must stop the algorithm before convergence to avoid an overly grainy reconstruction. To avoid such interaction between the user and the algorithm, we must incorporate a smoothing component into the statistical model. We accomplish this through a prior distribution that represent the expectation that extended sources are often smooth. As we shall see, our prior specification is designed to allow for both sharp edges and smooth extended regions.

4.2 Prior Distributions for Extended Sources

In order to regularize the source intensities by putting a prior distribution on them, the *EMC2* procedure uses a different representation of the image of intensities. The multiscale representation, rather than storing each pixel intensity, stores the total of all pixel intensities, and a series of four-way split proportions which, when multiplied together in the correct sequence, give back the pixel intensities. For example, for a 2 by 2 image, the multiscale representation consists of the sum of the four pixels and three of the four split proportions, the fourth being of course redundant. For a 4 by 4 image we add four more sets of split proportions to split each of the four pixels of the 2 by 2 image into four.

This image representation is advantageous in that we can use it to impose structure on the image by assigning prior distributions to each quantity, or to sets of split proportions. Then we may achieve regularization or smoothing at multiple scales in the image. In the *EMC2* procedure a Gamma prior distribution is used for the total intensity. The Gamma distribution is restricted to positive values, as is desirable for Poisson intensities, and is both flexible in terms of shape and mathematically tractable, and thus is an ideal choice for modeling the total intensity. We use Dirichlet distributions (vector generalizations of the Beta distribution) for the 4-way split proportions. Dirichlet distributions are nonnegative vector quantities, restricted to sum to one, and also are mathematically tractable when used here to model these quantities: they combine with the multinomial model for the counts per each split to yield Dirichlet posterior distributions for the split. The choices of models, gamma for the total intensity, and Dirichlet for the splits, are also minimax with respect to squared error loss functions, so that the worst possible error resulting from fitting the statistical model will be less than the worst case of any other statistical model, without additional distributional information. (Morris, 1983) Thus in the absence of any other information about their probability distributions, these models are the best choice.

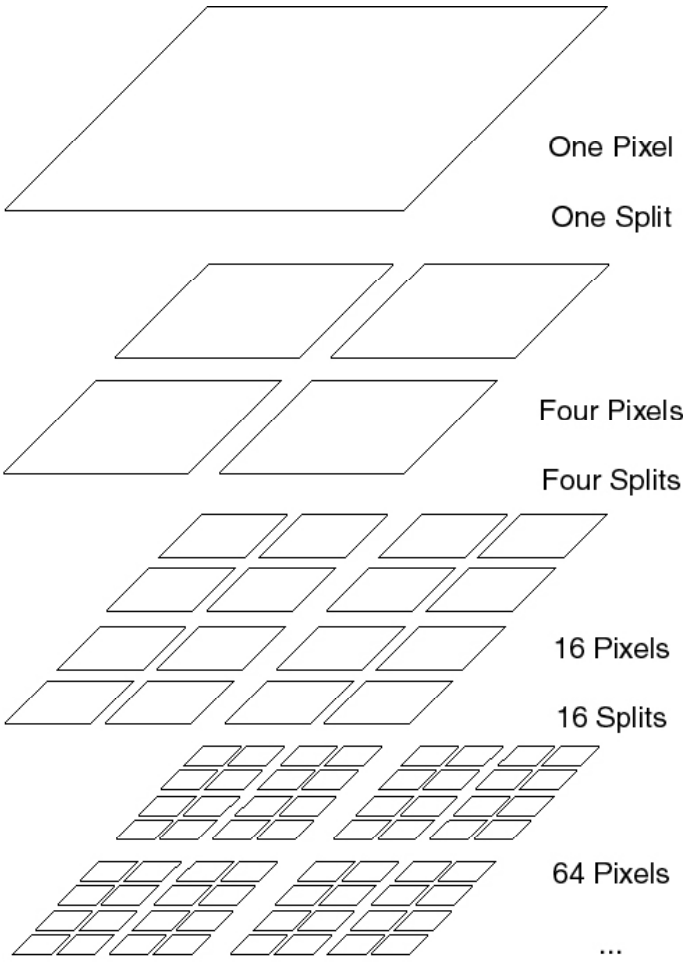


Figure 3: The multiscale image representation. The top layer is the total intensity, and each subsequent layer splits the blocks of its parent layer into four parts.

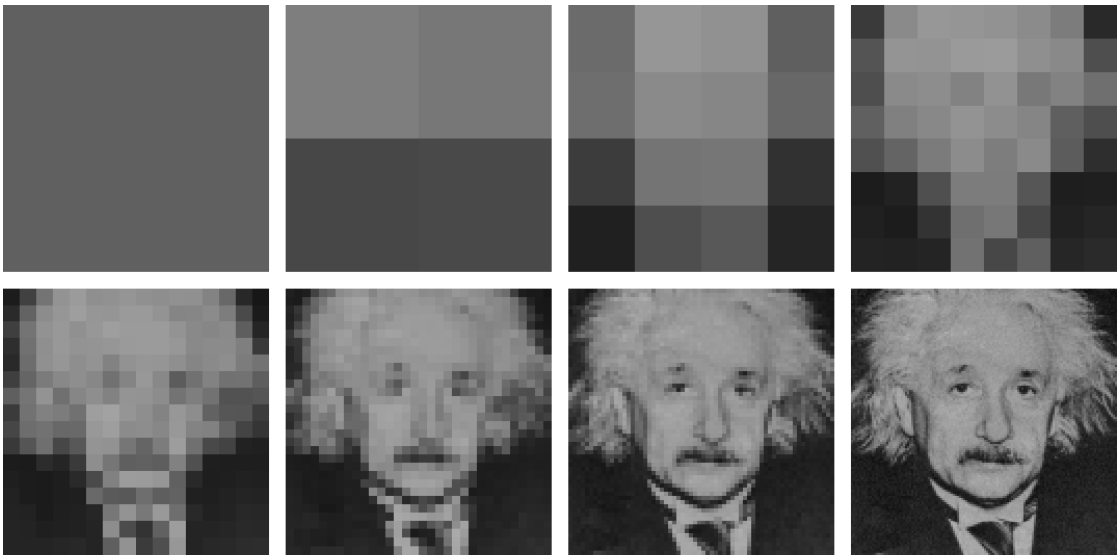


Figure 4: The eight levels of a multiscale representation of a 128 by 128 pixel image. The grayscale colormaps have been adjusted for each level. Each layer is produced from its parent layer by splitting each pixel block into four smaller ones.

The Dirichlet parameters function as smoothing or regularization parameters. They have the interpretation of adding their value of counts to each quadrant of a four-way split. Naturally, it would be possible to assign different values to every single Dirichlet parameter, or to group various parameters together and assign different values to them, but such model modifications add considerable complication to the implementation of the model fitting algorithm, and add to the already high level of complication in choosing the proper values for the parameters. This issue is discussed in (Nowak and Kolaczyk, 2000) in detail.

We model the observed event locations $\mathbf{Y} = \{Y_i\}$ as a convolution of multinomials about the ideal data $\mathbf{X} = \{X_i\}$, according to the PSF $\mathbf{P} = \{P_{ij}\}$:

$$\mathbf{Y}|\mathbf{X} \overset{\mathrm{d}}{\sim} \sum_j \text{Multinomial}(X_j, P_{j\cdot}). \tag{4.6}$$

We model the ideal data as Poisson with intensity rates
$\mathbf{\Lambda} = \{\lambda_i\} = \{G \prod_k D_{k,l_k(i),m_k(i)}\}$ in the multiscale representation:

$$X_i|\lambda_i \overset{\mathrm{d}}{\sim} \text{Poisson}(\lambda_i), i = 1, \ldots, N., \text{ i. e.} \tag{4.7}$$

$$X_i|G, \mathbf{D} \overset{\mathrm{d}}{\sim} \text{Poisson}\left(G \prod_k D_{k,l_k(i),m_k(i)}\right). \tag{4.8}$$

Here $l_k(i)$ and $m_k(i)$ index the splits and quadrants at level k containing pixel i.

Finally, to regularize the image we model the total intensity rate G and the split proportions $D_{klm}, m = 1, \ldots 4$ as Gamma and Dirichlet:

$$G \overset{\mathrm{d}}{\sim} \text{Gamma}(\gamma_0, \gamma_1); \tag{4.9}$$

$$D_{kl\cdot} \overset{\mathrm{d}}{\sim} \text{Dirichlet}(\alpha_k, \alpha_k, \alpha_k, \alpha_k); \tag{4.10}$$

This model is a two-dimensional generalization of the model used in the deconvolution technique proposed by (Nowak and Kolaczyk, 2000). Their model uses a similar regularization scheme, with a gamma prior for the overall intensity, and beta priors for each (one-dimensional, two-way) split. In their model, each regularization parameter, the two equal beta parameters, must be fixed by the user

before running the deconvolution. This procedure can produce excellent results if the beta parameters are specified well at each level, but results can vary considerably for different parameter values, even changing the scientific conclusion. Since error information is not available, the analyst cannot know with how much confidence inferences are being made. Still, this procedure is an improvement over the analysis with likelihood alone, and with a few enhancements, can overcome the aforementioned difficulties.

4.3 Fitting the Tuning Parameters

Since the value of the Gamma prior on the overall intensity governs only the overall flux of the image and not the ratios of any pixel or sets of pixels to any other, the settings on the parameters of that gamma distribution will not have any effect on the appearance of the restored image, only on the overall flux. Thus we choose to fix these parameters γ_0 and γ_1, to relatively uninformative values such that the expected total flux will be equal to the observed total flux.

In order to compensate for our uncertainty about the proper values for the tuning (Dirichlet) parameters, we add a common prior distribution for these parameters. Since allowable values for these parameters are restricted to positive values, we choose a gamma distribution for the common prior. The choice of gamma parameters has a paradoxical effect on the smoothing parameters: in order to specify a less informative prior on the image restoration, i. e., to let the data influence the outcome more with less regularization, we must put a more informative prior on the smoothing parameters, restricting them to smaller numerical values. We still allow considerable variation in the smoothing parameters, and computation of a posterior mean with the additional hyperprior distribution has the effect of averaging over the uncertainty in these parameters.

Thus, in the complete model, all parts of the model in Equations 4.6 stand, conditional on the values of $\{\alpha_k\}$, the smoothing parameters. We add a common

Gamma prior distribution to the set of smoothing parameters, to treat them as random quantities and enable fitting them:

$$\alpha_k \overset{\mathrm{d}}{\sim} \mathrm{Gamma}(\delta_0, \delta_1). \tag{4.11}$$

5 The *EMC2* Technique: Fitting the Model

We choose to draw a sample from the posterior distribution of the true image, rather than maximize the density. There are several reasons for this: the first and perhaps most important being that the unrestricted density is infinite for many values of the parameters, none of which we believe to be plausible values for the true image. Because of the boundedness of the four-way splits, when the smoothing parameter for a given split is less than unity and there are zero counts in some quadrants, the Dirichlet distribution will be U-shaped in certain cross sections, and the split for which zero intensity is given to the quadrant in question will have infinite posterior density. We could circumvent this problem by restricting the smoothing parameters to values greater than one, but this restriction might unnecessarily smooth other splits in mode densely populated areas of the image. In this case the *maximum a posteriori* (MAP) estimate of the true image is not a useful summary of the posterior, and is probably not unique. The posterior mean is a much better way to summarize the posterior, and will produce a restored image that is more " in the middle" of the range of values. Another reason for preferring a sampling scheme for the posterior rather than a maximization scheme, is that having a sample allows for all kinds of inference, particularly including confidence information about the inferences made. In other words, error bars can be put around the fitted intensities by using empirical quantiles from the sample.

Both the Richardson-Lucy and the Nowak-Kolaczyk algorithms use a maximization of the posterior distributions from their models as their fitted values. The Richardson-Lucy first calculates the expected values in $\mathbf{X}$, and then updates

the intensities with these same values, since the maximum likelihood estimates for Poisson intensities are simply the data values themselves. The expectations are taken from the conditional posterior of $\mathbf{X}$, which is given as follows:

$$X_i|\mathbf{Y},\mathbf{\Lambda} \overset{\mathrm{d}}{\sim} \sum_j \mathrm{Binomial}(Y_j, Q_{ji}), \tag{5.1}$$

where Q_{ji} is the probability that an event observed in pixel i originated from pixel j, i. e. $Q_{ji} = \lambda_j P_{ji}/(\sum_k \lambda_k P_{ki})$.

The Nowak-Kolaczyk algorithm then uses these $\mathbf{X}$ values to maximize the conditional posterior for $\mathbf{\Lambda}$, which is given as follows:

$$G|\mathbf{X} \overset{\mathrm{d}}{\sim} \mathrm{Gamma}(\gamma_0 + X_0, \gamma_1 + 1); \tag{5.2}$$

$$D_{kl\cdot}|\mathbf{X} \overset{\mathrm{d}}{\sim} \mathrm{Dirichlet}(\alpha_k + X_{kl1}, \alpha_k + X_{kl2}, \alpha_k + X_{kl3}, \alpha_k + X_{kl4}); \tag{5.3}$$

Thus the Nowak-Kolaczyk procedure adds another step to each Richardson-Lucy iteration, leaving the existing step intact.

The *EMC2* procedure again builds on the above steps. Rather than calculating expectations and maximizing, the *EMC2* procedure draws a sample from the densities in Equations 5.1 - 5.3. Then another step is added to draw the smoothing parameters. This step samples from the conditional posterior of the smoothing parameters, which is given as follows:

$$p(\alpha_j|\mathbf{X},\lambda) \propto \left[\prod_{\mathbf{k=1}}^{\mathbf{4^{j-1}}} \frac{\mathbf{\Gamma}(\mathbf{4}\alpha)}{\mathbf{\Gamma}(\alpha)} \prod_{\mathbf{l=1}}^{\mathbf{4}} \left(\frac{\lambda_{\mathbf{jkl}}}{\sum_{\mathbf{m=1}}^{\mathbf{4}} \lambda_{\mathbf{jkm}}}\right)^{\alpha_\mathbf{j}}\right] \mathbf{p}(\alpha). \tag{5.4}$$

We have found the conditional posterior density in Equation 5.4 to be too dependent on the Dirichlet values themselves, which were sampled using the current value of the smoothing parameters, so that the sampled value has extremely high correlation with the previous value, and the Markov chain will take a long time to cover the range of possible values. To remedy this situation, we instead draw from the conditional posterior distribution of alpha, first integrating out the Dirichlet

values themselves. This distribution is given by

$$p(\alpha_j|\mathbf{X}) \propto \prod_k \frac{\Gamma(4\alpha)\prod_m \Gamma(\alpha_j + X_{jkm})}{\Gamma(4\alpha + X_{jk})\Gamma(\alpha)^4} p(\alpha) \tag{5.5}$$

$$= \prod_k \frac{\prod_m \prod_{i=0}^{X_{jkm}-1}(\alpha_j + i)}{\prod_{i=0}^{X_{jk}-1}(4\alpha + i)}, \tag{5.6}$$

$$\text{where } X_{jk} = \sum_{m=1}^{4} X_{jkm} \tag{5.7}$$

We can approximate the curve with a t distribution centered at the mode and with the same curvature at the mode, and draw several iterations from a Metropolis-Hastings chain to obtain a draw from Equation 5.6. Figure 5 shows a comparison of a time series of draws for the two methods corresponding to Equations 5.4 and 5.6, and it is clear that the performance of the latter is far superior in mixing rate.

5.1 Assessing convergence

We use Gelman and Rubin's $\hat{R}$ (Gelman and Rubin, 1992) to assess convergence of the chains. Since each step of each chain is relatively expensive, we compute the statistic from just two chains. We keep track of the value of $\hat{R}$ for all of the smoothing parameters, as well as some or all of the pixel intensities. We declare convergence when the maximum $\hat{R}$ reaches a suitably low value, near 1.0, in particular 1.1, as recommended in (Gelman and Rubin, 1992). After we declare convergence in the chains, it remains to run the chains until the Monte Carlo error has been reduced to a suitably low level. We find that an average coefficient of variation of 0.05 provides a reasonable compromise between reducing the Monte Carlo error and the computing time for the procedure, but if the user desires more precision or faster computing time, the tolerance can be adjusted accordingly.

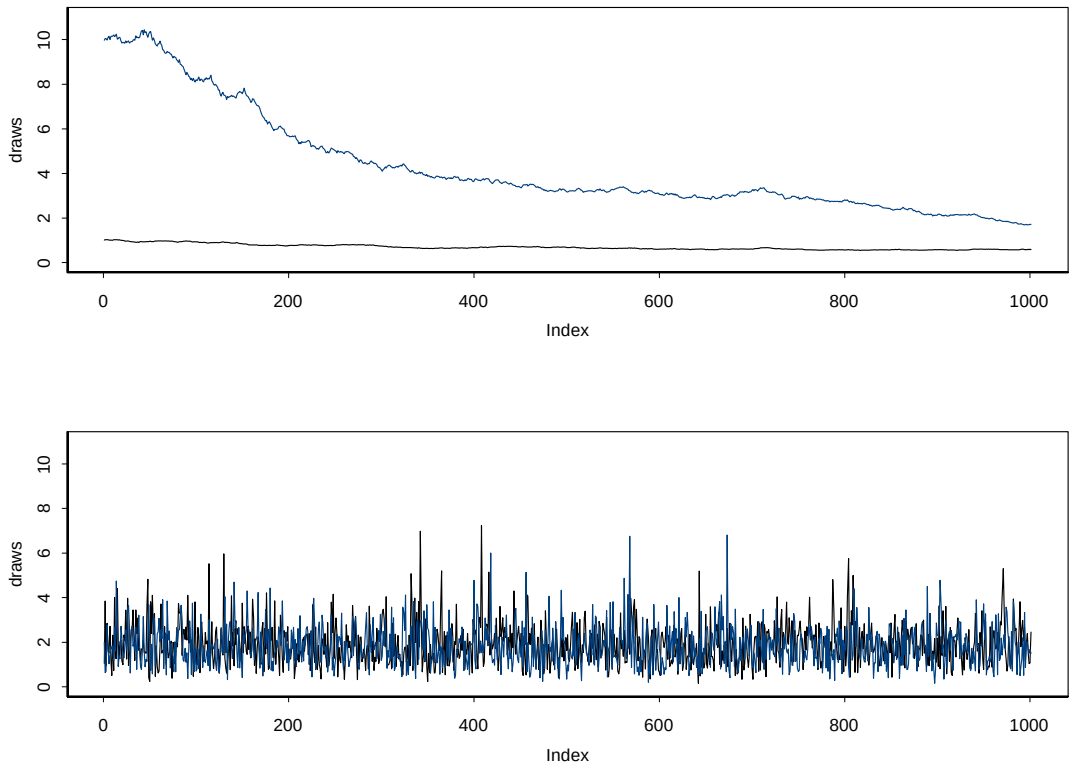


Figure 5: A comparison of the two methods for sampling the smoothing parameters for the highest resolution of a 256 by 256 image. The upper series draws from the density in Equation 5.4 and the lower from that of Equation 5.6. Both graphs show draws from two chains starting at values of 1 and 10, and clearly the lower set of chains mix immediately, compared to the upper set, which have not clearly attained convergence after 1000 draws. These draws were made fitting data from a Chandra observation of NGC4244.

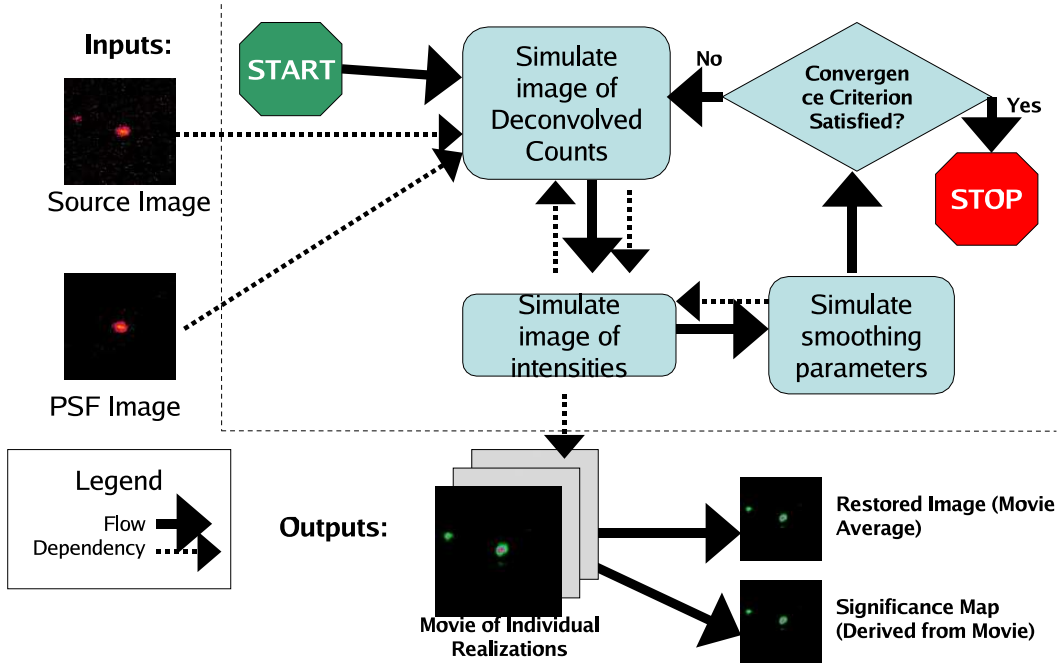


Figure 6: EMC2 Flow Diagram

5.2 Cycle Spinning

Fitting the model as above, artifacts of the multiscale representation are clearly visible in the resulting fit. The standard remedy for this problem is to average together multiple analyses of the image, with the origin of the multiscale decomposition of the image in a different place at each iteration. This can be done systematically or randomly. This technique is known as cycle spinning, and was described by Coifman and Donoho. (Coifman and Donoho, 1995)

5.3 The Algorithm

1. We start with the image of observed counts and locations, and the PSF.

2. We recreate the missing image of counts restored to their true locations, according to the following probabilities: if an event is observed in pixel j, the probability that it originated from pixel i is $\lambda_i P_{ij}/(\lambda * PSF)_j$, where $\lambda * PSF$ is the current version of the true image convolved with the PSF.

3. We reconstruct the multiscale version of the results of 2, and draw the intensities from the distributions in Equations 5.1 - 5.3.

4. We draw the smoothing parameters from the distributions in Equation 5.6.

5. Unless convergence has been assessed and sufficient draws have been made to reduce Monte Carlo error to desired precision, we go back to step 2. Otherwise we stop and output the iterations from step 3.

6 EMC2 Output and Inference

The primary output of *EMC2* is the sequence of images of intensities. The sequence of images from *EMC2* after convergence is a representative sample from the posterior distribution of the true image, and thus is the basis for all inference from the statistical model. We can form a principal estimate of the true image by averaging all of the images in the sequence. As a first pass we can easily calculate the standard deviation of the draws and rescale the averaged image, dividing by the standard deviations of pixels. This image will form a significance map, in which the pixels are test statistics for the hypotheses that pixel intensities are zero; for greater pixel intensities this makes sense, but due to the Poisson nature of the observed counts, we feel that there is much more in the restored image that is "real" than will show in such a significance map, thresholded to a reasonable cutoff, such as 3 (standard deviations). A more sophisticated analysis could be done by estimating background for each pixel and evaluating the proportion of output images whose pixel intensities are greater than the estimated background value. However, this opens up questions as to how to estimate background, and

how to quantify uncertainty in background estimation procedures, and thus will require more research to implement.

The output images themselves can be viewed in a movie, which by itself provides much information to the researcher about the restoration. Some pixels in the movie are persistent, whereas others flicker between very low and moderately low intensities. The persistent pixels would thus be more statistically significant, because they are more certain according to the model, whereas the flickering ones are less certain according to the statistical model. The movie can also be sorted pixel-by-pixel, thereby destroying correlation information from the data cube, but at the same time providing a useful summary of each pixel's marginal posterior density.

The three dimensional array of output images can be the basis for many other types of inference. For any image feature that can be assigned a quantitative summary we can calculate a posterior predictive P-value (PPP-value) to assess statistical significance. Regions of the image could be defined and compared with other regions to extract information regarding observed features in the restoration.

Since the model generally conserves total flux, it should be noted that pairs of pixels in the image generally tend to correlate negatively, especially when nearby. This can be seen by the splitting mechanism: if intensity is increased in one pixel, it must be decreased in an adjacent pixel to preserve the sum. This correlation structure is important in making correct inferences, but is also very difficult to summarize in a concise and informative way.

7 Examples

7.1 A Simulated Example

We constructed a simulation, based on a real PSF, of two point sources. The simulations were done at lower counts than was the original PSF, so we have no

hope of a perfect restoration. Figure 7 shows the simulated observation and the PSF, with two *EMC2* restorations of the source. The restorations differ in the settings of the prior. We have experimented with several different setting of $p(\alpha)$, and compared the results. The most obvious setting for this prior distribution is a gamma distribution. However, with the shape parameter set to 1.0, i. e., the exponential distribution produces numerical instability which causes the resulting Markov chain to fail. Other settings of the shape parameter can be made to work, but we prefer other priors entirely. The half normal, or Chi distribution with one degree of freedom, produces good results, and the prior with density proportional to $\exp(-\delta\alpha^3)$ produces excellent results. The restorations shown in Figure 7 used the latter prior with $\delta = 1000/3$ and $\delta = 10/3$. We can see clearly that the $\delta = 1000/3$ restorations are preferable, becuse the $\delta = 10/3$ produces too much smoothing in the restoration. Otherwise, the results are good, and the Poisson noise in the data make it impossible for the intensity in the restored image to shrink down to two pixels. We have shown the restored images on the log scale, to illustrate these "leftovers", but the pixels adjacent to the points really are much less in intensity than the two points themselves. On the linear scale the images do appear as two lit pixels with a black background.

7.2 Example 1: NGC 4244

Our first example is a Chandra/ACIS observation of a faint source, NGC 4244. The observation consists of approximately two thousand counts, and is clearly seen to be elongated. However, the observation was made off-axis on the telescope detector, and the PSF for this observation, simulated based on detailed knowledge of the instrument, can also clearly be seen to be elongated as well. Thus it is unclear from the raw data alone if the source is binary or not. As can be seen in Figure 9, the Richardson-Lucy deconvolution shows a single source, but WAVDETECT shows a binary source. The WAVDETECT program (Freeman *et al.*, 2002) is a

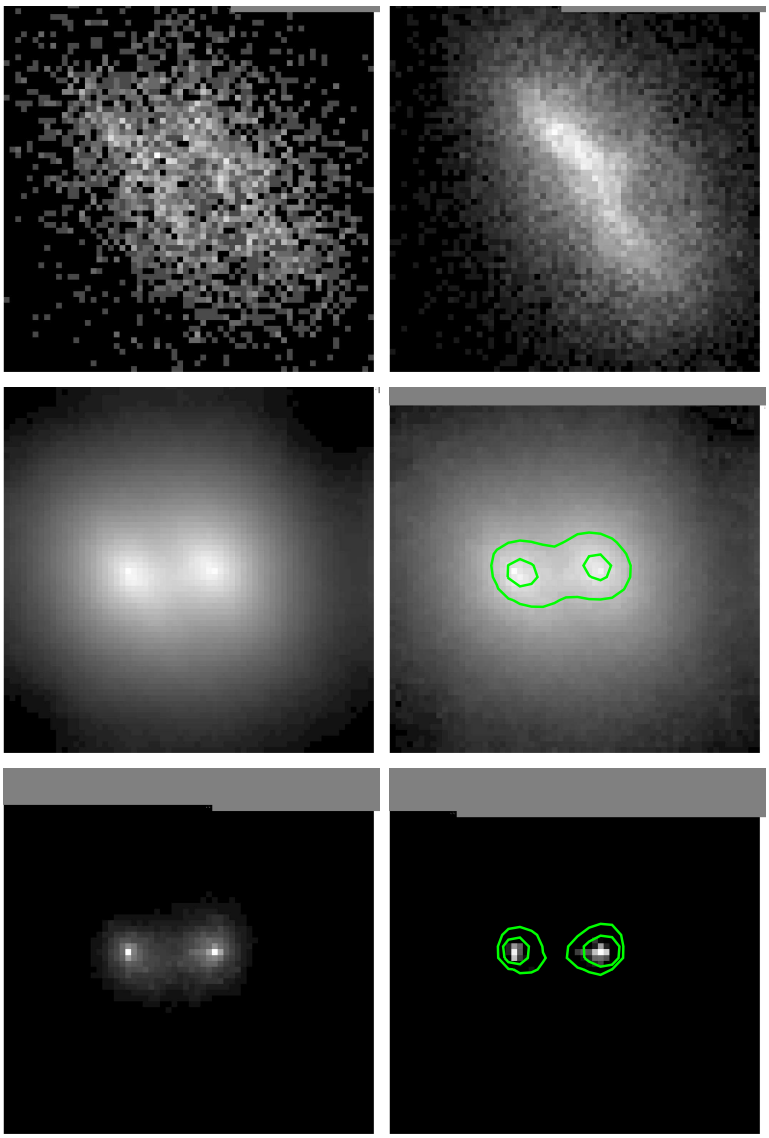


Figure 7: A simulated example. The top two panes are the data and PSF, respectively, shown on the square root scale to best illustrate the variability of the counts in these images. The next two rows of panes are the posterior mean and significance map (posterior mean divided by posterior standard deviation) of two *EMC2* restorations. Both of the restorations used a prior for the smoothing parameters of $p(\alpha) \propto \exp(-\delta\alpha^3)$, the first with $\delta = 1000/3$, and the second with $\delta = 10/3$. The contour lines shown on the significance maps have levels of 3.0 and 10.0, so we are quite confident that the nature of the source is binary.

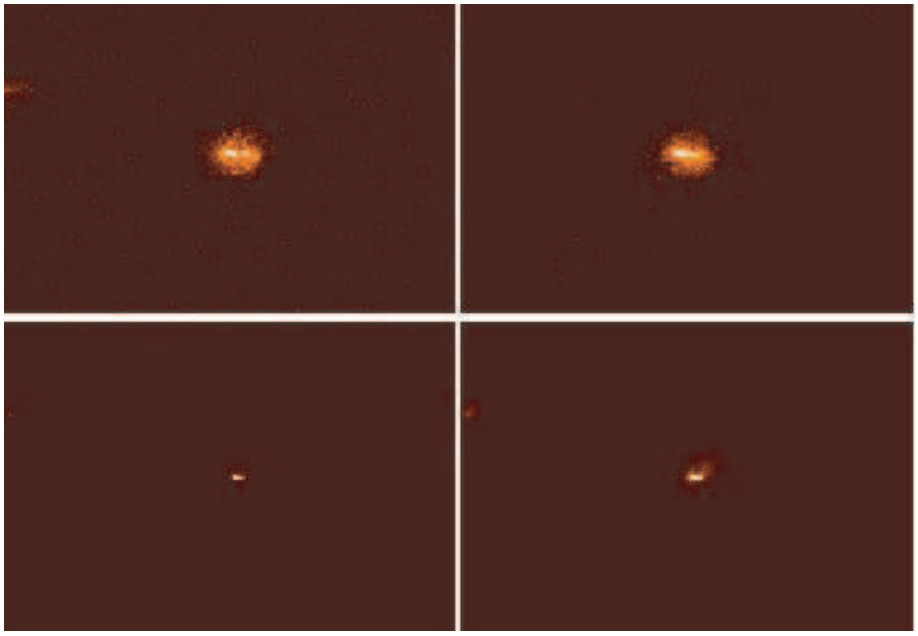


Figure 8: NGC4244, the object in example 1. The upper left pane is the original dataset, and the upper right pane is the PSF, simulated with the ChaRT/MARX simulator. The lower left pane is the posterior mean reconstruction, and the lower right pane is the significance map, thresholded to three standard deviations.

source detection routine, which does not perform deconvolution, so we are more inclined to believe the Richardson-Lucy result, although there is still some doubt as to the certainty of the conclusion.

Figure 8 shows the EMC2 deconvolution, and we can clearly see a single source. The upper two panels in Figure 8 are the observed counts and the PSF, and the lower two panels are the posterior mean reconstruction and the errormap thresholded to three standard deviations. The conclusion of a single source is also borne out by the errormap image as well as the movie of images.

There is also a readout streak in the instrument which extends to the upper

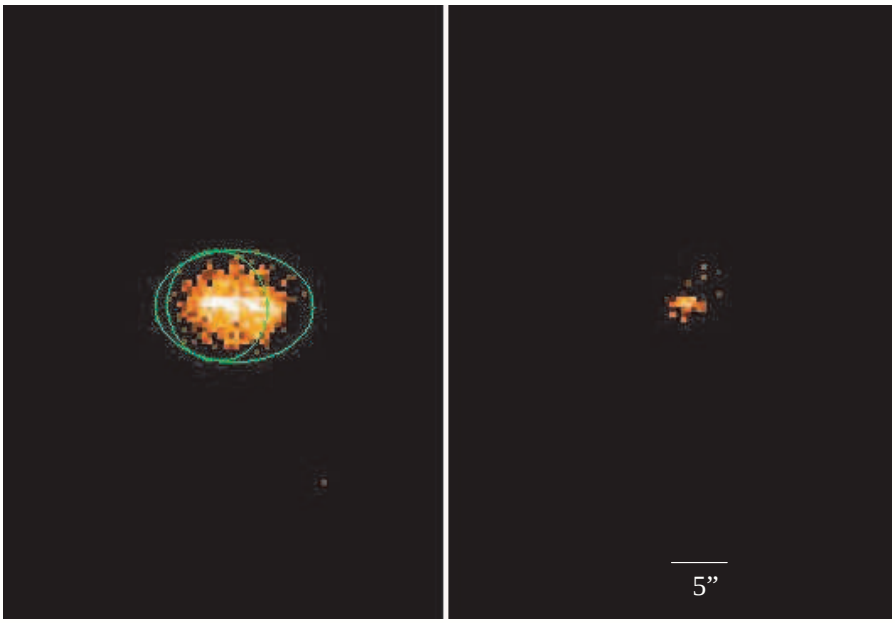


Figure 9: NGC4244, the object in example 1. The left-hand pane shows the WAVDETECT result, indicating a binary source, and the right-hand pane is the Richardson-Lucy deconvolution, clearly showing a single source.

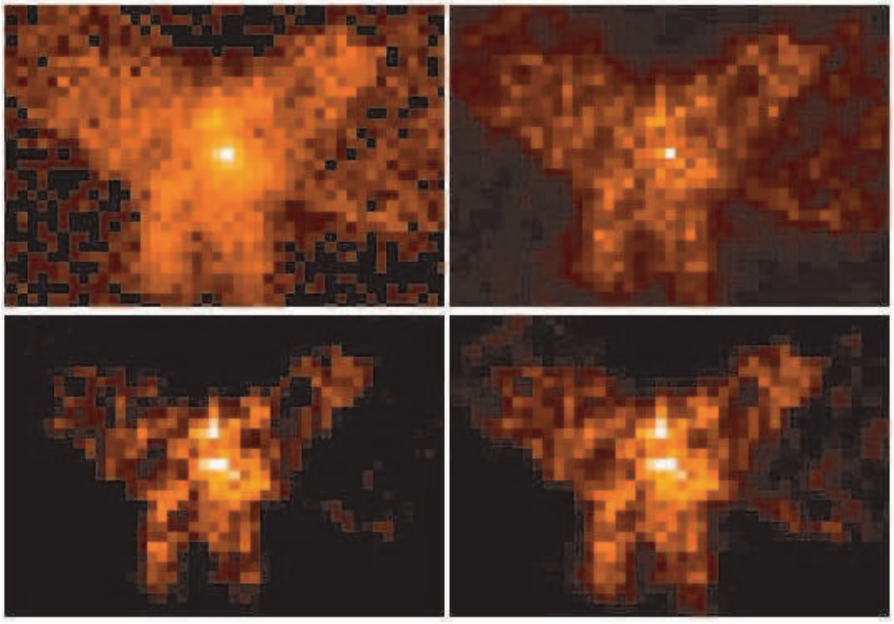
right of the image, and was not accounted for by the PSF simulation. This readout streak explains the faint trail extending away from the source to the upper right of both of the deconvolved images.

8 Example 2: NGC6240

Our second example is a Chandra X-Ray Observatory image of NGC 6240, a source now known to be a binary supermassive black hole. In Figure 8 we see in the upper left pane the observed data, and in the upper right pane the *EMC2* posterior mean reconstruction. Many features in the source are greatly sharpened and more visible in the reconstruction. The lower two panes are the errormap, thresholded at one and three standard deviations. Clearly the extended features, in particular the loops surrounding the central black holes, are real according to the errormap. We know the source is a binary black hole from Hubble Space Telescope (HST) images in the optical frequencies, and the new Chandra observations corroborate this finding. There is also much extended material in the source, forming loops which circle away from and back towards the source. These loops can clearly be seen in the HST images, and are somewhat less discernable in the lower-count X-ray data. Figure 11 shows an overlay of the *EMC2* deconvolution with the HST images. The X-ray observations provide valuable scientific information on the content of the various components of the source, and by correlating these parts of the image with other frequencies much can be learned about the object.

Figure 10 compares the *EMC2* restoration of the object with Richarson-Lucy deconvolutions, stopped at 20 and 100 iterations, and we can see the poor performance of the Richardson-Lucy deconvolutions for the extended material. The images tend to aggregate intensity into clumps, rather than smoothing them over extended regions.

Figure 11 shows the overlay of the *EMC2* posterior mean restoration with the HST optical image of the same source. The extended structure is the same in both



captionNGC6240, the object in example 2. The upper left pane shows the original dataset as recorded by the telescope. The upper right pane is the posterior mean restoration, clearly showing two sources in the middle, surrounded by extended loops of interstellar material. The lower two panes are the significance map, thresholded at three and one standard deviation, respectively.

images, which is additional evidence supporting the *EMC2* results.

9 Concluding Remarks

This work brings several new and important features to a deconvolution analysis, perhaps the most important being the quantification of the uncertainty of the outcome. The analysis itself is based on principled methods, and is performed entirely in one procedure, thus eliminating any error propagation that may occur by

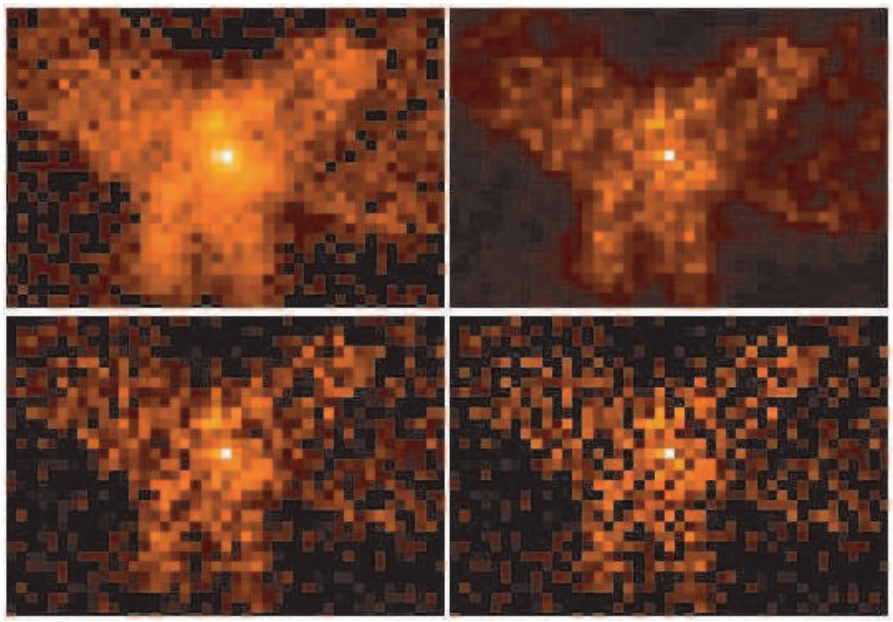


Figure 10: NGC6240, the object in example 2. The upper panes are the raw data and the posterior mean EMC2 restoration, respectively, as in figure 8. The lower two panes are the Richardson-Lucy deconvolutions, stopped at 20 and 100 iterations, respectively. The Richardson-Lucy restorations clearly undersmooth the image, aggregating the intensity at "clumps" in the image.

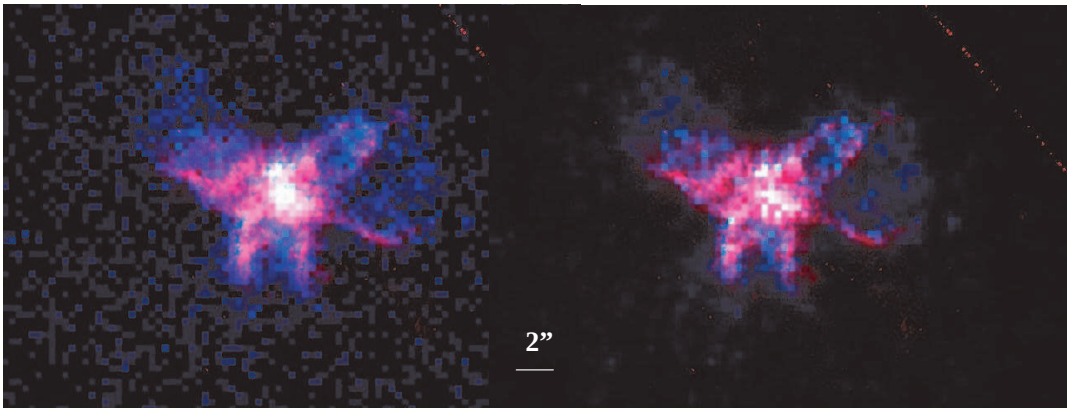


Figure 11: NGC6240, the object in example 2. The figure shown is an overlay of the HST optical frequency image with the EMC2 posterior mean restoration. Clearly there is much correlation in the extended structures, which provides additional confidence in the EMC2 results.

reprocessing data in multiple steps. The regularization technique works well and preserves the integrity of extended sources in the image. The results from *EMC2* for point sources match well those from the Richardson-Lucy algorithm, which is currently accepted as a standard for point-source deconvolutions, and the results for extended sources also match those from adaptive smoothing algorithms, which are generally used for extended material. The results also match the observations from other wavelengths, e. g. optical, which are not subject to the extreme scattering effects from the PSF. Many of the images used for publicity are composites of several procedures, one for point sources, one for extended material, and perhaps others; the *EMC2* procedure could conceivably replace all of those and be a consistent, scientifically sound procedure as well.

10 Future Work

In the current implementation of EMC2 the PSF is a single matrix, representing the scattering of a point source from anywhere on the entire image; however, the procedure described here requires only that we be able to calculate probabilities that an event with any true location is observed in any other location on the detector. This probability can come from one PSF, many PSFs in a mosaic image, or an unknown PSF that can be described probabilistically and simulated from. Thus the method is extendable to a broad range of deconvolution problems.

Other components of the image could also be specified and included in the analysis. This would amount to adding a step to the MCMC algorithm which splits the true counts according to the current intensities in the components for each pixel. Often astronomers have vague or strong prior information about an observation, which would be reasonable to include in an image analysis. We have worked with including point sources in the image at known locations, by including small Gaussian components at fixed locations. Such components could also be fit for location as well.

Part II

Generalized Linear Models with the NEF_6 Distribution

11 Introduction

The NEF_6 distribution is the only little-known Natural Exponential Family with Quadratic Variance Function (NEF-QVF). It is the natural exponential family generated by the Generalized Hyperbolic Secant (GHS) distribution, which in turn represents convolutions and divisions of the Hyperbolic Secant (HS) distribution, which has density proportional to the hyperbolic secant function. There are only six NEF-QVF. (Morris, 1982) The first five NEF-QVF distributions, $\mathrm{NEF}_1, \ldots \mathrm{NEF}_5$ are the Gaussian, Poisson, Gamma, Binomial, and Negative Binomial. The NEF_6 family has received little attention from the statistics community , perhaps because most computations related to the distribution are more complicated than the other five. However, the distribution enjoys all of the properties of natural exponential families with quadratic variance functions derived in (Morris, 1982) and (Morris, 1983), and as such deserves some attention from the statistics community for its modeling possibilities. Understanding this family of distributions completes our understanding of NEF-QVF theory, and departures from the other more familiar models may provide further understanding of intrinsic behavior of NEF-QVF models and what properties are specific to various members of the six NEF-QVF. The NEF_6 is the only NEF-QVF distribution with support on the entire real line, and possible nonzero skewness. Its skewness gives the distribution its potential to model datasets with extreme values, skewed error distributions, nonlinear conditional expectations, or other characteristics not well-matched by Gaussian models.

Special cases of the NEF_6 distribution have been known for some time. (Baten, 1934) introduced the Hyperbolic Secant (HS) distribution and its integer convolutions. The HS distribution is the unskewed case of the NEF_6 distribution where the parameter r is set to 1. (Harkness and Harkness, 1968) expressed the density of noninteger convolutions of the HS density as the product of two complex gamma functions, calling the result the Generalized Hyperbolic Secant (GHS) distribution. (Morris, 1982) generated the exponential family from the GHS distribution. The resulting distribution is the most general NEF_6 distribution, and was called the NEF-GHS distribution in that work. (Devroye, 1993) gave routines for obtaining random draws from Morris' NEF-GHS distribution.

Generalized linear models with the canonical links (McCullagh and Nelder, 1989) correspond to Natural Exponential Families; models with other link functions mapping the mean on to the linear combination of regressor variables correspond to exponential families, but not natural exponential families. Several results apply for Natural Exponential families that do not necessarily hold for general Exponential Families: the likelihood surfaces for the regression coefficients are unimodal, and the distributions of sufficient statistics, which are simply sums and cross-products of the observations, also come from the same NEF. Procedures which fit models with non-canonical link functions are not guaranteed to have stable convergence properties. Nevertheless, there is no particular *a priori* reason why we should prefer the canonical link, other than convenience, (McCullagh and Nelder, 1989) and NEF_6 models with the identity link might be worth investigating, because their properties are quite different than models with the arctangent link, which is the canonical link function for the NEF_6 Generalized Linear Model.

We believe that this density may provide a missing piece of the Generalized Linear Model/NEF framework, and may provide investigators with a means to analyze data with quadratic mean/variance relation, but whose variance does not entirely disappear for zero mean variables of the family. The NEF_6 Generalized Linear Model may also be useful for quasi-likelihood analysis, for quadratic vari-

ance functions with no real roots. It may be quite common for datasets to have standard deviations scaling approximately proportionally to the mean of the linear predictor, but which do not disappear entirely at zero.

12 Basic Properties of the NEF_6 Distribution

The natural parametrization of the NEF_6 density is

$$f(x|\theta, r) = \frac{2^{r-2}\cos^r(\theta)}{\pi r} e^{r\theta x} \mathrm{B}\left(\frac{r}{2} + i\frac{rx}{2}, \frac{r}{2} - i\frac{rx}{2}\right), \tag{12.1}$$

with respect to Lebesgue measure, in terms of the natural exponential parameter $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, and the "convolution parameter" $r \in (0, \infty)$. Here $\mathrm{B}(a, b)$ is the Beta function, which is equal to $\Gamma(a)\Gamma(b)/\Gamma(a+b)$. In the natural parametrization, the mean is $\tan(\theta)$, and it is often convenient to reparametrize the distribution in terms of $\mu = \tan(\theta)$. The product of the complex conjugate pair of gamma functions is real, and equal to the squared modulus of the gamma function evaluated on either argument, and reduces to a product of rational and hyperbolic functions for integer values of r. In the case where $r = 1$ the density reduces to

$$f(x|\theta, r = 1) = \cos(\theta)\frac{\exp(r\theta x)}{2\cosh(r\pi x/2)},$$

and hence the name "hyperbolic secant." We will follow the convention of (Morris, 1982) in parametrizing the density either by $X \sim \mathrm{NEF}_6(r, \mu)$ or $X \sim \mathrm{NEF}_6[\mu, (1+\mu^2)/r]$, the square brackets in the second parametrization indicating that its two entries are the mean and variance of the random variable. Either of these two parametrizations completely identifies the density in question.

Distributions with integer values of r are averages of r $\mathrm{NEF}_6(1, \mu)$ random variables; in other words, the convolution is scaled by $1/r$. This rescaling is desirable because maximizations for the parameters can be separated when obtaining maximum likelihood estimates. Noninteger values of r are also permitted, since the distribution is infinitely divisible. The exponential parameter θ can be reparametrized

as $\tan(\theta) = \mu \in (-\infty, \infty)$, giving

$$f(x|\mu, r) = \frac{2^{r-2} e^{rx \arctan(\mu)} \mathrm{B}(\frac{r}{2} + i\frac{rx}{2}, \frac{r}{2} - i\frac{rx}{2})}{\pi r (1+\mu^2)^{\frac{r}{2}}}. \tag{12.2}$$

The density 12.1 12.2 is unimodal and skewed for nonzero values of μ; the skewness is

$$\gamma_3 \equiv \kappa_3/\kappa_2^{3/2} = V'(\mu)/\sqrt{V(\mu)} = \frac{\mu}{\sqrt{1+\mu^2}} \frac{2}{\sqrt{r}}, \tag{12.3}$$

where κ_j is the j^{th} cumulant, and $V(\mu)$ is the variance function, here $(1+\mu^2)/r$.

The tails behave like Gamma densities. Specifically, using the identity from Abramowitz and Stegun (Abramowitz and Stegun, 1970) (6.1.45)

$$\lim_{|x|\to\infty} \left| \Gamma\left(\frac{r}{2} + i\frac{rx}{2}\right) \right| / \sqrt{2\pi} e^{-\frac{\pi}{2}\frac{|rx|}{2}} |rx|^{\frac{r}{2}-\frac{1}{2}} = 1; \tag{12.4}$$

we have, for large positive x, the density becomes proportional to a Gamma(r , $r(\pi/2 - \arctan(\mu))$) density, and for large negative x, the density becomes proportional to a Gamma(r , $r(\pi/2 + \arctan(\mu))$) density. The relative sizes of the proportionality constants are determined by μ, and through the fact that as $\mu \to \infty, \frac{\pi}{2} - \arctan(\mu) \sim \frac{1}{\mu}$ we have that as $\mu \to +\infty$, the density becomes that of a Gamma(r , r/μ) density, and as $\mu \to -\infty$, to the density of the negative of a Gamma(r , r/μ). Thus an NEF_6 could be viewed as a smooth transition between a Gamma random variable, a random variable with bell-shaped density and gamma tails, and the negative of a Gamma, with the skewness parameter μ governing the weight and scaling given to each of the two densities. This compromise can also be seen from the variance function

$$V(\mu) = (1+\mu^2)/r. \tag{12.5}$$

For large μ the behavior is asymptotically equal to μ^2/r, a Gamma variance function, and for small μ the variance function is asymptotically equal to $1/r$, a normal variance function. See Figures 12 and 12 for some examples of the density function, showing the effect of varying θ, which is also the parameter linearly related to the regressor variables.

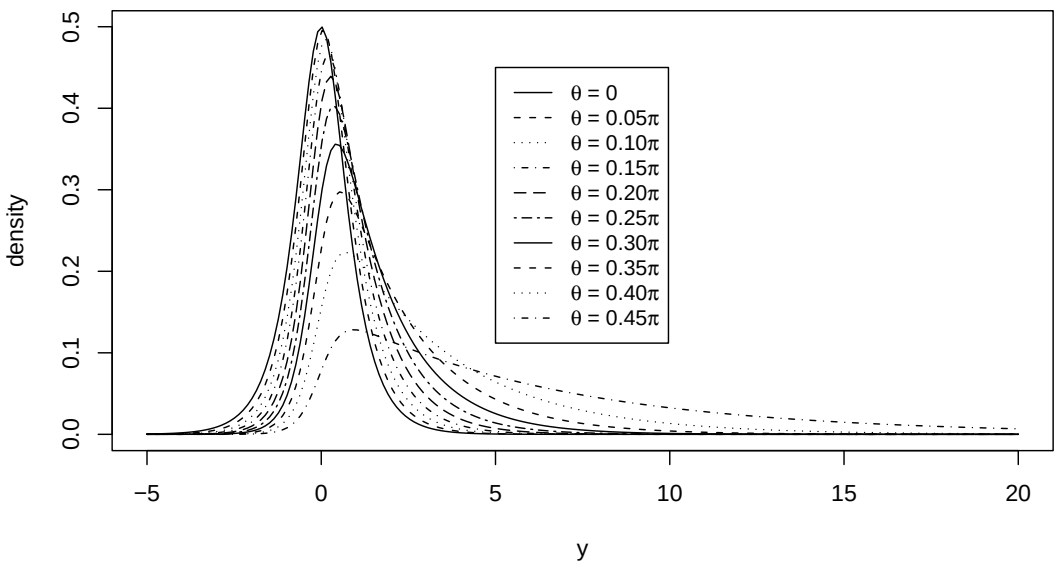


Figure 12: Some NEF_6 density functions (Equation 12.1), $r = 1$, illustrating the Gamma Limit Law as $\mu \to \infty$, $\theta = \arctan(\mu) \to \pi/2$.

The effect of increasing the value of r is to preserve the variance structure given by (12.5), while making the error distribution much more symmetrical, and asymptotically Gaussian as r increases.

The moment generating function for the NEF_6 distribution is given by

$$\mathrm{M}(t) = \mathrm{E}(e^{tX}) = (\cos(t/r) - \mu \sin(t/r))^{-r}, \tag{12.6}$$

and the characteristic function is given by

$$\phi(t) = \mathrm{E}(e^{itX}) = (\cosh(t/r) - i\mu \sinh(t/r))^{-r} \tag{12.7}$$

13 Generalized Linear Models

In this work we treat the canonical model, with link function $\tan(\theta)$, so that the linear predictor is restricted to $(-\pi/2, \pi/2)$. Since the link function blows up

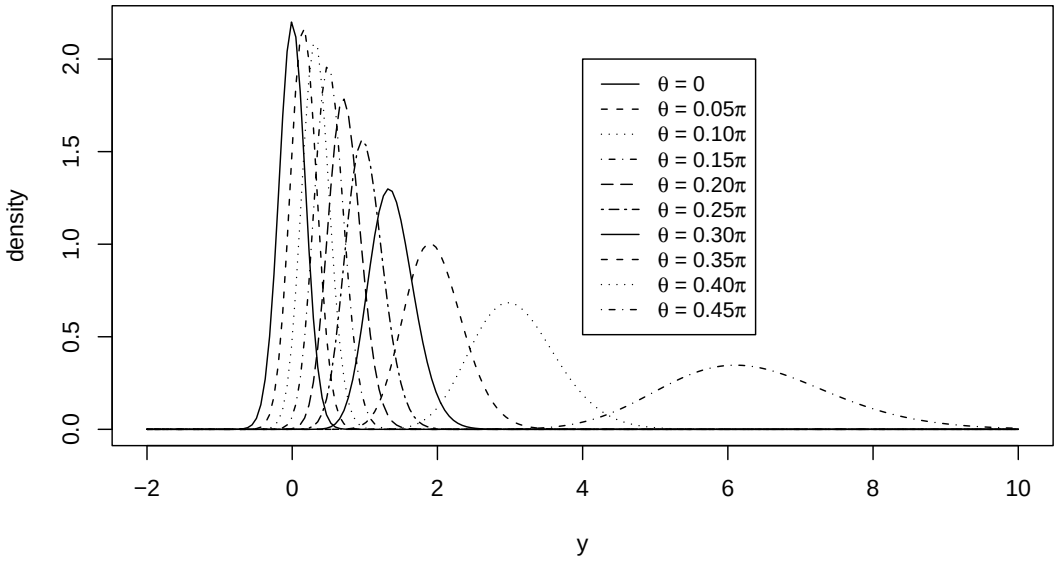


Figure 13: Some NEF_6 density functions, $r = 30$, illustrating the tendency towards Normal distributions, but preserving the quadratic variance structure

at either of the extremes, the fitting algorithm will generally tend to push the regression coefficients so that the linear predictor is well inside these bounds. We may consider other link functions in future work.

The Generalized Linear Model for the NEF_6 with the canonical link is given by

$$y_i \sim \text{NEF}_6(r, \theta = \beta' X_i), \tag{13.1}$$

where r and θ are the convolution and natural parameters as given in Equation 12.1. We may also fit or fix location and scale parameters for the model, replacing y_i with $(y_i - y_0)/s_0$, where y_0 is a location parameter and s_0 is a scale parameter. We will find these parameters necessary for the model to be useful.

For fixed r the model can be fit with an iteratively reweighted least squares algorithm. However, we wish to fit r as well, by maximum likelihood methods. It will be shown below, that we can fit the two parameters r and $\boldsymbol{\beta}$ separately, each in one step. There is a method of estimation given in McCullagh and Nelder

(McCullagh and Nelder, 1989) for r in the Gamma model, quite analogous to our model, in which r is fit via method of moments.

14 The Likelihood

The likelihood for the model is given by

$$\prod_{i=1}^{n} \frac{2^{r-2}}{\Gamma(r)} \frac{r}{s} \cos(\beta' X_i) \exp(r\theta y) \Gamma\left(\frac{r}{2} + i\frac{ry}{2}\right) \Gamma\left(\frac{r}{2} - i\frac{ry}{2}\right) \tag{14.1}$$

$$= \frac{2^{n(r-2)}}{(\Gamma(r))^n} \cos(\beta' X_i) \exp\left(\sum_{i=1}^{n} r\beta' X_i y_i\right) \prod_{i=1}^{n} \Gamma\left(\frac{r}{2} + i\frac{ry_i}{2}\right) \Gamma\left(\frac{r}{2} - i\frac{ry_i}{2}\right), \tag{14.2}$$

and thus the log likelihood for this model is

$$\begin{aligned} l(\boldsymbol{\beta}, r) &= n(r-2)\log(2) - n\log(\Gamma(r)) - n\log(r) + \log(\cos(\mathbf{X}\boldsymbol{\beta})) \\ &\quad + r\boldsymbol{\beta}'(\mathbf{X}'\mathbf{y}) + \sum_{j=1}^{n} \left\{ \log\left(\Gamma\left(\frac{r}{2} + i\frac{ry_j}{2}\right)\right) + \log\left(\Gamma\left(\frac{r}{2} - i\frac{ry_j}{2}\right)\right) \right\}. \end{aligned} \tag{14.3}$$

$$= rnD(\boldsymbol{\beta}) - n \cdot avg(\log f_{0,r}(y_j)), \text{ where} \tag{14.4}$$

$$D = \frac{(\mathbf{X}\boldsymbol{\beta})'\mathbf{y}}{n} + \frac{\sum_j \log(\cos(\boldsymbol{\beta}' X_j))}{n}, \tag{14.5}$$

$$\log f_{0,r}(y_j) = (r-2)\log(2) - \log(\pi) + \log B\left(\frac{r}{2} - i\frac{ry_j}{2}, \frac{r}{2} + i\frac{ry_j}{2}\right) - \log(r) \tag{14.6}$$

Thus we are not only modeling the error distribution as NEF_6, but the entire response as NEF_6 with mean equal to the tangent of the linear predictor. Thus the conditional expectation is a tangent curve, scaled and shifted horizontally, and also scaled and shifted vertically if the scale and location parameters are included. Thus the shape of the expected response may range from straight, if the response is in the center of the tangent curve, to curved in one direction, to an s-shaped curve, if the entire width of the tangent curve is included in the regression. A consequence of using the canonical link is that the model is not good for prediction based on out-of-sample values of the independent variable; extrapolation can lead to dangerous

situations. For such situations the identity link function would probably be a more sensible choice.

15 Fitting the Model

We fit the model by maximizing (14.3) with respect to the fitted parameters. To use a Newton-Raphson algorithm for these maximizations, we need first and second derivatives of (14.3). In the following formulas, $\Psi(\cdot)$ is the digamma function.

$$\frac{\partial l}{\partial r} = nD - n\Psi(r) + \sum_{i=1}^{n}\left\{\left(\frac{1}{2}+i\frac{y_i}{2}\right)\Psi\left(\frac{r}{2}+i\frac{ry_i}{2}\right)+\left(\frac{1}{2}-i\frac{y_i}{2}\right)\Psi\left(\frac{r}{2}-i\frac{ry_i}{2}\right)\right\} + \frac{n}{r} \tag{15.1}$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = rn\frac{\partial D}{\partial \boldsymbol{\beta}} = r\mathbf{X}'(\mathbf{y}-\tan(\mathbf{X}\boldsymbol{\beta})) \tag{15.2}$$

$$\frac{\partial^2 l}{\partial r^2} = -n\Psi'(r) + \sum_{i=1}^{n}\left\{\left(\frac{1}{2}+i\frac{y_i}{2}\right)^2\Psi'\left(\frac{r}{2}+i\frac{ry_i}{2}\right)+\left(\frac{1}{2}-i\frac{y_i}{2}\right)^2\Psi'\left(\frac{r}{2}-i\frac{ry_i}{2}\right)\right\} - \frac{n}{r^2} \tag{15.3}$$

$$\frac{\partial l}{\partial r\partial \boldsymbol{\beta}'} = \mathbf{X}'\left(\mathbf{y}-\tan(\mathbf{X}\boldsymbol{\beta})\right) \tag{15.4}$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = -r\mathbf{X}'\mathrm{diag}(\sec^2(\boldsymbol{\beta}'\mathbf{X}))\mathbf{X} \tag{15.5}$$

In the above equations $\mathbf{X}$ is a matrix of the values of X as its rows, and X_i is a column vector corresponding to the i^{th} row of $\mathbf{X}$.

Any root of (15.2) will be unique, since it is well known that the regression model with the canonical link for natural exponential families produces a concave log-likelihood. (McCullagh and Nelder, 1989)

Note that the expectation of $\partial^2 l/\partial r\partial\boldsymbol{\beta}$ is zero because of the formulation of the likelihood (14.4) with $\mathrm{NEF}_6(r,\mu)$ representing the average of r $\mathrm{NEF}_6(1,\mu)$ random variables, rather than the convolution.

The model is fit by maximizing the likelihood with r and $\boldsymbol{\beta}$. The parameter r is fit on the log scale, in other words we fit $\rho = \log(r)$ rather than fitting r directly,

because the likelihood is better approximated with a normal distribution at the mode on this scale. Fitting on the log scale of course does not matter in terms of the value found for the mode, but a Newton-Raphson algorithm will converge faster and be more stable on this scale, and the information matrix calculated on this scale will be different and produce confidence intervals with more accurate nominal coverage. Because of the structure of the log likelihood as expressed in (14.4) we can maximize the log likelihood separately for ρ and $\boldsymbol{\beta}$, and we need not iterate. The information matrix is then calculated at the mode using Equations 15.3 - 15.5 and the log transformation: $\partial l/\partial\rho = r\,\partial l/\partial r$ and $\partial^2 l/\partial\rho^2 = r\,\partial l/\partial r + r^2\,\partial^2 l/\partial r^2$. The inverse of the negative information is then the approximate covariance matrix of the fitted parameters at the mode.

Rather than maximizing (14.3) for fitting the model, we can also calculate the mode of the adjusted likelihood, where the likelihood is adjusted with a REML-type correction for fitting r in the presence of nuisance parameters, and the information is similarly calculated at the adjusted likelihood mode. This adjustment is as follows: the approximate marginal likelihood for r, integrating out $\boldsymbol{\beta}$ is given by

$$p(\rho) = \int p(\rho, \boldsymbol{\beta}) d\boldsymbol{\beta} \propto |\widehat{\mathbf{H}_\rho}|^{-1/2} L(\rho, \widehat{\boldsymbol{\beta}}), \tag{15.6}$$

where $\hat{\mathbf{H}}_\rho$ is the information matrix evaluated at $\hat{\boldsymbol{\beta}}$. From (15.5) we see that the determinant of $\hat{\mathbf{H}}_\rho$ is proportional to r^{-p}, where p is the number of regression parameters, and hence the REML adjustment is made by multiplying the likelihood by $r^{p/2}$. This adjustment should cause the resulting estimator to perform better than the unadjusted estimator, and is completely analogous to replacing n with $n - p$ in the denominator of the estimate for σ^2 in a normal model.

16 A Simulation Example

We started with 20 evenly spaced x values, between -1 and 1, and regression coefficient 0.4π with zero intercept, to ensure that the linear predictor is well

within the allowable range. We then use the generated linear predictor to simulate NEF_6 variates with $r = 1$ and μ equal to the tangent of the linear predictor. Thus the data appear as in Figure 14.

We also produce univariate and multivariate plots of the likelihood, which are shown in Figures 15 and 16, respectively. The profile likelihoods for r given $\boldsymbol{\beta}$ and for $\boldsymbol{\beta}$ given r are the same as the conditional likelihoods because of the separation of variables in the likelihood.

The likelihood is well-behaved, and the fact that the linear predictor is bounded seems to present little problem for the analysis; the poles of $\tan(\theta)$ at $\theta = \pm\pi/2$ tend to push the linear predictors away from the bounds during the optimization. The log likelihood is indeed close to quadratic near its mode, as can be seen from the preceding figures.

16.1 Performance of the MLE

We created 1000 datasets, simulated as above, and fit each of them with an NEF_6 regression model. The variability of the resulting parameter estimates can be seen in Figure 17. As expected the REML estimates for r are better centered around the true values than are the unadjusted ML estimates. We also used these datasets to estimate the true coverage of the nominal 95% confidence intervals, created with the Gaussian approximation to the likelihood or adjusted likelihood, at its mode. The coverage results are shown in Table 1.

17 Model Diagnostics

Residual plots and analyses can be somewhat problematic for non-normal GLMs. Simply plotting the residuals will produce a clearly heteroskedastic dataset that may be difficult to diagnose without intimate knowledge of the error distribution.

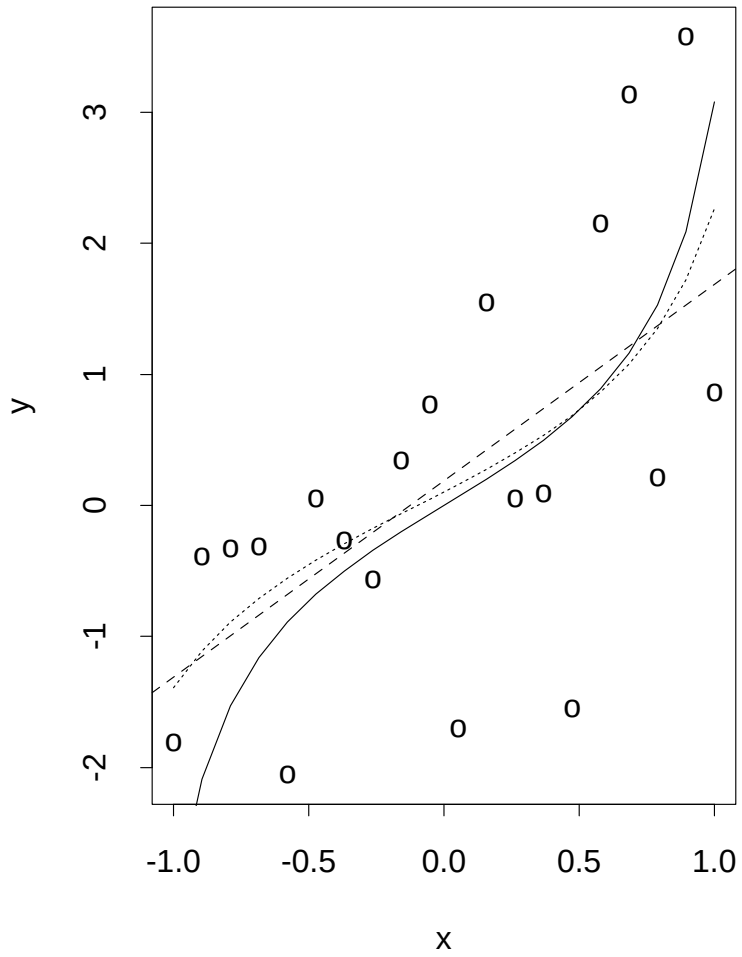


Figure 14: One simulated dataset for the NEF_6 canonical regression model. The true and fitted (nonlinear) conditional regression means are shown with solid and dotted lines, respectively, as well as the fitted ordinary least squares regression line (dashed).

The simulation used $r = 1$ with $\boldsymbol{\beta} = (0, 0.4\pi)$, and the fitted values were 1.113 (ML r), 0.944 (REML r), and (0.103,1.051) ($\boldsymbol{\beta}$).

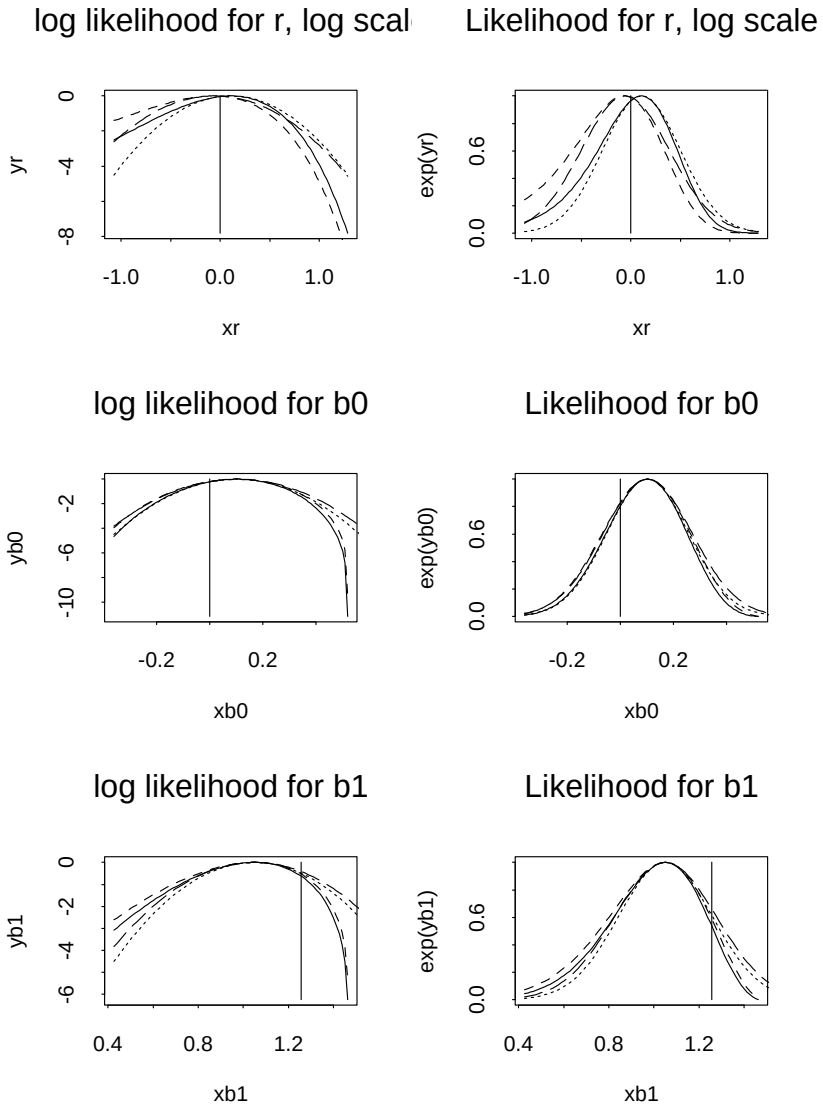


Figure 15: The univariate log likelihoods and likelihoods for the NEF_6 Regression fit, applied to the data shown in Figure 14. The likelihood functions, unadjusted and adjusted, and their quadratic approximations are shown, in solid, short dashed, dotted, and long dashed lines, respectively. Note that the log likelihoods for the regression coefficients begin to drop more sharply than its quadratic approximation on the side closer to the boundary of the linear predictor. Note that for these examples the profile likelihood and the conditional likelihood are the same, since the fitting procedure can separate the fitting of r and $\boldsymbol{\beta}$

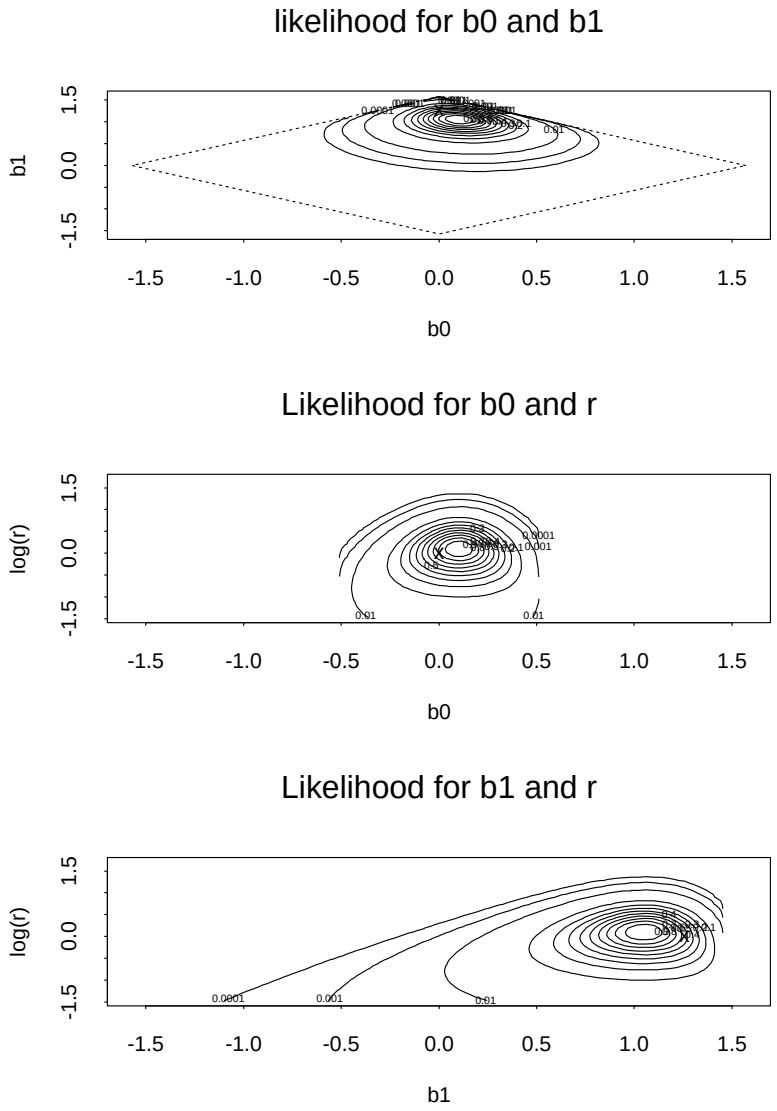


Figure 16: The pairwise likelihood functions for the fitted parameters. As expected, the regression parameters are negatively correlated, and r is roughly uncorrelated with the regression parameters, although the likelihood deviates from elliptical symmetry further from the mode. The true values of the estimands are marked with "x". The diamond marked with a dashed line in the upper panel is the boundary of the allowable parameter space for the regression coefficients.

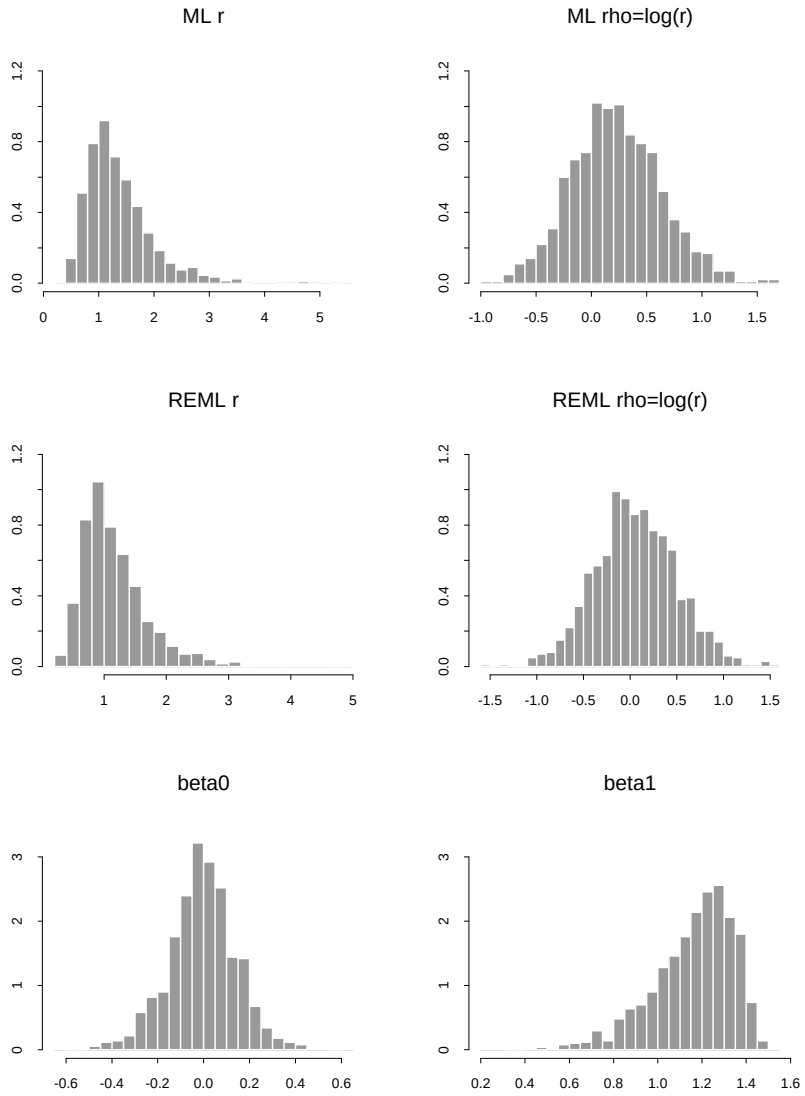


Figure 17: Parameter fits for 1000 simulated datasets. The true values are $r = 1$, i. e. $\log(r) = 0$, $\beta_0 = 0$, and $\beta_1 = 0.45\pi$. The estimation procedure can be seen to be reasonably unbiased in this case, and the variability can be seen in the histograms.

Actual Coverage of Nominal 95% Intervals				
Parameter	r (ML)	r (REML)	β_0	β_1
Coverage	0.881	0.936	0.941	0.930

Table 1: Coverage of nominal 95% intervals, based on a simulation of 1000 datasets. The true values used were 1.0 for r, and $(0, 0.45\pi)$ for $\boldsymbol{\beta}$. Here we can see, for this example, the improved coverage of the REML intervals for r over the ML intervals.

Pearson residuals attempt to resolve this be rescaling each residual by its standard deviation; however, the conditional distribution is still non-normal, and in the NEF_6 case will be skewed differently at different values of the linear predictor. This nonnormality may still be difficult to assess. Anscombe and Deviance residuals are attempts to deal with these issues using approximations, but may be even more difficult to interpret (McCullagh and Nelder, 1989). We can produce model diagnostic plots especially tailored for these models, by plotting sorted quantiles for the residuals within their conditional distributions against uniform quantiles. Such plots, although they may not correctly diagnose all problems, can provide a good indication of poor model fit. We will show some such plots later in Section refsec:example.

18 Other Models

Up to now we have fit the model (13.1) with a fixed value of the scale parameter s_0, so the fit will be different for different scalings of the data. The only way model (13.1) can account for extreme observations is to adjust the skewness in one direction; adding a scale parameter may allow the model to account for larger variances without extreme skewness. The log likelihood for such a model is given in Equation 18.1.

$$l(\boldsymbol{\beta}, r, s_0) = n(r-2)\log(2) - n\log(\Gamma(r)) - n\log(r) + n\log(s_0) + \log(\cos(\mathbf{X}\boldsymbol{\beta})) \tag{18.1}$$

$$+ r\boldsymbol{\beta}'(\mathbf{X}'\mathbf{y}/s_0) + \sum_{j=1}^{n}\left\{\log\left(\Gamma\left(\frac{r}{2} + i\frac{ry_j}{2s_0}\right)\right) + \log\left(\Gamma\left(\frac{r}{2} - i\frac{ry_j}{2s_0}\right)\right)\right\}. \tag{18.2}$$

$$= rnD(\boldsymbol{\beta}) - n \cdot avg(\log f_{0,r}(y_j/s_0)), \text{ where} \tag{18.3}$$

$$D = \frac{(\mathbf{X}\boldsymbol{\beta})'\mathbf{y}}{ns_0} + \frac{\sum_j \log(\cos(\boldsymbol{\beta}' X_j))}{n}, \tag{18.4}$$

$$\log f_{0,r}(y_j) = (r-2)\log(2) - \log(\pi) + \log B\left(\frac{r}{2} - i\frac{ry_j}{2s_0}, \frac{r}{2} + i\frac{ry_j}{2s_0}\right)$$

$$- \log(r) + \log(s_0) \tag{18.5}$$

However, there is a potential issue with fitting both convolution and scale parameters, which can be seen as follows: the limit as $\mu \to \pm\infty$ is a gamma distribution, for which the scale and convolution parameters are one and the same. Also, the convolutions parameter r appears in the likelihood everywhere the scale parameter s_0 does. Thus we might expect the two parameters to play similar roles in fitting the model. Figures 18 and 19 illustrate a simulated fit for both r and s_0, with a dataset of size $n = 20$ simulated from the model (13.1), and indeed, we find that the contour plot of the likelihood against $\log(r)$ and $\log(s)$, the scale parameter, near the MLE, the likelihood surface has a diagonal feature between the two parameters, indicating some interchangeability between the two parameters. This feature extends over a substantial range on the log scale: roughly, an interval containing 95% of the area under the likelihood surface has a width of one, corresponding to a multiplicative factor of e on s_0, which is a substantial range for a fitted value on the order of $s = 1000$. This feature could easily become a ridge under more extreme circumstances. The ridge in the likelihood would also adversely affect the fitting procedure, which often must pass through many

iterations before convergence, and might be subject to numerical irregularities as a result of these many iterations. So, although we can fit both of these two parameters, we are not sure of the robustness of the fitting procedure, and we suspect that this model may be too flexible for smaller datasets, which might not contain enough information to accurately fit both of these parameters. It is quite possible that some transformation of these two parameters might have less correlated likelihoods or posteriors, and this is a topic for future work. Our work thus far has been focused on relatively small datasets, and it is quite possible that with a larger dataset it will be possible to fit more parameters, and more effectively decide among models. Thus we mention this extension of the work here for completeness, but we will postpone further investigation of fitting scale parameters for future work.

19 Example: Estimating Beta for Stock Returns

A standard problem in finance is the estimation of beta relative to some benchmark. According to Grinold and Kahn (Grinold and Kahn, 1995), beta for a portfolio P is defined as the covariance of the excess returns on P and the excess returns on the benchmark divided by the variance of the excess returns on the benchmark. In other words, it is the regression coefficient when the excess benchmark returns are the explanatory variable and the excess portfolio returns are the response variable. Excess returns are defined as total returns minus the total returns on risk-free assets over the same time period. These betas are important, as they reflect the relation of a portfolio to a benchmark, in particular an entire market, and can be used with pricing models to put together more effectively diversified portfolios. Excess returns are positive or negative, and we expect them to demonstrate skewed behavior, so modeling with a distribution such as NEF_6 does not seem unreasonable.

Figure 20 shows the monthly excess returns from April 1979 to December 1989

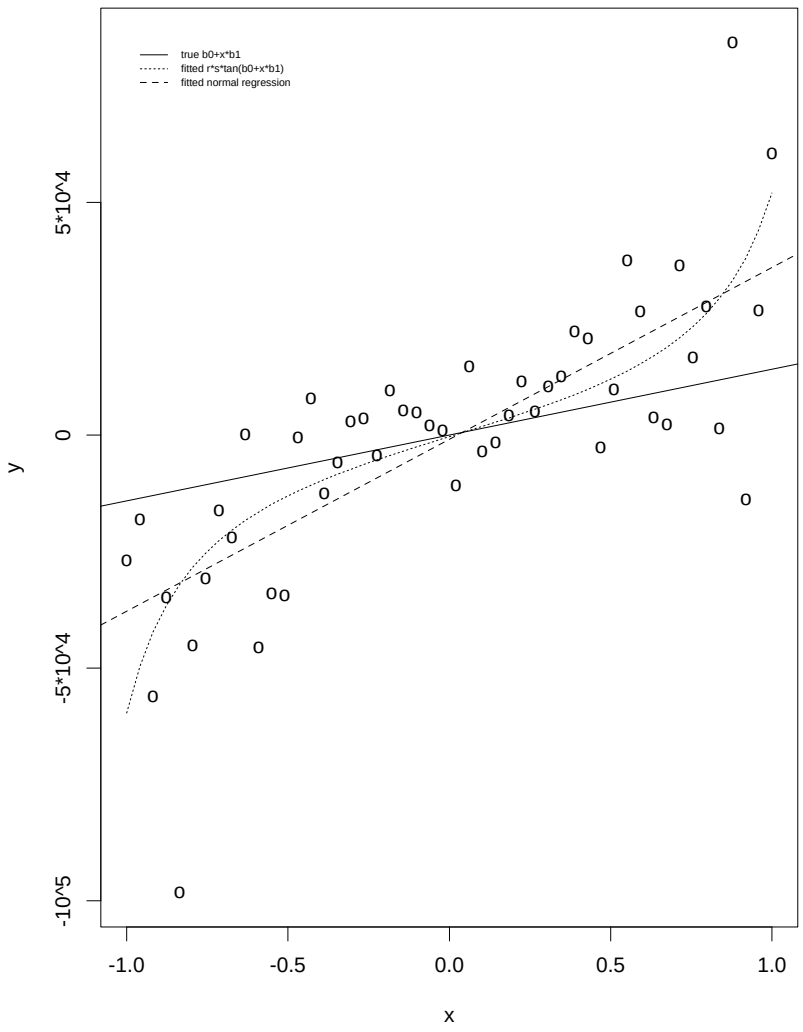


Figure 18: A plot of a simulated dataset (n=50) from the model (13.1), $s = 10000$, $r = 1$. The normal regression is too steep at the center, as would be expected from the outliers at the extremes.

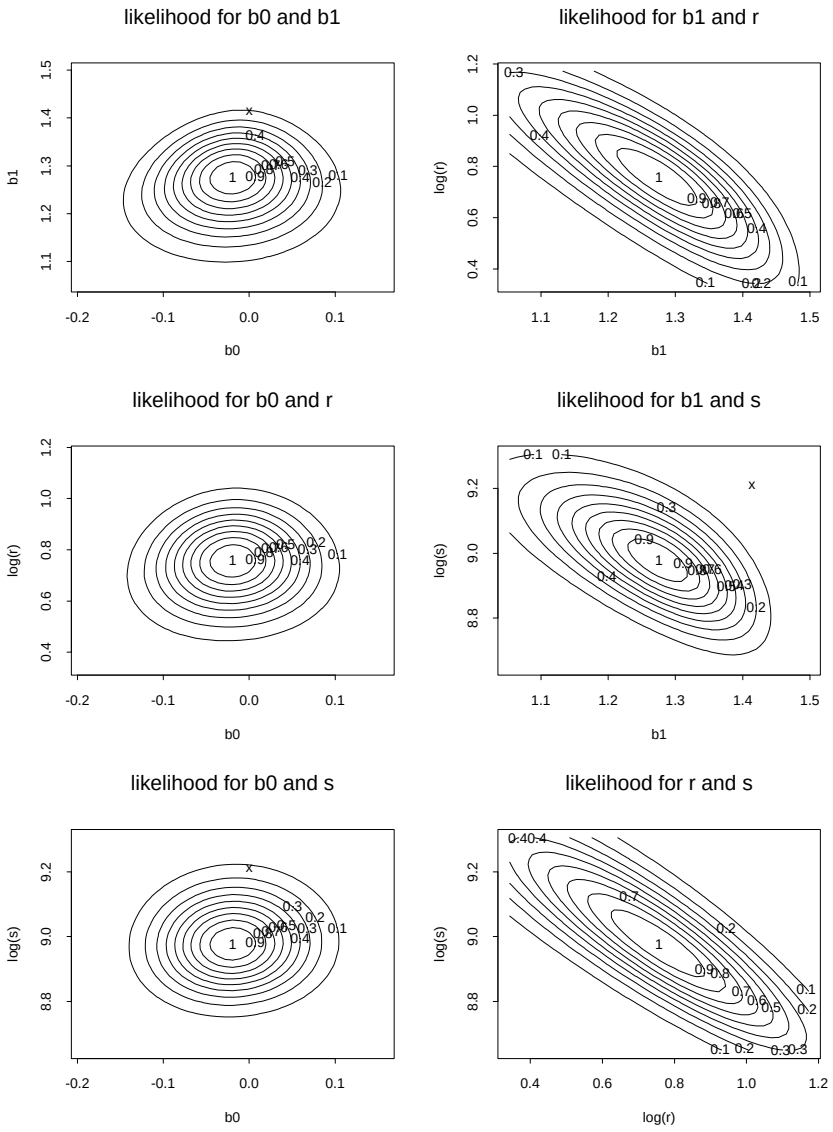


Figure 19: The bivariate conditional likelihoods for the fit of the data in figure 18. The likelihoods for $\rho = \log(r)$ and $\sigma = \log(s)$ show a strong interaction between the two parameters.

(129 time periods) of Oil City Petroleum, Incorporated, and the monthly excess returns of the market of the same period. We can see immediately that there is a heteroschedastic variance structure in the data, with the standard deviation of the response increasing as the explanatory variable increases. This variance does not decrease to zero at the low end. There is also one extreme value in both values, corresponding to March of 1987. This dataset was used in the *SPLUS Guide to Statistics* (Division, 1999) to illustrate the use of robust regression techniques, in relation to the outlier. We will use our various NEF_6 models to fit the data, and compare the results to more traditional techniques.

We can clearly see that inclusion or exclusion of the outlying point in traditional OLS regression methods has a large impact on the outcome of the procedure. We can hope that the NEF_6 regression procedures might be able to accommodate the outlier by fitting a skewed conditional distribution at the extreme value, and less skewness at the more central points where most of the data are clustered. This might produce an estimate of beta that is more resistant to outliers, and whose predictive distributions will be more faithful to both the variance structure and the skewness of the dataset. Although such a hope is not immediately borne out by our model fits, we can use these fits to learn about the shortcomings of the model presented here, and determine how to overcome these shortcomings.

Figures 20-28 show the fits for models not including location parameters, but with scale parameters set arbitrarily to three values. The figures are divided into three sets of three, each set being a plot of the data and the conditional means of the NEF_6 models, univariate conditional likelihoods, and bivariate conditional likelihoods, setting other parameters to their estimates. Figures 20 - 22 show fits with s_0 set to 0.01. In order to compensate for this small scale parameter, the fitted values of r are also quite small, less than 0.01. These distributions with small r can be extremely skewed, as they would need to be to accommodate extreme values in the data, but have most of their density in the middle peak. The fits including the outlier have even smaller fitted values of r. The model diagnostics shown in

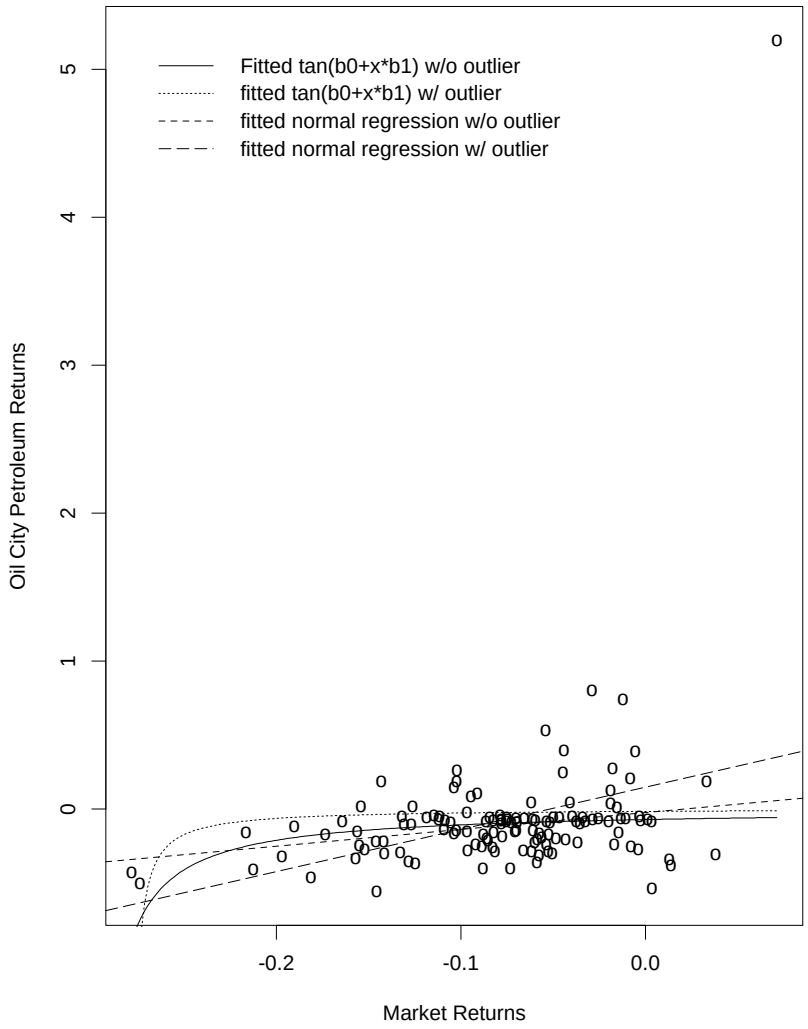


Figure 20: Some model fits for r with s_0 set to 0.01. The model fits, excluding the outlier, are thus: $\hat{r}_{ML} = 0.0482, \hat{r}_{REML} = 0.0472, \hat{\beta}_0 = -1.47, \hat{\beta}_1 = 0.457$; including the outlier, $\hat{r}_{ML} = 0.00997, \hat{r}_{REML} = 0.00815, \hat{\beta}_0 = -1.18, \hat{\beta}_1 = 1.95$. The NEF_6 model fits are particularly poor at the extremes of the independent variable, and definitely perform worse in a least-squares sense than a normal regression model

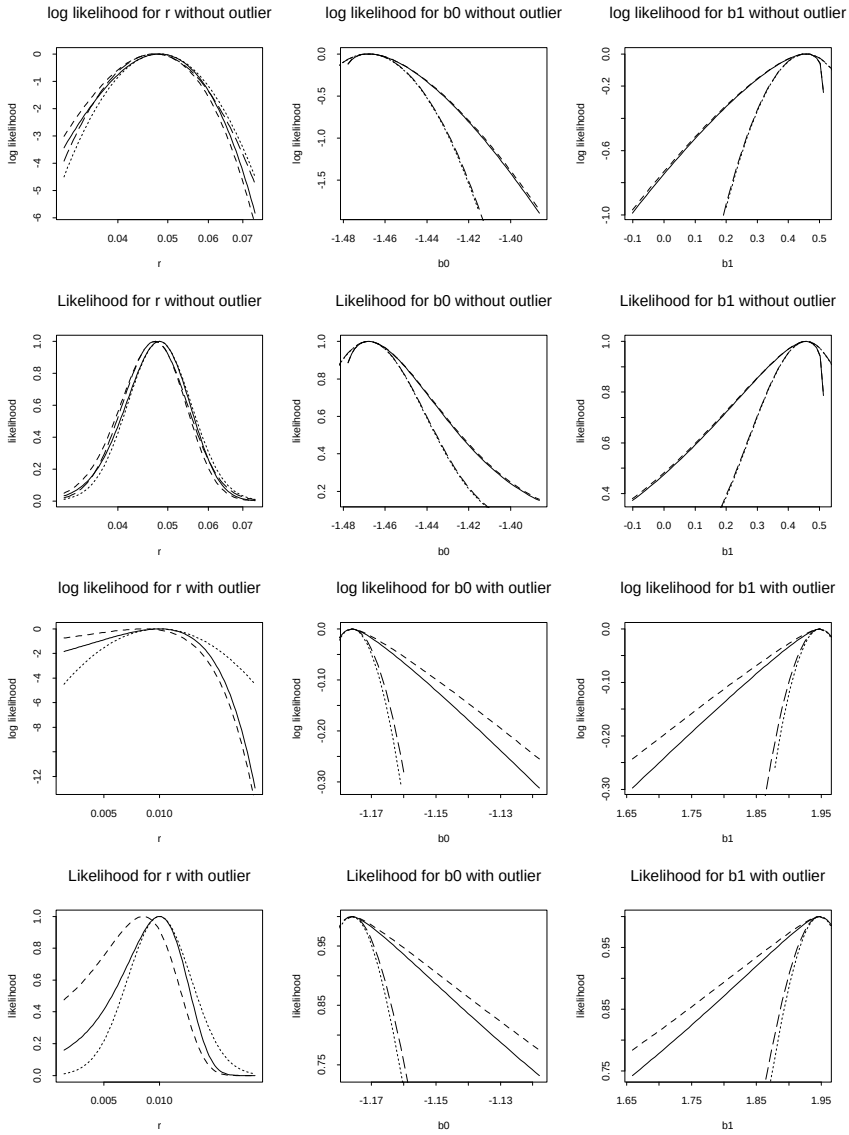


Figure 21: Conditional Likelihoods for the fitted parameters, s_0 fixed at 0.01. In each graph the other parameters are set to their estimates. The solid lines are the likelihoods, using the ML estimates of r, the short dashed lines are the quadratic approximations at the mode. The long dashed lines are the restricted likelihoods, using the REML estimates of r, and the dotted line is its corresponding quadratic approximation. The lower six panes, representing analyses which include the outlier, show a significant disruption in model fit.

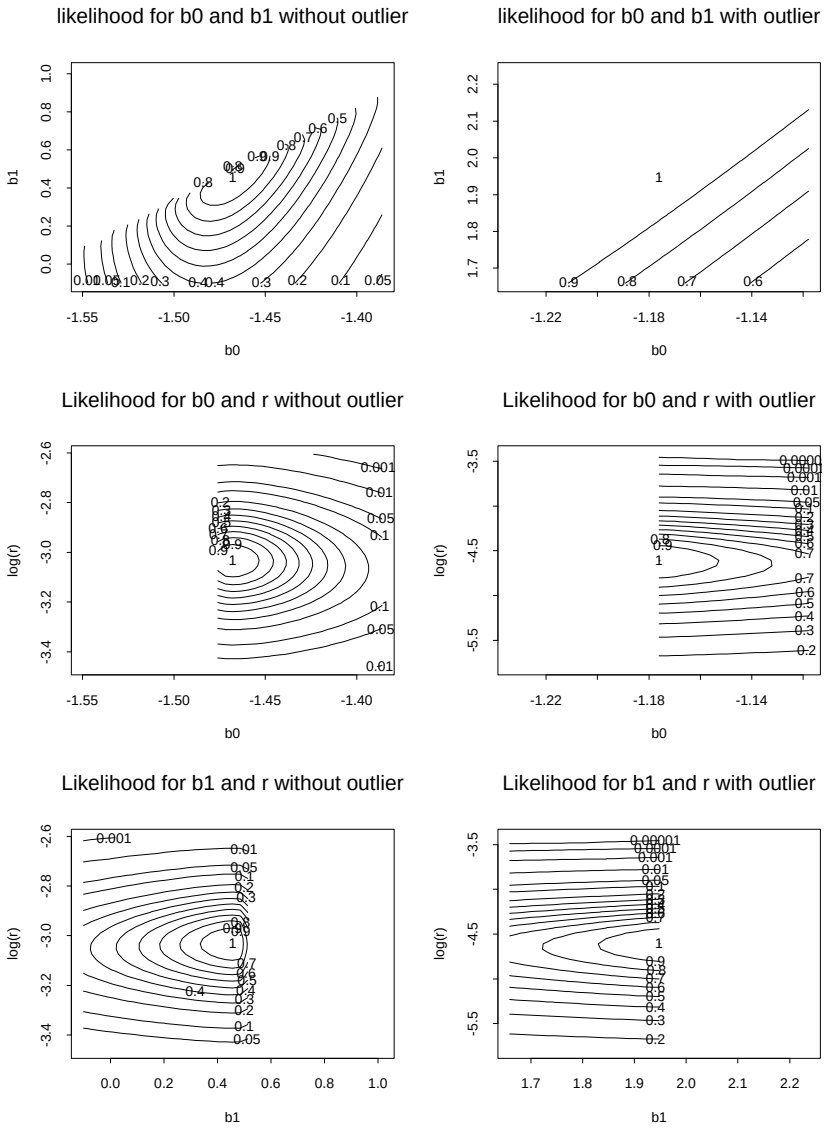


Figure 22: Bivariate Conditional Likelihoods for the fitted parameters, with s_0 fixed at 0.01. All other parameters are fixed at their maximum likelihood estimates. Further evidence of the poor model fit is seen in the proximity of the maxima to the boundaries of the parameter space.

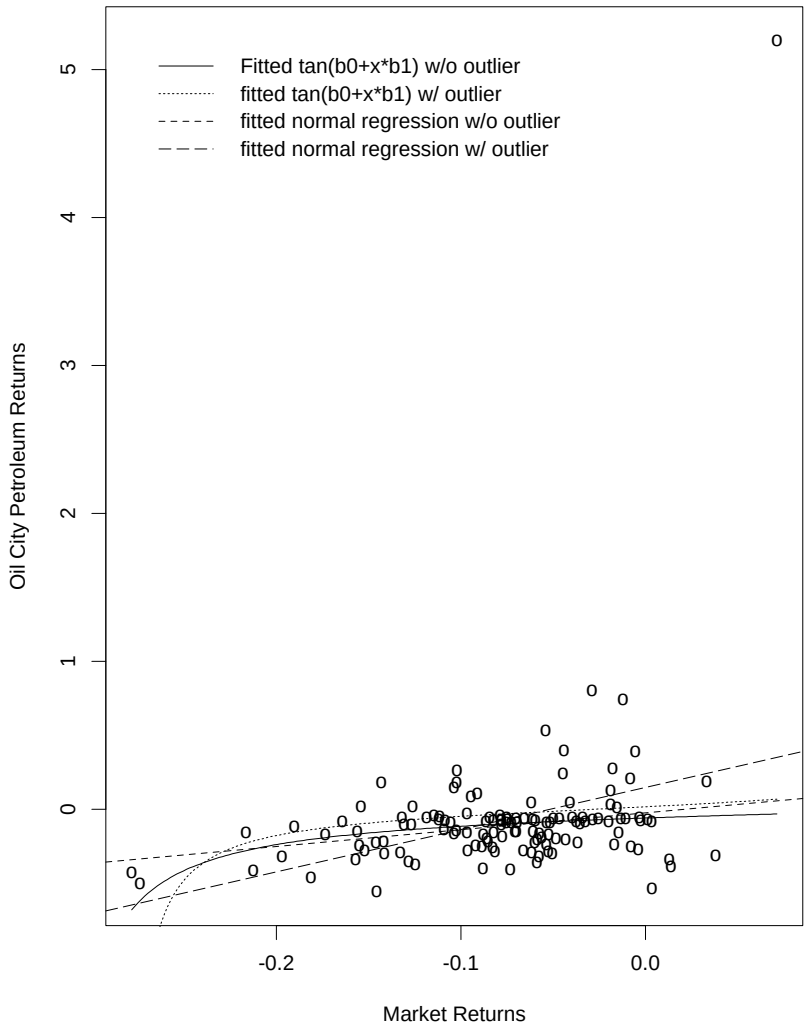


Figure 23: Some model fits for r with s_0 set to 0.1. The model fits, excluding the outlier, are thus: $\hat{r}_{ML} = 0.424, \hat{r}_{REML} = 0.411, \hat{\beta}_0 = -0.785, \hat{\beta}_1 = 3.190$; including the outlier, $\hat{r}_{ML} = 0.185, \hat{r}_{REML} = 0.171, \hat{\beta}_0 = -0.306, \hat{\beta}_1 = 6.117$. The model fits are still quite poor, especially at the extremes of the independent variable.

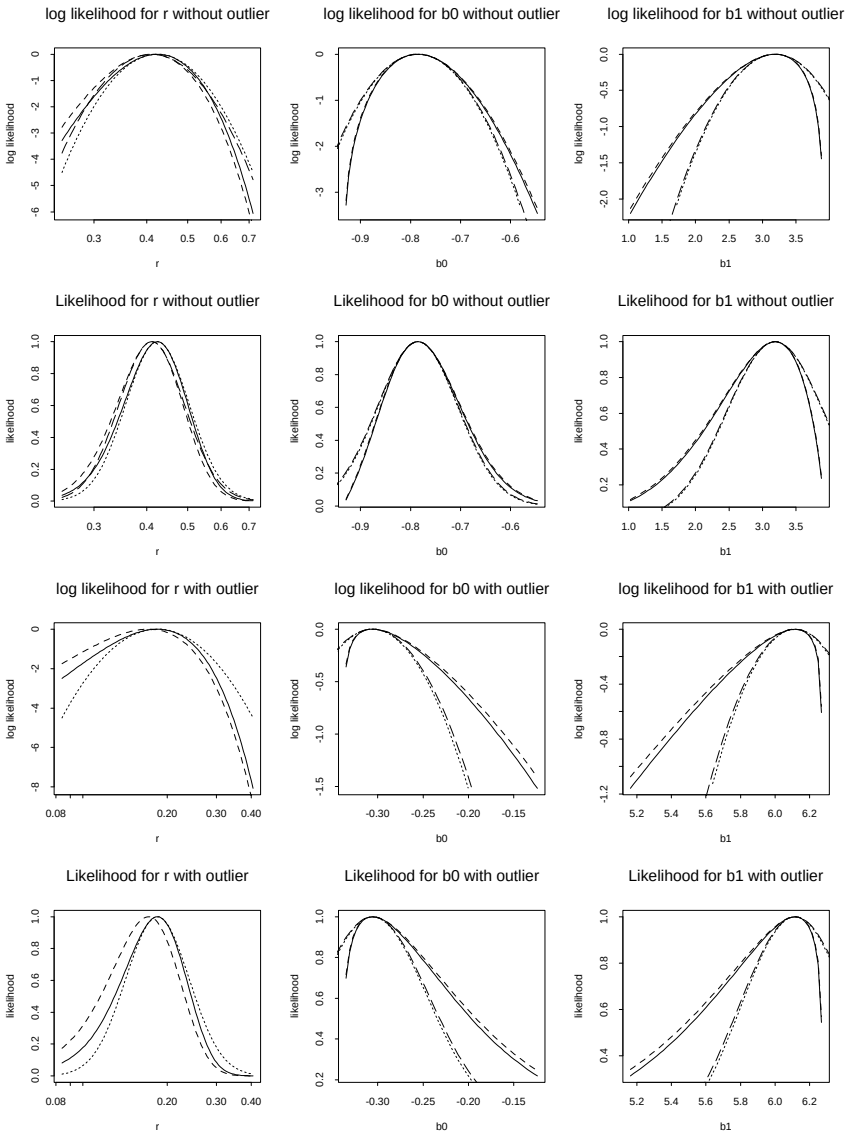


Figure 24: Conditional Likelihoods for the fitted parameters, s_0 fixed at 0.1 In each graph the other parameters are set to their estimates. The solid lines are the likelihoods, using the ML estimates of r, the short dashed lines are the quadratic approximations at the mode. The long dashed lines are the restricted likelihoods, using the REML estimates of r, and the dotted line is its corresponding quadratic approximation.

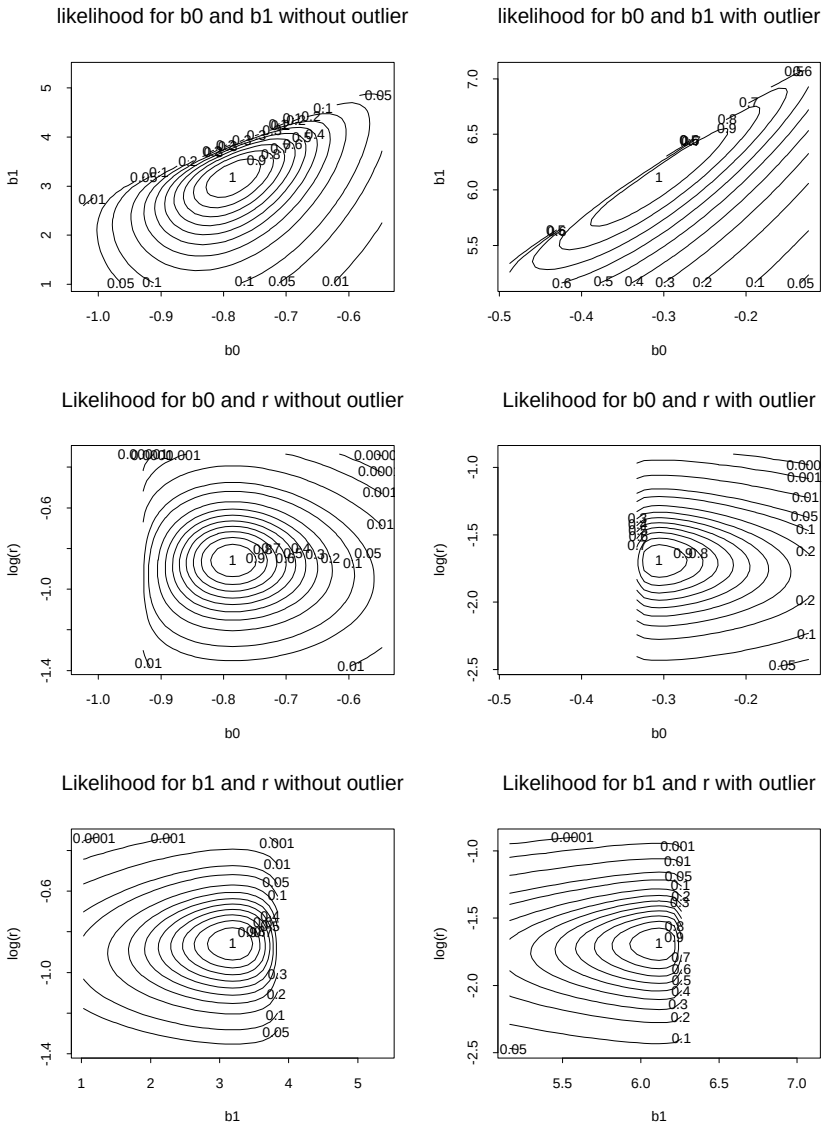


Figure 25: Bivariate Conditional Likelihoods for the fitted parameters, with s_0 fixed at 0.1. All other parameters are fixed at their maximum likelihood estimates. The maxima here are further away from the parameter space boundaries than in the previous fit, but still quite close. In any case, normal approximations are likely to be poor.

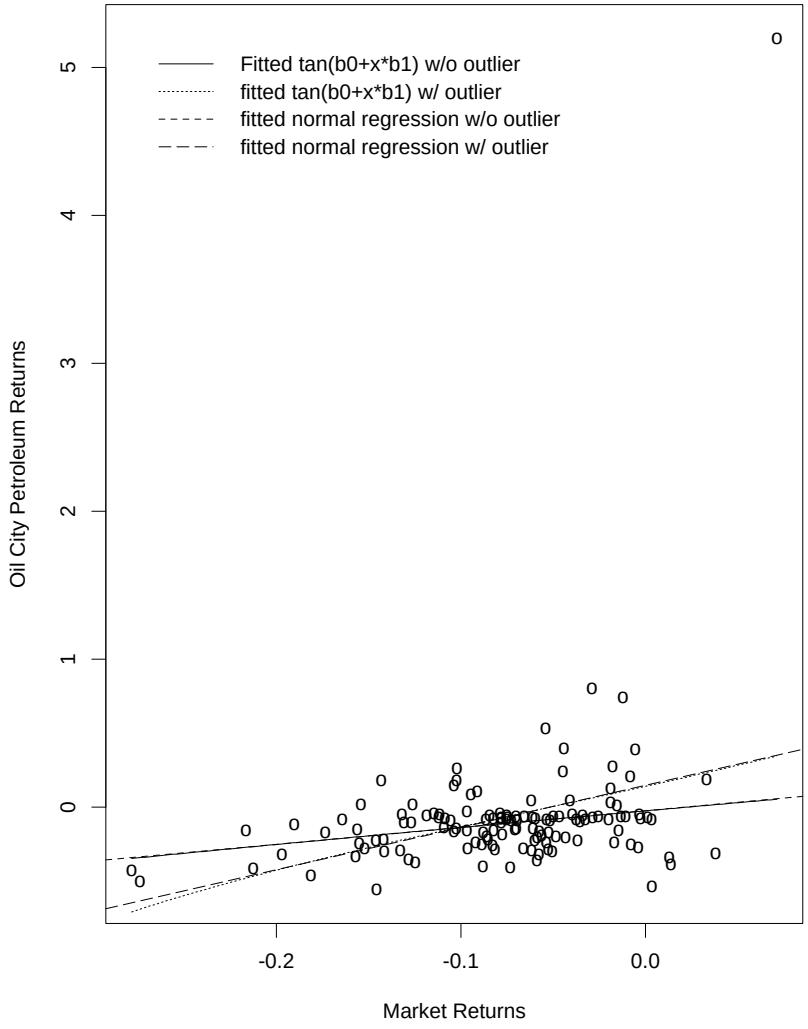


Figure 26: Some model fits for r with s_0 set to 1. The model fits, excluding the outlier, are thus: $\hat{r}_{ML} = 25.1, \hat{r}_{REML} = 24.7, \hat{\beta}_0 = -0.112, \hat{\beta}_1 = 1.11$; including the outlier, $\hat{r}_{ML} = 8.73, \hat{r}_{REML} = 8.57, \hat{\beta}_0 = -0.0692, \hat{\beta}_1 = 2.714$. The NEF_6 fits are extremely close to the normal fits here, and do not take advantage of the flexibility of the nonlinear conditional mean of the model.

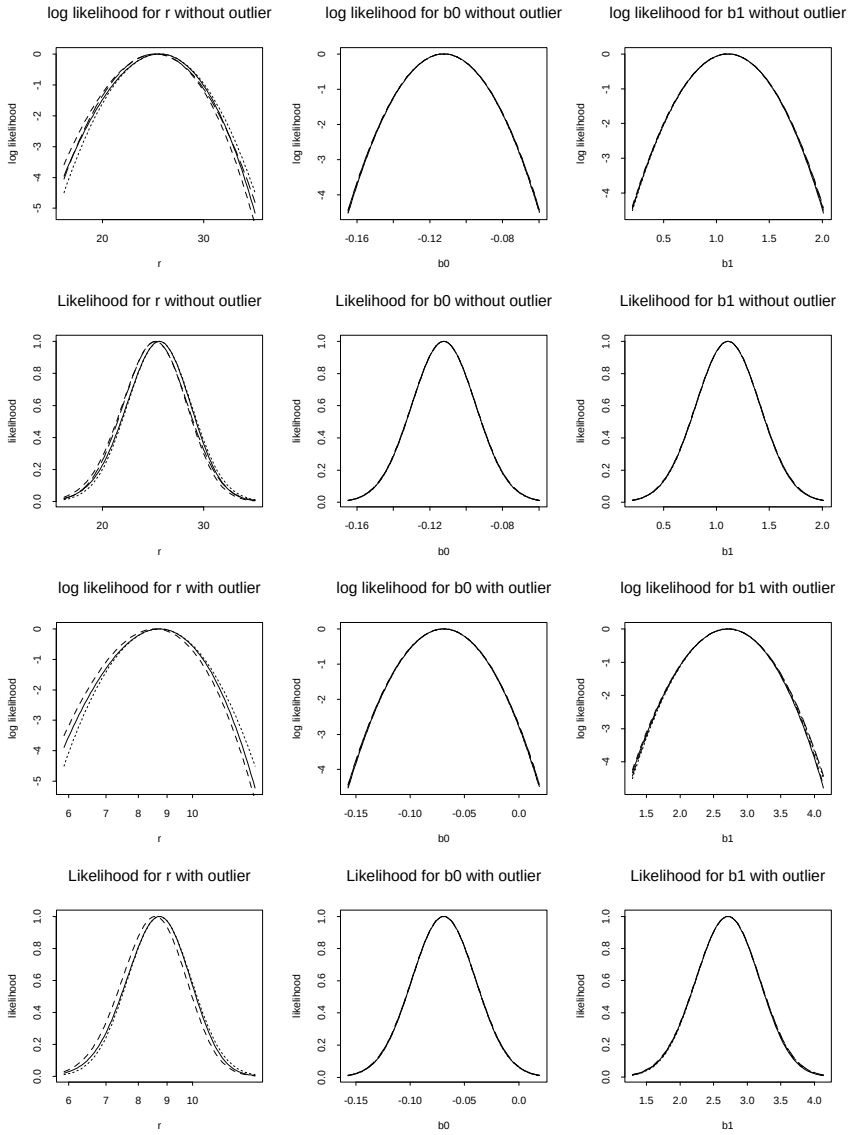


Figure 27: Conditional Likelihoods for the fitted parameters, s_0 fixed at 1. In each graph the other parameters are set to their estimates. The solid lines are the likelihoods, using the ML estimates of r, the short dashed lines are the quadratic approximations at the mode. The long dashed lines are the restricted likelihoods, using the REML estimates of r, and the dotted line is its corresponding quadratic approximation. The normal approximations are quite good, because the model being fit is quite close to normal.

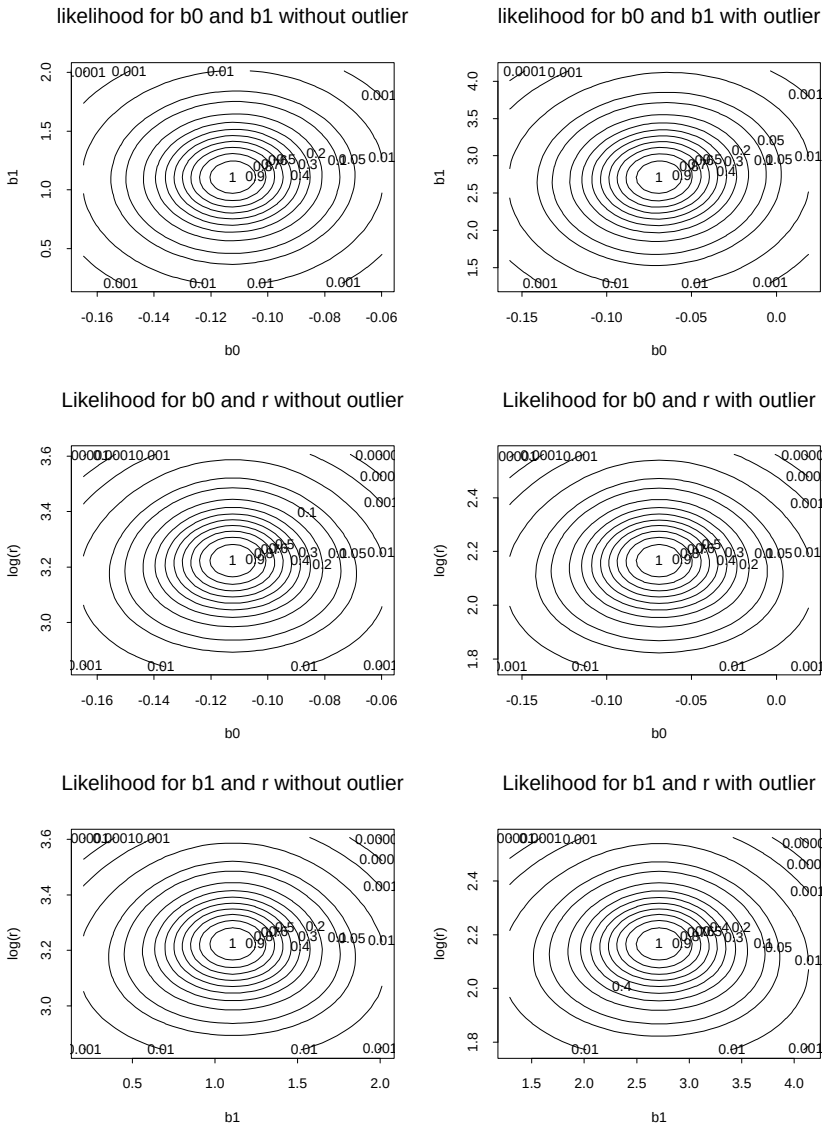


Figure 28: Bivariate Conditional Likelihoods for the fitted parameters, with s_0 fixed at 1. All other parameters are fixed at their maximum likelihood estimates. In this case all of the normal approximations are quite good, as the model fit is very close to a normal model fit.

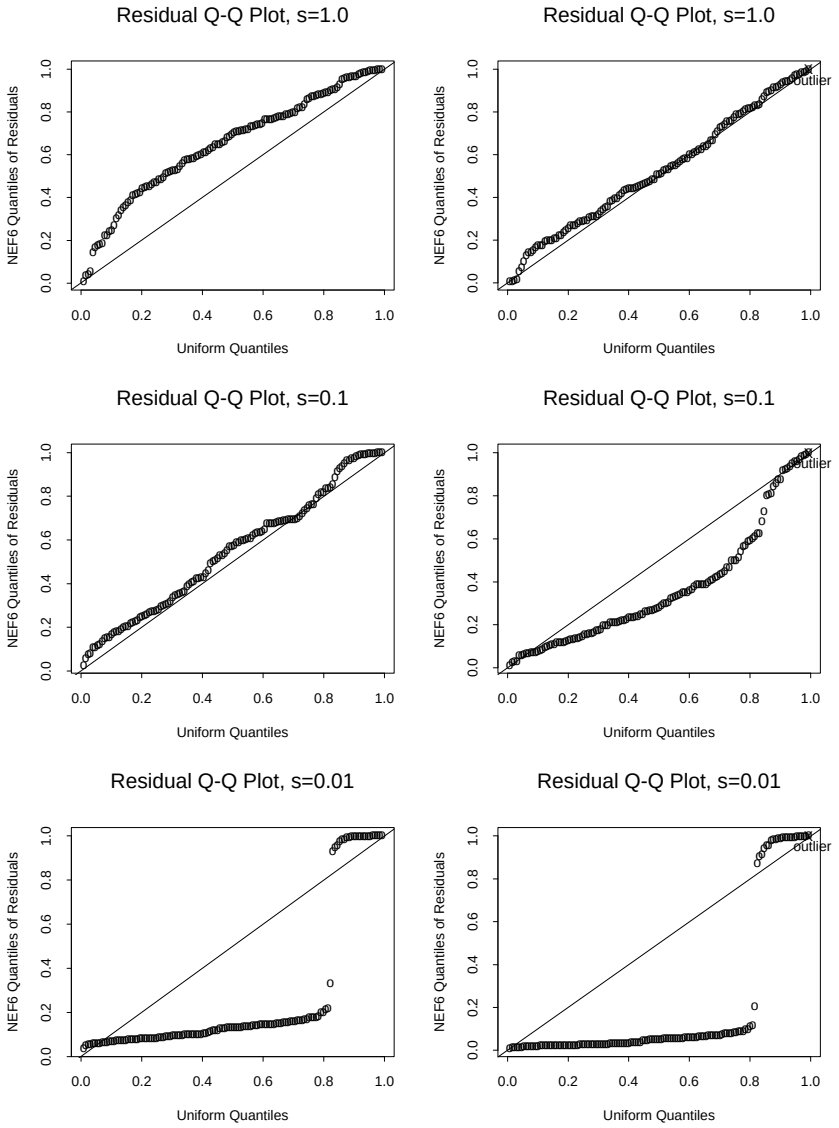


Figure 29: Diagnostic plots for the three fits. The left-hand panes are the fits without the outlier, and the right-hand panes are with the outlier. These plots clearly provide evidence of poor model fit, especially for $s = 0.01$, where all of the residuals' NEF_6 quantiles are in the lower or upper end of the unit interval. Interestingly the fit is much poorer without the outlier than with, for the $s = 1.0$ case, and neither fit is very good for $s = 0.1$. The outlier is the extreme point in all three fits.

Figure 29 show the poorness of fit for this model, because the residuals all fall in the low or high ends of their fitted distributions. The next three figures, 23 - 25 show the fits with s_0 set to 0.1. These models provide a better fit than the previous three, but there are still some issues. The two points on the left of Figure 23 are not fit well, and the conditional expectation does not curve up to fit the outlier when it is included in the analysis. The model is unable to fit the increasingly positive values on the right with positively skewed distributions because of the many negative values in this region. The likelihoods are still far from Gaussian, and although this is not necessarily a bad thing, we might be concerned that so many of the estimates are so close to the parameter space boundaries. The third set, Figures 26 - 28 have s_0 set to 1.0. These three fits end up with fitted values of r that are quite large, and thus, the model fits are close to the normal model fits, as would be expected. For the improved fits below we used a value of 0.2 for s_0, which seemed to be close to optimal.

In some of these model fits, e. g. Figures 20 and 23, we can see that the slope of the conditional expectation in the central part of the fit is relatively unaffected by the inclusion of the outlier, unlike the OLS fits. However, the model cannot be described as "resistant", because inclusion of the outlier significantly affects the model fit. All of the model fits are relatively poor, particularly at the lower end of the independent variable, where there is a sharp dip in the conditional expectation. We will discuss below why this is the case, and how to remedy this problem. We can still learn something about the model from these fits, though. The likelihoods show a very interesting fact, which carries through on all analyses of these data: the parameter space for β is bounded by a two-dimensional quadrilateral in such a way that the linear predictors of none of the points may reach a greater absolute value than $\pi/2$. When the outlier is included in the regression, the mass under the likelihood surface is shifted to be closer to this boundary. This occurs in both model fits, and the reason is clear: The model needs to use the more skewed sections of the model to fit the data, and the skewed conditional distributions appear when

the linear predictor is closer to $\pi/2$, i. e. the regression coefficients are closer to their boundaries. The model fitting procedure, however, is still quite stable, and if need be caution can be taken that the fitting procedure does not produce iterations outside the allowable parameter space, by checking at each iteration if such a step is called for, and damping the suggested jump or backtracking slightly in the algorithm. In the thousands of model fits we have done it has not been necessary to do this. It is quite interesting to note the variation among fits of the data with various scalings. When the scale parameter is large, e. g. in this case 1, the NEF_6 density easily accommodates the outlier, so it fits the model with the central part of the parameter space for the regression coefficients, and the likelihood itself is very well approximated by a normal curve. In fact, the curves of conditional expectations differ very little from their corresponding OLS regression lines, for the range of the independent variable in the sample. Thus in this case we are not capitalizing on the advantageous features of the NEF_6 model, and the fixed value of s_0 is probably too high. For smaller values of the scale parameter, i. e. when the dependent variable values input into the model are much greater, the conditional distribution must change more to account for these extreme values, as they are much further out in the tail of the conditional density. Thus the model fits a value for the linear predictor much closer in absolute value to $\pi/2$. In fact, when the fixed value of s_0 is small, the fitted value of r is large, and vice-versa. The fitted value is trying to make up for the incorrect scaling of the data.

The model diagnostic plots shown in Figure 29 corroborate the poorness of fit for these data. All of the panes show problems at the lowest end of the distribution, and big problems are indicated by the large gaps in the coverage of the error distribution in the lower panes with $s = 0.01$.

As mentioned above, the model fits are poor in terms of fitting the data, and often probably worse in a least-squares sense than the traditional OLS models. The reason for this is as follows: the only way the model has to fit location is to change the skewness of the conditional distribution. However, a positively skewed

conditional distribution with no location change cannot easily accommodate negative values, of which there are many. Simply looking at the dataset, we would expect a positive skewness in the conditional distribution for nearly all of the range of the independent variable, hence there should not be a preponderance of negative values in the response variable, to fit these values with a NEF_6 error distribution. Hence we might expect a better model fit if we subtract the right constant from the response variable before fitting the model. Since the variables themselves were constructed as differences between stock prices, or averages of stock prices, and a baseline risk-free rate, it is not unreasonable to subtract another baseline value from these values to fit the data to the model with NEF_6 error distribution. In fact it is clear that this offset should be fitted in the model like the other parameters. As a preliminary effort to verify that adding a location parameter improves model fit, we fit the model with offsets in the response variable, and we can indeed obtain a much closer fit in the conditional expectation to the data, especially on the left where the fits were poor without the offset. We will include the location parameter as a fitted parameter in the model in future implementation. Figures 30 - 32 show the model fits and likelihoods for the models fit with offsets in the independent variable. We can see in the likelihood plots that the fitted values are more comfortably away from the boundaries of the parameter space.

The diagnostic plots for these final model fits, shown in Figure 33, show much improvement over the previous fits, as expected. Although they are not perfect, it should be noted that two of the model parameters, the location and scale of the response variable, were hand-selected, and an optimized fit of these parameters could probably provide an even better fit. However, there are not huge systematic departures from the diagonal line in these plots, and so the Q-Q plots are acceptable from a model diagnosis standpoint.

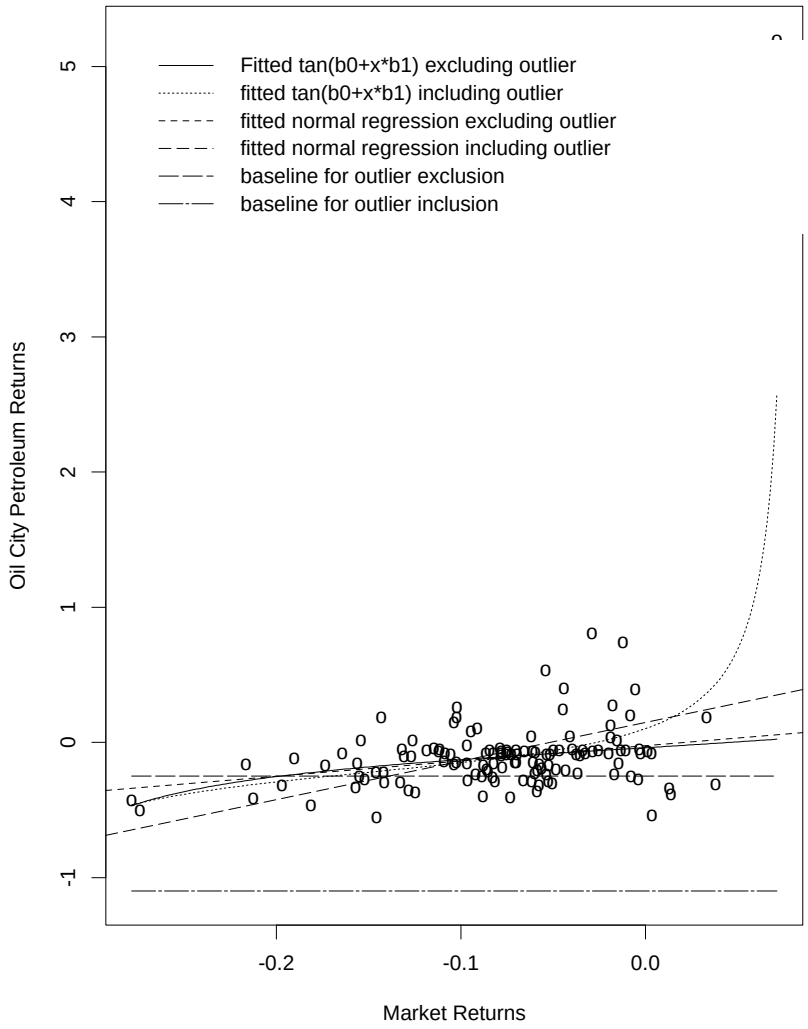


Figure 30: Some model fits for r with s_0 set to 0.2, and offsets as shown in the response variable. The model fits, excluding the outlier, are thus: $\hat{r}_{ML} = 1.41, \hat{r}_{REML} = 1.11, \hat{\beta}_0 = -0.278, \hat{\beta}_1 = 4.221$; including the outlier, $\hat{r}_{ML} = 1.37, \hat{r}_{REML} = 1.34, \hat{\beta}_0 = 0.491, \hat{\beta}_1 = 6.80$. The offsets used were -0.25 for the outlier excluded, and -1.10 for the outlier included. Note particularly the improvement of fit for the extreme values on the left and right of the frame.

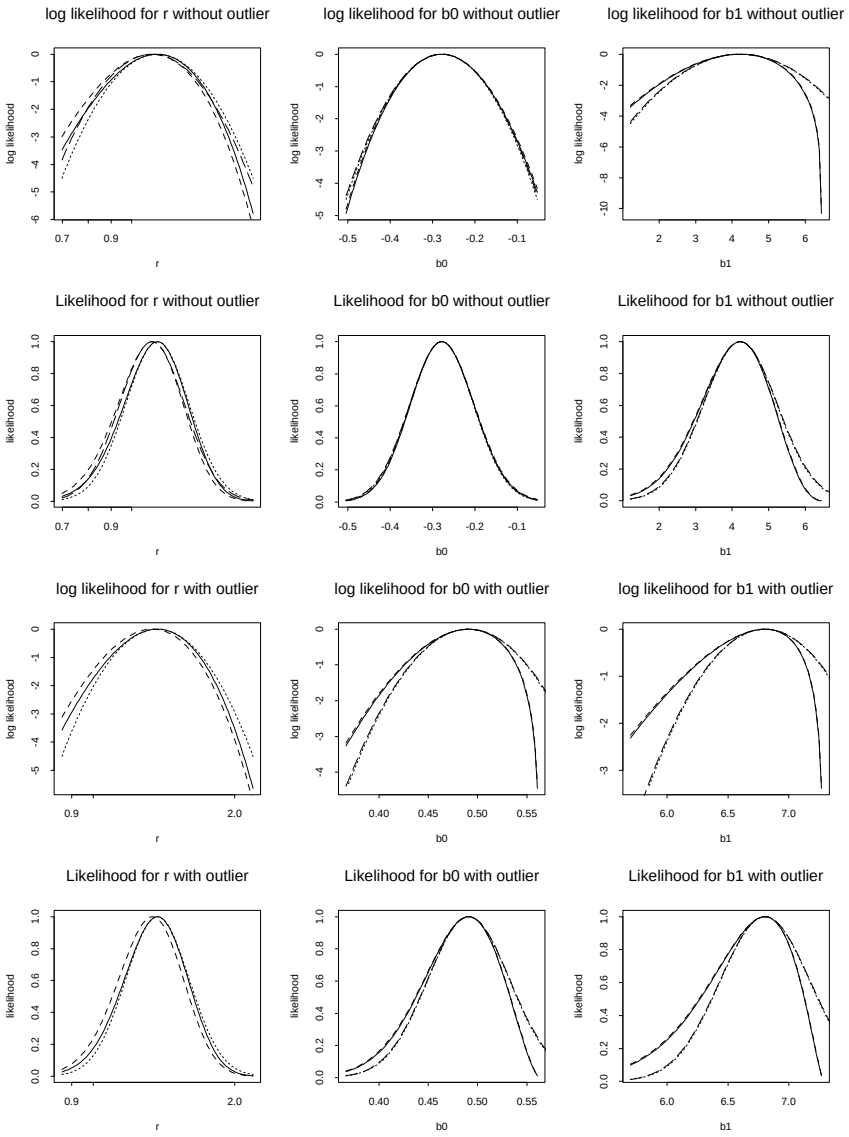


Figure 31: Conditional Likelihoods for the fitted parameters, s_0 fixed at 0.2, with offsets as described in (30). In each graph the other parameters are set to their estimates. The solid lines are the likelihoods, using the ML estimates of r, the short dashed lines are the quadratic approximations at the mode. The long dashed lines are the restricted likelihoods, using the REML estimates of r, and the dotted line is its corresponding quadratic approximation.

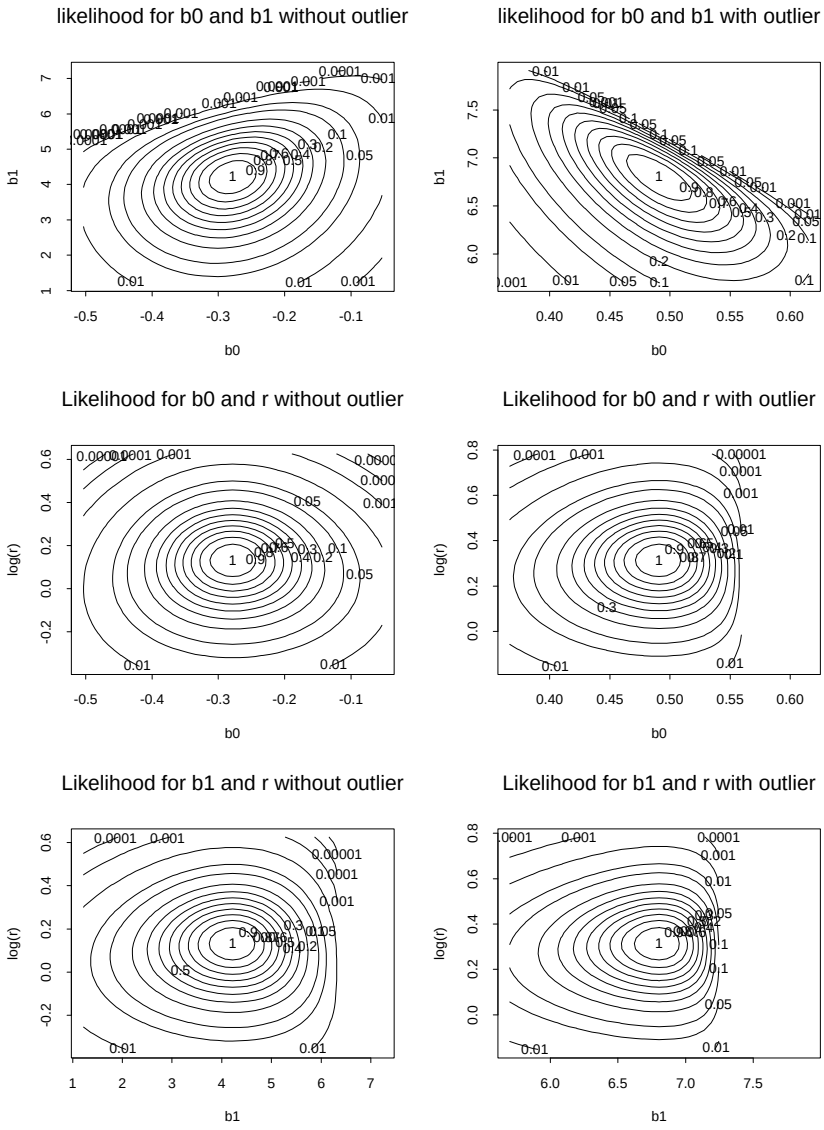


Figure 32: Bivariate Conditional Likelihoods for the fitted parameters, with s_0 fixed at 0.2, with offsets as described in (30).

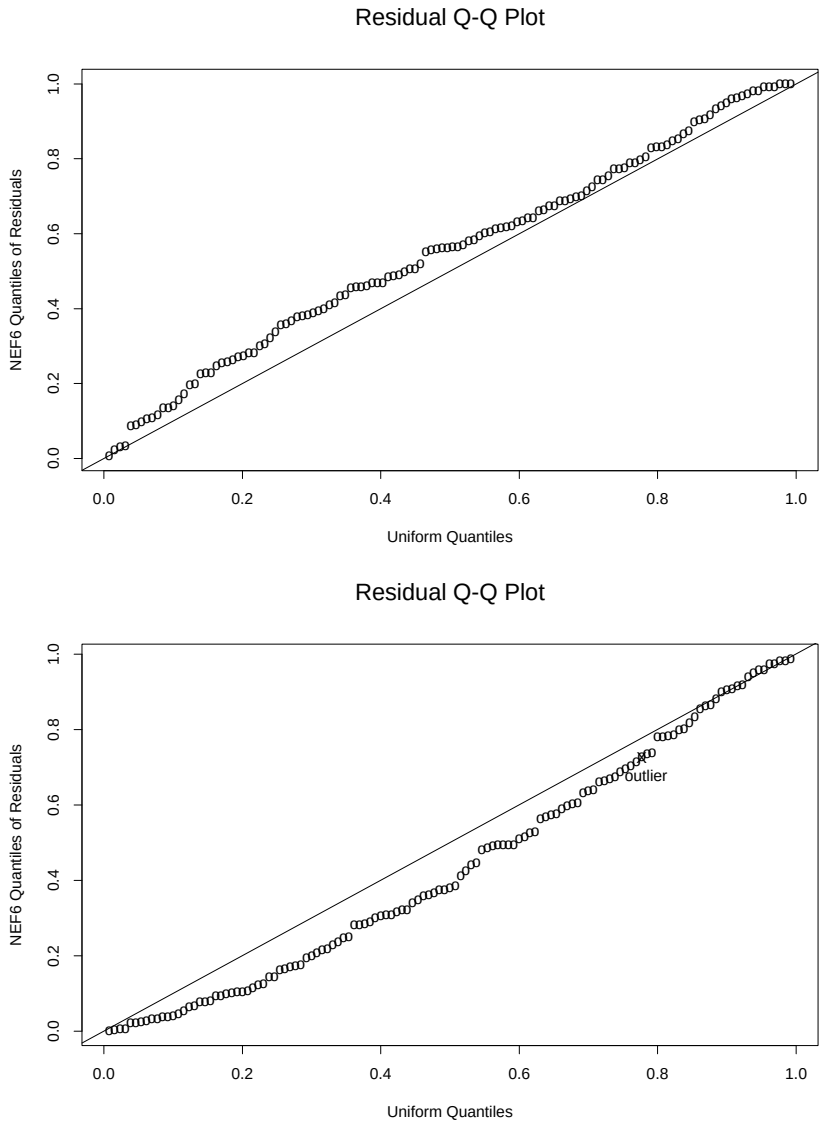


Figure 33: Residual diagnostic plots for the model fit with location and scale parameters in the response variable. Clearly the coverage of the error distribution is superior to the previous fits, and interestingly, the outlier is no longer an extreme point in the quantiles.

20 Conclusion and Discussion

We feel that this model can serve as a compromise between Gamma and Normal models, and can model conditionally and unconditionally skewed data. For data with quadratic variance structure, where the quadratic variance function has no real roots, this model may be appropriate. Although such datasets may be hard to find, the model is also of purely theoretical interest because it completes the set of NEFs with quadratic variance functions, and completes the set of elementary generalized linear models. Further research will definitely include fitting more types of datasets, and learning what we can from either the goodness or badness of the model fits.

This model can also be thought of as a missing component of quasi-likelihood analysis for quadratic variance functions with no real roots, i. e. asymptotically constant coefficients of variation whose variances do not disappear entirely at mean zero. The correct analysis for quasi-likelihood with such a variance function is the same as the analysis presented here.

Future work and extensions of these models might include investigating different link functions, such as the identity link; fitting of scale and location parameters for the response variable, addition of a Bayesian prior, and application of the model to more datasets. Other datasets may provide deeper insight into the strengths and shortcomings of the model (13.1), as did the datasets shown here.

Part III

The Skew-t Distribution: Properties and Computations

21 Pearson Families

The Pearson families arise as the set of all log densities whose first derivative is the quotient of a linear function and a quadratic, with respect to Lebesgue measure. There are basically only three such families, two of which in the general case are affine transformations of famous distributions. Special cases and limits of these three families form all of the other types named by Pearson. The three principal types are: Type I, or Beta, Type VI, or F, and the Type IV, or Skew-t. The Normal, Gamma (Type III), and Student-t are all special or limiting cases of the above three. Thus properties of all the Pearson families are well known, and software is readily available for computation, except in the case of the Type IV. Further properties of the Pearson families are available from many sources, e. g. Kendall and Stuart (1977) and will not be further discussed here.

22 The Pearson Type IV distribution

The type IV density is heavy-tailed, and is a skewed generalization of the symmetric Student-t distribution, which includes the Student-t as a special case. The Skew-t is not to be confused with the noncentral-t, which is generally obtained as the distribution of the usual t statistic for a normal population under the alternative hypothesis of nonzero mean. There is no such characterization for the Skew-t. The generalization occurs by multiplying the Student-t density by a factor of $\exp(r\delta \arctan(x))$. Thus the Skew-t family is an exponential family, but

not a natural exponential family; i. e. it is a nonlinear transformation of a natural exponential family. That natural exponential family is the family of densities proportional to

$$\cos^{r-2}(x)\exp(r\delta x), \tag{22.1}$$

with respect to Lebesgue measure, and with support on $(-\pi/2, \pi/2)$. This transformation (tangent) proves to be quite useful in calculating integrals of the density function.

23 Basic Properties

The density function of the Skew-t random variable, with respect to Lebesgue Measure, is (Nagahara, 1999)

$$p(x|\mu,\tau,\delta,r) = \frac{\Gamma(\frac{r}{2}-\frac{r\delta i}{2})\Gamma(\frac{r}{2}+\frac{r\delta i}{2})}{(\Gamma(r/2))^2}\frac{\tau^{r-1}}{B\left(\frac{r}{2}-\frac{1}{2},\frac{1}{2}\right)}\frac{\exp\{r\delta\arctan((x-\mu)/\tau)\}}{\{(x-\mu)^2+\tau^2\}^{r/2}}, \tag{23.1}$$

$$= \frac{\Gamma(\frac{r}{2}-\frac{r\delta i}{2})\Gamma(\frac{r}{2}+\frac{r\delta i}{2})\tau^{r-1}}{\Gamma(\frac{r}{2})\Gamma\left(\frac{r}{2}-\frac{1}{2}\right)\sqrt{\pi}}\frac{\exp\{r\delta\arctan((x-\mu)/\tau)\}}{\{(x-\mu)^2+\tau^2\}^{r/2}}, \tag{23.2}$$

$$-\infty < x < \infty$$

Here $i = \sqrt{-1}$ and the product of the two Gamma functions in the numerator is real because the second term is the complex conjugate of the first, so the product is the squared norm.

The above are proper densities on the entire real line for $r > 1$, $\delta \in (-\infty, \infty)$, $\mu \in (-\infty, \infty)$, and $\tau > 0$. For the remainder of this paper, unless otherwise indicated we will set the location parameter $\mu = 0$ and the scale parameter $\tau = 1$, since for other values of these parameters all calculations can be made with a standardizing transformation. With this, and by an application of the duplication formula for Gamma functions (Abramowitz and Stegun, 1970), we rewrite the density as

$$p(x|\delta,r) = \frac{B(\frac{r}{2} - \frac{r\delta i}{2}, \frac{r}{2} + \frac{r\delta i}{2})(r-1)2^{r-2}}{\pi} \frac{\exp\{r\delta \arctan(x)\}}{(1+x^2)^{r/2}}. \tag{23.3}$$

Also, the identity [3]

$$\arctan(x) = \frac{i}{2}\text{Log}\left(\frac{1-ix}{1+ix}\right) \tag{23.4}$$

allows us to write the density as

$$p(x|\delta,r) = \frac{B(\frac{r}{2} - \frac{r\delta i}{2}, \frac{r}{2} + \frac{r\delta i}{2})(r-1)2^{r-2}}{\pi} (1+ix)^{-\frac{r}{2} - i\frac{r\delta}{2}} (1-ix)^{-\frac{r}{2} + i\frac{r\delta}{2}}. \tag{23.5}$$

Note here that the parameterization does not follow exactly that of the traditional Student-t random variable; for the latter variable reparametrizes degrees of freedom via the transformation $\nu = r - 1$. Also, the Student-t random variable is rescaled, dividing the random variable which would result from the appropriate Skew-t density by the square root of its degrees of freedom. So when $\mu = 0$, $\tau = \nu$, $r = \nu + 1$, and $\delta = 0$, (23.1)-(23.5) become Student-t densities with ν degrees of freedom. Since it is desirable that the symmetric case match the traditional Student-t variable, we employ such a rescaling for the tables of the distribution in the appendix. We choose not to use this rescaling in the mathematical development to restrict unnecessary parameters from the mathematical development. However, the tables in the appendix do use this parametrization, also rescaling the distributions by $(1+\delta^2)^{-1}$ in the interest of having the $\delta = 0$ column match the Student-t random variable and including a $\delta = \infty$ column as well as a $\nu = \infty$ row. These limiting cases will be discussed further in Section 23.2

Many other parametrizations have been given for the Skew-t random variable. The degrees of freedom parametrization, shown above, is commonly used for the Student-t variable. If $r > 2$, we can parametrize in terms of the mean: $\mu_0 = r\delta/(r-2)$, and $m = r - 2$. This parametrization matches the form for Pearson conjugate Prior Distributions given in Morris (1983). Frequently the literature

[3]Here $\text{Log}(z)$ is the principal branch of the complex logarithm, i. e. $\text{Log}(z) = log(|z|) + i\arg(z)$, where $-\pi < \arg(z) \leq \pi$

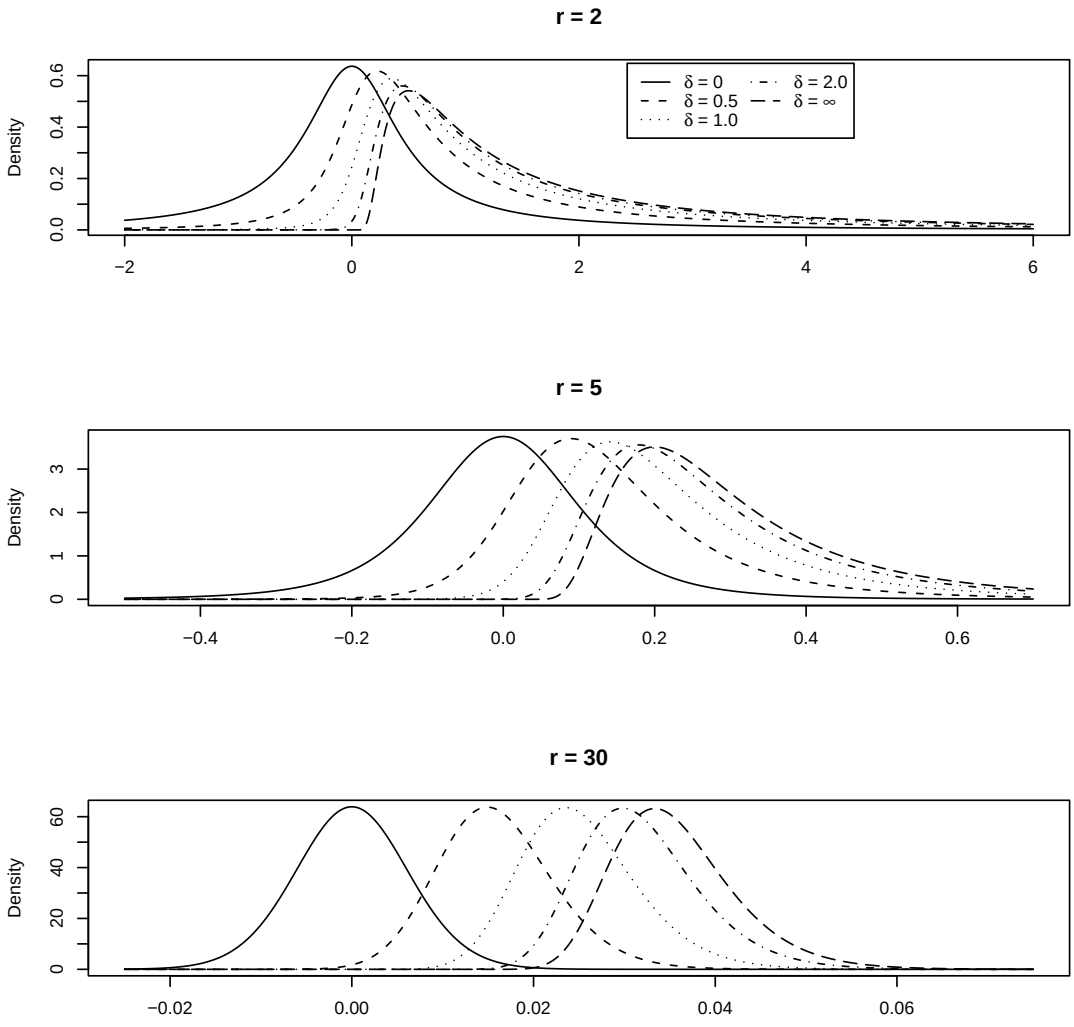


Figure 34: Various Skew-t densities, scaled to illustrate Reciprocal Gamma Limit Law

on Pearson families uses a single parameter for the product $r\delta = m\mu_0$; see, for example, Kendall and Stuart (1977).

23.1 Moments and Cumulants

Existence of moments and cumulants follows the pattern of the Student-t distribution; for $r \geq 2$ the mean exists, and every integer greater than unity guarantees the existence of an additional moment or cumulant. The first four moments are (Nagahara, 1999)

$$\mu = \mathrm{E}[X] = \frac{r\delta}{r-2}, \quad (r > 2) \tag{23.6}$$

$$Var[X] = \frac{1}{r-3}\left\{1 + \left(\frac{r\delta}{r-2}\right)^2\right\} = \frac{1+\mu^2}{r-3}, \quad (r > 3) \tag{23.7}$$

$$\mathrm{E}[(X - \mathrm{E}[X])^3] = \frac{4\mu}{r-4} Var[X], \quad (r > 4) \tag{23.8}$$

$$\mathrm{E}[(X - \mathrm{E}[X])^4] = \frac{3}{r-5}\left\{1 + \frac{r+4}{r-4}\mu^2\right\} Var[X], \quad (r > 5) \tag{23.9}$$

$$\gamma_3 \equiv \text{Skewness} = \frac{4\sqrt{r-3}}{r-4}\frac{\mu}{\sqrt{1+\mu^2}} \tag{23.10}$$

Thus δ, or equivalently μ determines the skewness of the Skew-t.

These and higher-order moments can be found from the recursive relation, derived from a general recursive relation for all Pearson families, Kendall and Stuart (1977) and Morris (1983)

$$n\mu\prime_{n-1} + r\delta\mu\prime_n + \{n + 2 - r\}\,\mu\prime_{n+1} = 0, \text{ if } r > n + 2. \tag{23.11}$$

23.2 Limit Laws

Similarly to the Student-t random variable, an appropriately scaled Skew-t will converge in distribution to a Normal random variable as $r \to \infty$, with δ fixed. As $\delta \to \infty$, with r fixed, the appropriately scaled version of a Skew-t will converge

in distribution to a reciprocal-Gamma variable. This can be seen as follows: As $\delta \to \infty$, the natural observation, with density proportional to $\cos^{r-2}(x)\exp(r\delta x)$, converges to $\pi/2 - G$, where G is $(r\delta)^{-1}$ times a Gamma$(r-1)$ random variable. This occurs because the probability mass is gathering near $\pi/2$, where $\cos(x) \doteq \pi/2 - x$. Near $x = \pi/2$, $1/(\pi/2 - x) \approx \tan(x)$, so the tangent of the natural observation, or the original Skew-t variable must converge to $1/(\pi/2-(\pi/2-G)) = 1/G$. Thus we can think of the Skew-t distributions as compromising continuously between the Student-t and the reciprocal Gamma laws, both well known as heavy-tailed distributions, the Student-t being symmetrical and the reciprocal Gamma representing the extreme case of skewness.

23.3 Characteristic Functions

23.3.1 The Characteristic Function of the Student t

There is little information in the literature about the characteristic function of the Student-t distribution. A few authors are aware of the specification of the characteristic function of the Student-t distribution in terms of a modified Bessel function of the second kind, $K_{\nu/2}(\cdot)$. Here, $\nu/2$ is the order of the modified Bessel function, and ν will be the degrees of freedom of the Student-t variable. This function is typically defined as one of two linearly independent solutions to the Modified Bessel Equation $z^2\frac{d^2w}{dz^2} + z\frac{dw}{dz} + (z^2 + (\nu/2)^2)w = 0$. The function $K_{\nu/2}(t)$ satisfies the integral relation (Abramowitz and Stegun, 1970)

$$K_{\nu/2}(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi}}\left(\frac{2}{t}\right)^{\nu/2}\int_0^\infty \frac{\cos(tx)}{(2+t^2)^{\frac{\nu+1}{2}}}dx, \tag{23.12}$$

$$\Re(\nu) \geq -1, t > 0. \tag{23.13}$$

Because the Student-t density is symmetric, we recognize the integral (23.13) as half of the Fourier transform $\phi_X(t)$ (for positive t) of the Student-t distribution, scaled by $\sqrt{\nu}$ and unnormalized. Making the necessary adjustments, we have

$$\phi_X(t) = \mathrm{E}(e^{itX}) = 2\frac{K_{\frac{\nu}{2}}\left(\frac{|t|}{\sqrt{\nu}}\right)\left(\frac{|t|}{2\sqrt{\nu}}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}, \tag{23.14}$$

where $K_{\nu/2}(\cdot)$ represents the modified Bessel function of the second kind, and of order $\nu/2$; [4]

The series expansions for $K_{n+\frac{1}{2}}(z)$ terminate for integer values of n, yielding expressions for $\phi(t)$ in terms of elementary functions for odd degrees of freedom, matching the expressions given for the characteristic function of the Student t with odd degrees of freedom, as given in (Johnson *et al.*, 1995).

23.3.2 The Characteristic Function of the Skew-t distribution

We can actually deduce the characteristic function of the Skew-t(r, δ) distribution. Since the characteristic function involves confluent hypergeometric functions, which are generally unfamiliar to the Statistics community, we will provide a brief introduction here. These functions are traditionally defined as solutions to Kummer's equation

$$x\frac{d^2y}{dx^2} + (b - x)\frac{dy}{dx} - ay = 0$$

. Plugging a general power series expansion for y into Kummer's equation yields a solution

$$\mathrm{M}(a; b; x) \equiv {}_1\mathrm{F}_1(a; b; x) \equiv 1 + \frac{a}{b}x + \frac{a(a+1)}{b(b+1)2!}x^2 + \frac{a(a+1)(a+2)}{b(b+1)(b+2)3!}x^3 + \ldots \tag{23.16}$$

(Abramowitz and Stegun, 1970) The before and after subscripts on ${}_1\mathrm{F}_1(a; b; x)$ indicate one coefficient each in the numerator and denominator, multiplicatively incremented with each term. This series is absolutely convergent for all a, b, and x, excluding nonpositive integer values for b.

[4] In (Abramowitz and Stegun, 1970), 26.7.2, the Characteristic Function of the Student-t is reported as

$$\frac{Y_{\frac{\nu}{2}}\left(\frac{|t|}{\sqrt{\nu}}\right)\left(\frac{|t|}{2\sqrt{\nu}}\right)^{\frac{\nu}{2}}}{\pi\Gamma\left(\frac{\nu}{2}\right)}, \tag{23.15}$$

which we believe to be erroneous.

A second form of solutions of Kummer's equation is given by a linear combination of two of the first form:

$$\mathrm{U}(a;b;x) \equiv \frac{\Gamma(1-b)}{\Gamma(1+a-b)}\mathrm{M}(a;b;x) + \frac{\Gamma(b-1)}{\Gamma(a)}x^{1-b}\mathrm{M}(1+a-b;2-b;x) \quad (23.17)$$

$$= \frac{\pi}{\sin(\pi b)}\left\{\frac{\mathrm{M}(a;b;x)}{\Gamma(1+a-b)\Gamma(b)} - \frac{x^{1-b}\mathrm{M}(1+a-b;2-b;x)}{\Gamma(a)\Gamma(2-b)}\right\} \quad (23.18)$$

$\mathrm{U}(a;b;x)$ remains bounded as $|x| \to \infty$, and is of particular interest here due to this fact.

Theorem 23.1. *The characteristic function of the Skew-t(r,δ) distribution is given by*

$$\phi(t) = \mathrm{E}(e^{itX}) = \frac{\Gamma\left(\frac{r}{2} - i\frac{r\delta}{2}\right)}{\Gamma(r-1)}e^{-t}(2t)^{r-1}\mathrm{U}\left(\frac{r}{2} - i\frac{r\delta}{2}; r; 2t\right), \quad (23.19)$$

for positive t, and the complex conjugate of (23.19) for negative t.

Proof. From Equation 5.4 of (Morris, 1983), any $h(\mu)$ with continuous derivative $h'(\mu)$ on Ω, which in the Skew-t case is $\Omega = (-\infty, \infty)$, and with

$$\lim_{\mu\to\pm\infty} h(\mu)V(\mu)g(\mu) = 0$$

satisfies

$$E(\mu - \mu_0)h(\mu) = \frac{1}{m}Eh'(\mu)V(\mu). \quad (23.20)$$

Here μ is the Skew-t variable, and $V(\mu)$ is the variance function, in our case $1+\mu^2$. If we let $h(\mu) = \exp(it\mu)$, which satisfies all of the sufficient conditions, (23.20) implies a differential equation for $\phi(t)$, the characteristic function of the Skew-t,

$$\phi''(t) - (r-2)\frac{\phi'(t)}{t} + \left(\frac{r\delta}{t}i - 1\right)\phi(t) = 0, \quad (23.21)$$

whose general solution is a confluent hypergeometric function with complex parameters, namely

$$\phi(t) = C_1e^{-t}(2t)^{r-1}\mathrm{M}(r/2 - r\delta i/2; r; 2t) + C_2e^{-t}(2t)^{r-1}\mathrm{U}(r/2 - r\delta i/2; r; 2t) \quad (23.22)$$

(Abramowitz and Stegun, 1970), 13.1.35

Setting $\phi(0) = 1$, we have $C_2 = \Gamma(r/2 - r\delta i/2)/\Gamma(r-1)$, and C_1 undetermined. However, $C_1 = 0$ since the function $e^{-t}\mathrm{M}(a;b;t)$ has a nonzero horizontal asymptote as $t \to \infty$ and hence the entire first term of the solution (23.22) is of exact order t^{r-1}, and characteristic functions attain their maximum moduli at zero. □

Corollary 23.2. *The Characteristic Function of the Student-t distribution is given by*

$$\phi_X(t) = \mathrm{E}(e^{itX}) = 2\frac{K_{\frac{\nu}{2}}\left(\frac{|t|}{\sqrt{\nu}}\right)\left(\frac{|t|}{2\sqrt{\nu}}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}, \tag{23.23}$$

.

Proof. From Equation 1.8.7 (Slater, 1960),

$$U(a;2a;t) = e^{t/2}\frac{t^{\frac{1}{2}-a}}{\sqrt{\pi}}K_{a-\frac{1}{2}}\left(\frac{t}{2}\right), \text{ where} \tag{23.24}$$

K_ν is the modified Bessel function of the second kind of degree ν, in the standard notation. This establishes the result through a different method than above, substituting $\delta = 0$ into Equation 23.19. □

The characteristic function (23.19) indeed yields the correct moment structure for the Skew-t random variable when differentiated and set to zero. Equation 2.1.28 from (Slater, 1960) is helpful in performing these differentiations:

$$\frac{d}{dx}\{x^{b-1}U[a;b;x]\} = (b-a-1)x^{b-2}U[a;b-1;x]. \tag{23.25}$$

Thus the solution (23.19) is consistent with the solution for the Student-t characteristic function.

24 Computational Algorithms

The CDF of the Skew-t distribution can be calculated explicitly in closed form for integer values of r by using the recursive relations

$$\begin{aligned}&(r^2\delta^2+(r-2)^2)\int \cos^{r-2}(\theta)\exp(r\delta\theta)d\theta\\&=(r\delta\cos(\theta)+(r-2)\sin(\theta))\,e^{r\delta\theta}\cos^{r-3}(\theta)+(r-2)(r-3)\int\cos^{r-4}(\theta)e^{r\delta\theta}d\theta\end{aligned}\tag{24.1}$$

the basic integrals for $r=3$ and $r=2$

$$\int\cos(\theta)\exp(r\delta\theta)d\theta=\frac{\exp(r\delta\theta)}{1+r^2\delta^2}(r\delta\cos(\theta)+\sin(\theta)),\tag{24.2}$$

and

$$\int\exp(r\delta\theta)d\theta=\frac{\exp(r\delta\theta)}{r\delta}.\tag{24.3}$$

Thus, to compute the CDF for an integer value of r, we first use (24.1) to reduce the value of the exponent in the cosine term under the integral to 0 or 1, and then use (24.2) or (24.3) to evaluate the remaining term. Since the recursive use of (24.1) will produce a terminating series, we check that the recursion will not introduce unnecessary numerical instability. In fact, we can write the formula for k applications of the recursion in a single formula as follows:

$$\begin{aligned}&\int\cos^{r-2}(\theta)\exp(r\delta\theta)d\theta\\&=e^{r\delta\theta}\frac{(\cos^{r-2}\theta)(\mu_0+\tan\theta)}{(r-2)(1+\mu_0^2)}+e^{r\delta\theta}\frac{(r-3)(\cos^{r-4}\theta)(\mu_1+\tan\theta)}{(r-2)(r-4)(1+\mu_0^2)(1+\mu_1^2)}\\&+e^{r\delta\theta}\frac{(r-3)(r-5)(\cos^{r-6}\theta)(\mu_2+\tan\theta)}{(r-2)(r-4)(r-6)(1+\mu_0^2)(1+\mu_1^2)(1+\mu_2^2)}+\cdots\\&+\frac{(r-3)(r-5)\ldots(r-2k-1)}{(r-2)(r-4)\ldots(r-2k)(1+\mu_0^2)(1+\mu_1^2)\ldots(1+\mu_{k-1}^2)}\times\\&\int\cos^{r-2k}(\theta)\exp(r\delta\theta)d\theta,\end{aligned}\tag{24.4}$$

where $\mu_j=r\delta/(r-2-2j)$ is the mean of a Skew-t$(r-2-2j,r\delta/(r-2-2j))$ variable. From (24.4) we see that all terms are a multiple of order one of the

previous term, and no instability will be introduced in the form of small numbers in the denominator. The only potential for numerical instability is in the evaluation of the exponential and cosine terms, which should be evaluated on the logarithmic scale and adding the logarithm of the normalizing constant before exponentiating to avoid underflow or overflow.

Thus Equation 24.4 solves the problem of evaluating the CDF for integer r. However, we still need an algorithm for non-integer values of r across a range of two, since values of r outside of this range can be moved to this range through (24.1) or its inverse. It is actually advantageous to use a range of values greater than two, since for values less than two the integrand is infinite at $\pm\pi/2$. Otherwise it is advantageous not to have the probability mass too concentrated, as will occur for larger values of r, so we choose to apply (24.1) recursively to adjust the value of r under the integral so that $2 \leq r < 4$. The rest of this section will be devoted to the problem of calculating the CDF for non-integer values of r.

Most computational routines are derived from asymptotic series and/or rational function expansions for functions. When the function is indexed by zero or one parameters, this approach may lead to fast and highly accurate calculations. However, in this case, the family of distributions is indexed by two fundamental parameters, because while the scale and location parameters can be transformed out, the degrees of freedom (r) and exponential parameters (δ) remain. Thus, when generating a series approximation, the terms become increasingly difficult to compute, because more and more derivatives of products of exponential, polynomial, and Gamma functions must be taken. Most notably, the coefficients of the series itself must be recalculated every time. The accuracy of the series approximation will depend on (1) the abscissa at which the CDF is being calculated, (2) the values of the parameters, and (3) the accuracy of the routines used to compute polygamma functions. These computational problems might feasibly be overcome, but the problem of error control also arises, and errors for Taylor series are only known to an order of magnitude. Calculations from the same series at different

abscissas could result in greatly different magnitudes of error in the calculation.

Because of these problems with series, we develop a direct quadrature method to provide uniformity of calculation methods, the easy management of quadrature error to within arbitrary tolerance, and the possibilities of using vectorization to calculate distributions or quantiles of vectors almost as quickly as a single value.

Since evaluation of the CDF requires integrating over an infinite range, we transform the variable of integration by $\theta = \arctan(Y)$. Thus the range of the transformed variable θ is $(-\pi/2, \pi/2)$ and its density is

$$p(\theta|\delta, r) = \frac{B(\frac{r}{2} - \frac{r\delta i}{2}, \frac{r}{2} + \frac{r\delta i}{2})(r-1)2^{r-2}}{\pi} \cos^{r-2}(\theta) \exp(r\delta\theta) \tag{24.5}$$

The quadrature problem amounts to evaluating

$$\int_{-\frac{\pi}{2}}^{y} \cos^{r-2}(\theta) \exp(r\delta\theta)\, d\theta. \tag{24.6}$$

The arctangent transformation for calculating cumulative probabilities works well when the upper limit of the integral being evaluated is not too close to one of the boundaries $\pm\pi/2$. Accuracy may be lost when this happens, since the tangent of $\pi/2 \pm \theta$ is asymptotically equal to $1/(\pi/2 \pm \theta)$ as $\theta \to \mp\pi/2$. However, we find that only in extreme cases does this happen, and the method presented here is adequate for all but the most extreme cases.

Note that for $r < 2$ this integrand is singular at $\theta = \pm\pi/2$. We circumvent this problem with the recurrence relation

$$\begin{aligned}
&(r-1) \int_{-\frac{\pi}{2}}^{\theta} \cos^{r-2}(\phi) \exp(r\delta\phi)\, d\phi \\
&= -(\delta \cos(\theta) + \sin(\theta)) \mathrm{e}^{r\delta\theta} \cos^{r}(\theta) + r(\delta^2 + 1) \int_{-\frac{\pi}{2}}^{\theta} \cos^{r}(\phi) \exp(r\delta\phi)\, d\phi.
\end{aligned} \tag{24.7}$$

Although (24.7) might be used to reduce the value of r, here we increase r to ensure a finite integrand across the range of integration.

24.1 Evaluating the CDF for noninteger r

We use an adaptive quadrature scheme, calculating the midpoint and trapezoid rules for each subinterval, and refining the set of quadrature points wherever they disagree. This method can fail if the three points in question are collinear, which will not normally happen since the function is relatively smooth, with only two points of inflection. The integrand is not symmetric about the points of inflection, so it is almost never the case that integer divisions of the range of integration result in collinear points along the functions. We start with some arbitrary number of points for the quadrature, and refine as necessary until the desired precision times the fraction of the range each interval occupies is obtained within each interval. This error calculation is conservative; since the Simpson's rule is returned the actual error is considerably smaller than the tolerance. If a vector of points is sent to the function, the first subdivision is at the points in the (sorted) vector; thus no calculations are wasted, and the entire vector of distribution values can be calculated almost as fast as a single value. Note also that if the value whose CDF is to be evaluated is greater than zero, the shorter upper tail can be evaluated, and subtracted from unity to yield the answer, perhaps more efficiently than the "forward" calculation.

24.2 Evaluating the Quantile function

Here we take advantage of the CDF machinery already in place. We start with bounds $(-\pi/2, \pi/2)$ for the values, and bisect the interval. If the half interval's distribution function is greater than the desired quantile, the new bounds are $(-\pi/2, 0)$. Otherwise the new bounds are $(0, \pi/2)$, and the value of the quadrature is subtracted from the desired quantile value to obtain a new target value, and the code then attempts to find a value y such that the integral from 0 to y of the integrand is this new target. The above procedure is iterated until convergence. Thus the interval over which quadratures must be evaluated always decreases by

half, and the total of the lengths intervals over which quadratures must be evaluated before convergence is always equal to π. If a vector of quantiles is requested, the procedure first sorts the values, and for each successive value reduces the input by the previous input value, and starts the quadrature at the previous quantile. Thus redundancy in the calculations is minimized.

25 Obtaining Random Variates

For the purpose of obtaining random variates, we seek an accept-reject algorithm with several features: uniformly bounded rejection rates for all parameter values of the distribution; easily calculable parameter values for the envelope densities, and thus easily obtainable variates for the envelope densities; and easily obtainable rejection ratios. To these ends, we suggest several different families of envelope densities, each with its own area in the parameter space with superior performance. We first describe the various methods, and then compare them.

It should first be noted that there are some boundary cases where random variates can be obtained through widely available methods. Firstly, if $\delta = 0$, the density is a scaled t density with degrees of freedom $r - 1$, so that we may obtain a variate by sampling a t_{r-1} variate and dividing by $\sqrt{r-1}$. Secondly, in the case where $r = 2$, the natural observation $\Theta = \arctan(Y)$ has density proportional to $\exp(r\delta\theta)$, and thus can be sampled directly by transformation of a uniform variate.

25.1 Normal envelopes

One theoretically interesting property of the normal distribution is that exponentially tilting the density, i. e. adding a linear term to the log density simply results in another normal density with a location change. We can take advantage of this feature and construct a normal envelope for the density of the natural observation of the skew-t density with no exponential term, i. e. a density proportional to $\cos^{r-2}(\theta)d\theta$. Since the log ratio of the densities as a function of θ is a constant plus

$(r-2)\log(\cos(\theta))+\theta^2/(2\sigma^2)$, the ratio of the cosine density to the normal will have its maximum at $\theta = 0$ if $\sigma^2 > 1/(r-2)$. Hence we restrict our search to values of σ^2 less than $1/(r-2)$. The optimal value is indeed in this range, and because of the symmetry and smoothness of the ratio surface, viewed as a function of σ^2 and θ, must occur at a point where both partial derivatives are zero, namely where $\theta = \pm\sigma$ and $\theta\tan(\theta) = 1/(r-2)$. Because the function $\theta/\tan(\theta)$, which is concave on $(-\pi/2, \pi/2)$, with its maximum at $(0, 1)$ and contains the points $(\pm\pi/2, 0)$, is well-approximated by a quadratic, some approximations are possible which are close to optimal but can be evaluated without optimization. Setting the quadratic $1-(4/\pi^2)\theta^2$ equal to $(r-2)\theta^2$ yields the approximation $\sigma^2 = 1/(r-(2-4/\pi^2)$.

Normal envelopes can also be generated via other methods. Either the mean or the mode of the exponentially tilted cosine density can be matched to the normal distribution, and the variance of the envelope can be found by optimization or by matching moments. We have found the moment-matching method to give significantly suboptimal performance, so we will not consider it here. However, there is something to be gained by allowing the mean/mode of the normal envelope density to adjust to the cosine density, as the performance of the above central envelope method deteriorates for even moderate values of $r\delta$.

We will, however, consider matching the normal envelope to the mode of the exponentially tilted cosine density, i. e. $\mu = \arctan(r\delta/(r-2))$. When the log ratio is viewed as a function of θ, this has the effect of adding a linear term in θ to the log ratio, which has the effect of raising and elongating one of the ridges, which were formed by the double maximum in the symmetric case when $\sigma^2 < 1/(r-2)$ on the surface of the ratio of densities as a function of θ and σ. As the coefficient on this linear term grows, the minimum of the maximum ratio eventually no longer occurs at a point where the partial derivatives are both zero; although the partial derivative with respect to θ will still be zero, the partials at the two (equal) maxima with respect to σ will be of opposite signs. Thus the optimal value will occur at a point where there are two equal maxima in the $\theta - \sigma^2$ surface. Figure 25.1 shows

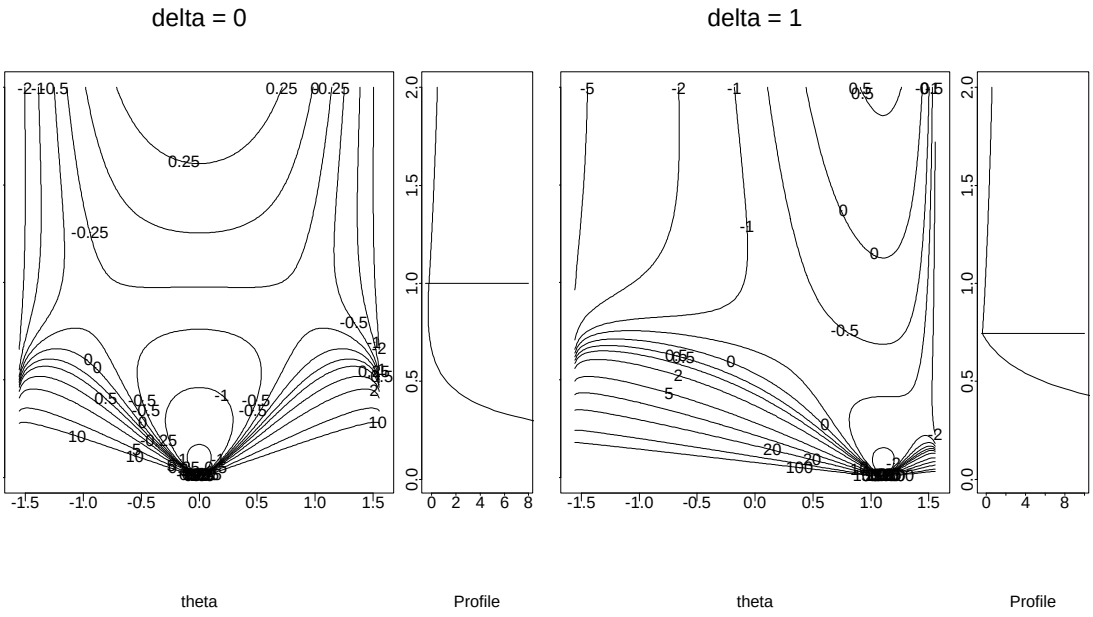


Figure 35: Log ratios of mode-matched normal densities to the densities (22.1). The vertical axes index the standard deviation of the normal densities, σ. Shown on the side of each surface plot is the profile, or maximum of the cross-section fixing the value of σ. For larger δ the minimum profile occurs at a "fold", i. e. where Equation 25.2 is satisfied.

the effect of increasing the value of δ on the log ratio surface. Also shown is the profile of the height of the surface, i. e. the maximum height over θ for fixed σ. We can see from this graphic that for small values of δ the minimum ratio profile occurs at a smooth point in the profile, whereas for larger δ the minimum occurs at a "corner" of the profile, i. e. the point on the surface corresponding to the profile jumps at the minimum. Examining the formula for the log ratio in this case,

$$C(r, \delta) + \log(\sigma) + \frac{(\theta - \mu)^2}{2\sigma^2} + (r - 2)\log(\cos(\theta)) + r\delta\theta, \tag{25.1}$$

we see that there is only one place where the above condition can happen: where the linear terms cancel, in which case we have $\theta\mu/\sigma^2 = r\delta\theta$, or

$$\sigma^2 = \frac{\mu}{r\delta} = \frac{\arctan\left(\frac{r\delta}{r-2}\right)}{r\delta}. \tag{25.2}$$

Thus in the case where the exponential tilt $r\delta$ is large enough, the optimal normal envelope matching the mode of the distribution is much simpler to calculate than in the symmetric case. However, we find that the performance gets worse as $r\delta$ increases in absolute value. Nevertheless, this degradation of performance is of little consequence, because in situations with larger exponential tilts we will use the gamma envelope. Because of the ease of choosing the parameters of the normal density, and its good performance, we will recommend the envelope $N(\mu, \sigma^2)$, where $\mu = \arctan(r\delta/(r-2))$, and $\sigma^2 = \mu/r\delta$. It remains to choose the normalizing constant, which is found by optimizing the ratio of densities. Note that the problem above is equivalent to finding the maximum ratio for $\delta = 0$, with the particular value of σ determined above. Taking the log of the ratio and differentiating yields the optimality condition $(r-2)\tan(\theta) = \theta/\sigma^2$. We find the approximation

$$\tan(\theta) \doteq \frac{\frac{\pi^2}{4}(\theta - C\theta^3)}{\frac{\pi^2}{4} - \theta^2} \tag{25.3}$$

useful in approximating the optimal value in closed form. Using $C = 1/13.3$ produced values which were always greater than the true maximum ratios when multiplied by 1.1, for $r < 10,000$ and $\delta < 40$ when the resulting quadratic equation was solved for θ When the desired region of coverage for this approximation is reduced, the constant by which the answers must be multiplied to always exceed the true values can be much closer to one, or the answer can be passed through a couple of Newton iterations to obtain the desired precision.

Keep rations for the algorithm, using the approximation, are shown in Figure 36. It is clear in this figure that the normal algorithm's performance is best for less skewed distributions in the family.

The normal envelopes are obviously only useful in the cases where the density is log-concave, i. e. where $r > 2$. The case where $r = 2$, as discussed above, has

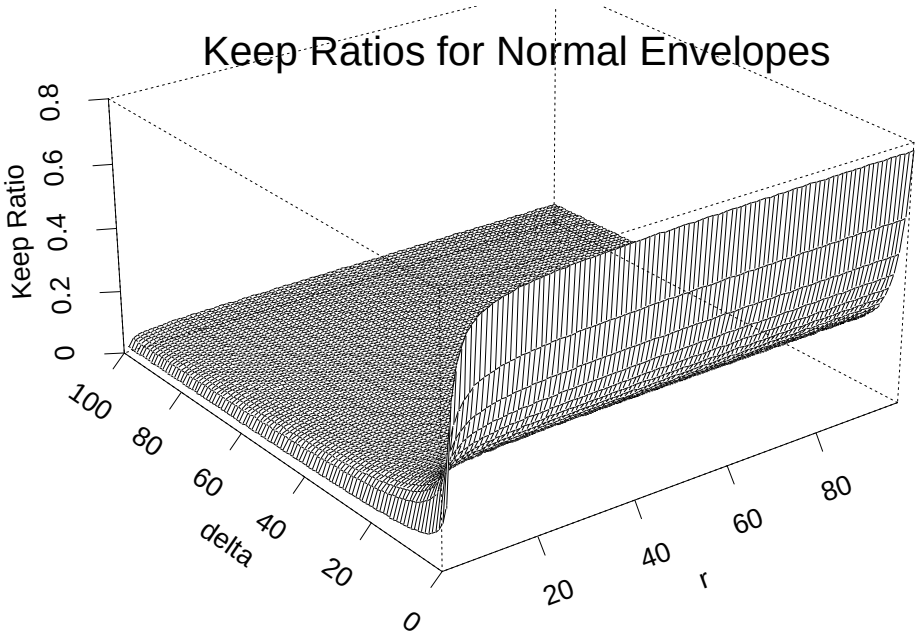


Figure 36: Keep ratios for Normal envelopes. The best keep ratios here are for less skewed densities.

exponential density and can be handled analytically. However, we still require a method for $1 < r < 2$ and for large exponential tilts. These cases will be taken care of by the beta and gamma densities.

25.2 Beta envelopes

Beta densities are a natural choice for an envelope function since they have bounded support on both sides and generally match the shape of the cosine-exponential densities. The ratio of the two densities is bounded as long as the two parameters of the beta distribution are both less than $r - 1$. Otherwise the ratio tends toward infinity one or more of the support boundaries. Beta densities with moments (or modes) matched to the cosine densities often have parameters which exceed this threshold, so we have found the best approach is to set one parameter to its maximum value, $r - 1$, and optimize on the other parameter. For zero exponential tilt the optimal value for the other parameter is also its maximum value, so that the density is symmetrical. As the exponential tilt increases, the optimal value lies somewhere between the value which would match the mode or mean and the maximum possible value $r - 1$. Note here that matching on the "mode" is impossible when $r < 2$, since there is no mode, but matching on the point of minimum density requires values of the beta parameters for which no density exists. For values of r less than two, we have found good performance choosing the optimal θ as an equal-weighted mean when r is close to 1, and more heavily weighted, e. g. 6/7, toward the mean-matched value when r is closer to two. We have implemented a linear interpolation to form a weighted mean for intermediate values of r, with maximum ratios less than 10 for values of δ extending to 2000. Although we do not here present a method which is uniformly bounded in the parameter space, it is highly unusual to require a deviate from a distribution which is more extreme than the ranges presented here, and as $\delta \to \infty$ the density approaches a linearly transformed gamma density, which could be used as an approximate density in

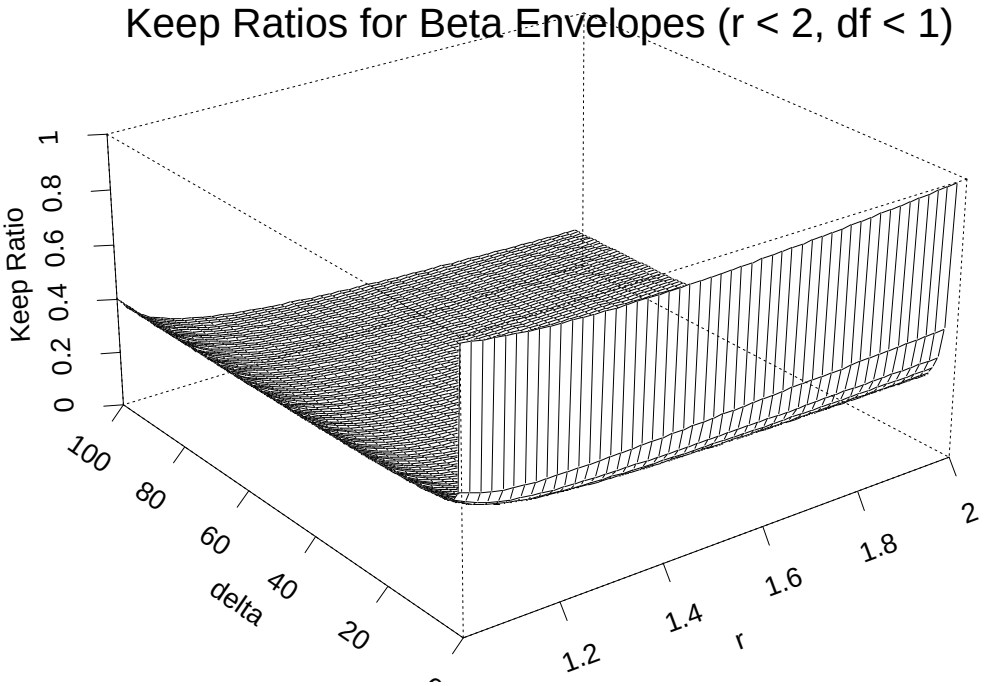


Figure 37: Keep ratios for Beta envelopes, with $1 < r < 2$. These densities are analogous to t distributions with less than one degree of freedom. The best keep ratios here are again for less skewed densities.

such cases. Note that in the $r < 2$ case the gamma density cannot be used as an envelope, since the density does technically also approach infinity at $\theta = -\pi/2$, but because of the exponential tilting this pole at $-\pi/2$ is for all purposes equivalent to a delta function with very small overall mass. Again it should be stressed that the $1 < r < 2$ case is similar to the Student-t with less than one degree of freedom; rarely would it be called for in a simulation. Keep ratios for beta envelopes are shown in Figure 37, and can be seen to perform better for less skewed densities.

25.3 Gamma Envelopes

The function $\cos(\theta)$ is asymptotically equal to $\pm(\pi/2 - \theta)$ at $\pm\pi/2$, so when the cosine density includes a steep exponential term, it is natural to use a gamma density, with appropriate location and scale changes, as an envelope. Without

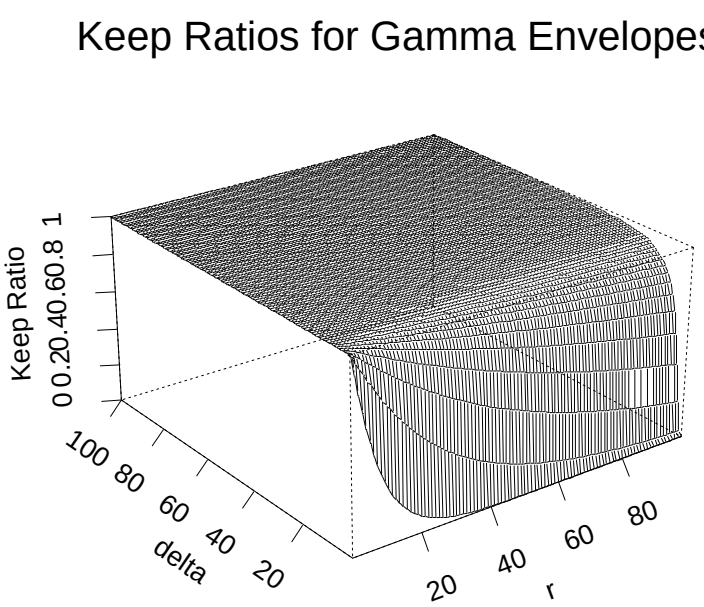


Figure 38: Keep ratios for Gamma envelopes. The best keep ratios here are for more skewed densities.

loss of generality we will consider positive coefficients in the envelope. Then the envelope density is that of $\pi/2$ minus a gamma random variable, and the maximum ratio will always occur at the right-hand edge, and is easily computable as the ratio of the normalizing constants for the two densities. These gamma envelopes work extremely well as δ increases, which is the area for which the other envelopes do not work well. Unfortunately they do not work for cases in which $r \leq 1$. Keep ratios are shown in Figure 38.

25.4 Piecewise Exponential Envelopes

In the case where $r > 2$, the density is log-concave; thus an envelope can be constructed using piecewise exponential curves. Several points on the log density and their slopes and intersections must be calculated, and then the sample from the envelope can be computed analytically. This method can provide the tightest en-

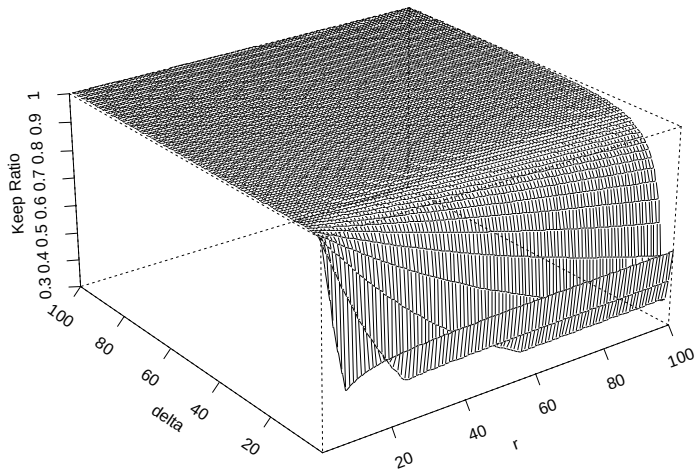


Figure 39: Keep ratios for the combined Normal/Gamma method, for $r > 2$. The worst case is approximately 0.4, quite acceptable for a rejection sampling scheme.

velope if enough points are chosen, but there is significant overhead in terms of the calculation. Therefore we present the method here only in the interest of completeness; the other methods, when patched together, provide enough flexibility to obtain deviates rather efficiently.

25.5 Choosing Between the Methods

For $r > 2$ we find that the Normal envelopes outperform the Beta envelopes in the regions for which the Gamma envelopes perform poorly. The Gamma envelopes are obviously preferable for large δ. The line along which the transition from normal to gamma should take place can be approximated by $\delta = \log(0.3 + r/2.2)$, i. e. using the normal envelope if $2.1 \exp(\delta) < r$, and the gamma envelope otherwise. See Figure 39 for a perspective plot of the keep ratios for the combined Normal/Gamma method for $r > 2$.

26 Applications

The Skew-t distribution is theoretically interesting in that it completes the set of Pearson families. Thus all application which use Pearson systems could benefit from the availability of tools to compute the Skew-t distributions. This includes various schemes to approximate distributions based on moments. For example, the Pearson-IV distribution has been used widely to approximate point defects in ion implantations for silicon diode and integrated circuit technologies (Hobler and Selberherr, 1988). The Skew-t distribution is the only available distribution which captures the appropriate skewness and kurtosis for the quantities being modeled.

The Skew-t distribution is the conjugate prior for the mean of the NEF_6 distribution, and thus is useful for modeling with the NEF_6 distribution, an absolutely continuous distribution which ranges from bell-shaped to extremely skewed. The mean of the distribution indexes the skewness, and has a Skew-t conjugate prior. Thus the Skew-t distribution may prove extremely useful in modeling data with outliers, with the possibility of treating the entire data set, including the outliers, with one distribution.

The Skew-t is also useful for approximating other densities, as it may be used to match four moments, along with the Beta and F distributions. Pearson's system was originally designed to match moments for datasets, and such systems of analysis were generally rejected by the statistics community, since third and especially fourth moment estimates from samples are particularly vulnerable to sample variation, and thus are poorly estimated from samples. We propose the use of the Pearson families to approximate densities whose theoretical, i. e. exact moments are known or calculable, as is the case for some posterior distributions in Bayesian analyses; see for example, (Morris, 1988). In that paper, Morris uses other Pearson families, specifically the beta family, for this purpose, but the computations described there can easily be extended to use the Skew-t distribution, which may be more appropriate for certain cases.

Skew-t distributions could also be used for robust analyses, i. e. long-tailed and/or skewed data could be modeled, and estimates for quantities of interest could be produced which are robust to skewness as well as kurtosis, a standard reason for using the Student-t distribution to model quantities which might otherwise be modeled with the Normal distribution. Thus the Skew-t could provide added robustness for analyses where the Student-t distribution is used in place of Gaussian or other thinner-tailed distributions.

27 Summary

We feel that given the new computational resources available to statisticians, the Skew-t distribution is worth investigating. The primary reason more theory and models have not been developed was always the difficulty involved in making calculations, but with modern computers, this difficulty is easily surmounted. This paper summarizes what is known about this distribution and gives algorithms for the basic computational tasks of calculating the CDF, quantiles, and drawing random variates. Many more computational routines are possible given parametric models which use the Skew-t distribution, and will undoubtedly be developed over time.

We have included in the following pages tables for various quantiles of the distribution. These are meant to complement the usual tables, although in practical applications, the computer functions for the quantiles should be used, especially because space considerations in this paper limit the parameter values, quantiles, and decimal accuracy of the quantiles shown.

28 Appendix: Tables of the Skew-t Distribution

28.1 Use of the Tables

To find a quantile of a Skew-t(r, δ) distribution, as parametrized in (23.3):

1. Calculate $\nu = r - 1$.

2. Locate the quantile on the table.

3. Multiply the entry on the table by $(1+\delta^2)/\nu$ and add δ to the result.

Quantiles for Skew-t Distributions

Note: Skew-t Distributions have been shifted to the left δ and rescaled by $\sqrt{\nu/(1+\delta^2)}$

ν:	δ: / α:	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	∞
1	0.01	-31.821	-4.860	-1.241	-0.811	-0.695	-0.646	-0.621	-0.606	-0.596	-0.590	-0.585	-0.566
	0.025	-12.706	-2.345	-0.879	-0.634	-0.555	-0.520	-0.501	-0.489	-0.482	-0.477	-0.473	-0.458
	0.05	-6.314	-1.416	-0.643	-0.474	-0.413	-0.385	-0.369	-0.359	-0.353	-0.349	-0.346	-0.332
	0.5	0.000	0.727	1.257	1.526	1.661	1.733	1.776	1.804	1.822	1.835	1.844	1.885
	0.95	6.314	17.78	26.91	31.61	33.98	35.27	36.04	36.53	36.86	37.09	37.25	37.99
	0.975	12.71	36.49	55.25	64.90	69.77	72.42	73.99	75.00	75.67	76.14	76.48	78.00
	0.99	31.82	92.59	140.3	164.8	177.1	183.8	187.8	190.4	192.1	193.3	194.2	198.0
2	0.01	-6.965	-2.423	-1.293	-1.011	-0.909	-0.861	-0.835	-0.819	-0.809	-0.802	-0.797	-0.775
	0.025	-4.303	-1.710	-1.023	-0.830	-0.756	-0.720	-0.699	-0.687	-0.679	-0.674	-0.670	-0.653
	0.05	-2.920	-1.275	-0.807	-0.662	-0.604	-0.575	-0.558	-0.548	-0.542	-0.537	-0.534	-0.520
	0.5	0.000	0.466	0.764	0.912	0.987	1.028	1.052	1.067	1.078	1.085	1.090	1.114
	0.95	2.920	5.798	7.915	9.008	9.566	9.874	10.06	10.17	10.25	10.31	10.35	10.53
	0.975	4.303	8.808	12.08	13.77	14.63	15.10	15.38	15.56	15.68	15.77	15.83	16.10
	0.99	6.965	14.74	20.33	23.19	24.64	25.45	25.93	26.23	26.43	26.58	26.68	27.15
3	0.01	-4.541	-2.122	-1.365	-1.131	-1.038	-0.993	-0.967	-0.952	-0.942	-0.935	-0.930	-0.908
	0.025	-3.182	-1.625	-1.113	-0.944	-0.874	-0.839	-0.820	-0.808	-0.800	-0.794	-0.790	-0.773
	0.05	-2.353	-1.272	-0.899	-0.769	-0.714	-0.686	-0.670	-0.660	-0.653	-0.649	-0.646	-0.632
	0.5	0.000	0.369	0.595	0.707	0.763	0.794	0.812	0.824	0.832	0.837	0.841	0.859
	0.95	2.353	4.039	5.247	5.870	6.190	6.366	6.471	6.539	6.584	6.616	6.639	6.741
	0.975	3.182	5.615	7.341	8.229	8.684	8.934	9.084	9.179	9.243	9.289	9.321	9.466
	0.99	4.541	8.296	10.93	12.28	12.97	13.35	13.58	13.72	13.82	13.89	13.940	14.16
4	0.01	-3.747	-2.030	-1.423	-1.216	-1.130	-1.087	-1.062	-1.047	-1.038	-1.031	-1.026	-1.004
	0.025	-2.776	-1.606	-1.176	-1.023	-0.957	-0.924	-0.905	-0.893	-0.885	-0.880	-0.876	-0.859
	0.05	-2.132	-1.285	-0.962	-0.842	-0.790	-0.763	-0.747	-0.738	-0.732	-0.727	-0.724	-0.710
	0.5	0.000	0.314	0.504	0.597	0.644	0.669	0.684	0.694	0.701	0.705	0.709	0.723
	0.95	2.132	3.370	4.240	4.690	4.921	5.048	5.124	5.173	5.205	5.228	5.245	5.319
	0.975	2.776	4.494	5.695	6.313	6.629	6.804	6.908	6.975	7.020	7.051	7.074	7.175
	0.99	3.747	6.265	8.010	8.903	9.360	9.612	9.762	9.858	9.923	9.968	10.00	10.15
5	0.01	-3.365	-1.993	-1.471	-1.282	-1.201	-1.159	-1.136	-1.122	-1.112	-1.105	-1.101	-1.080
	0.025	-2.571	-1.605	-1.225	-1.083	-1.020	-0.989	-0.970	-0.959	-0.951	-0.946	-0.942	-0.926
	0.05	-2.015	-1.299	-1.009	-0.897	-0.847	-0.821	-0.806	-0.797	-0.791	-0.787	-0.784	-0.770
	0.5	0.000	0.278	0.444	0.526	0.567	0.589	0.602	0.611	0.617	0.621	0.623	0.636
	0.95	2.015	3.017	3.713	4.071	4.256	4.357	4.418	4.457	4.483	4.501	4.515	4.574
	0.975	2.571	3.930	4.870	5.353	5.600	5.737	5.819	5.871	5.906	5.931	5.949	6.028
	0.99	3.365	5.299	6.626	7.306	7.653	7.845	7.960	8.033	8.082	8.116	8.142	8.253

Quantiles for Skew-t Distributions

Note: Skew-t Distributions have been shifted to the left δ and rescaled by $\sqrt{\nu/(1+\delta^2)}$

	δ:	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	∞
ν:	α:												
6	0.01	-3.143	-1.978	-1.510	-1.335	-1.258	-1.218	-1.196	-1.182	-1.173	-1.166	-1.162	-1.141
	0.025	-2.447	-1.609	-1.264	-1.131	-1.071	-1.040	-1.023	-1.012	-1.005	-1.000	-0.996	-0.980
	0.05	-1.943	-1.312	-1.046	-0.940	-0.893	-0.868	-0.854	-0.845	-0.839	-0.835	-0.832	-0.819
	0.5	0.000	0.252	0.402	0.475	0.512	0.532	0.544	0.552	0.557	0.560	0.563	0.574
	0.95	1.943	2.798	3.386	3.689	3.844	3.930	3.981	4.014	4.036	4.051	4.063	4.112
	0.975	2.447	3.590	4.374	4.776	4.982	5.095	5.164	5.207	5.236	5.257	5.272	5.338
	0.99	3.143	4.738	5.824	6.380	6.664	6.821	6.915	6.975	7.015	7.043	7.064	7.155
7	0.01	-2.998	-1.972	-1.544	-1.379	-1.305	-1.268	-1.246	-1.232	-1.223	-1.217	-1.213	-1.193
	0.025	-2.365	-1.616	-1.296	-1.170	-1.113	-1.084	-1.067	-1.056	-1.049	-1.044	-1.041	-1.025
	0.05	-1.895	-1.325	-1.076	-0.976	-0.930	-0.906	-0.892	-0.884	-0.878	-0.874	-0.871	-0.858
	0.5	0.000	0.232	0.370	0.437	0.471	0.489	0.500	0.507	0.512	0.515	0.517	0.528
	0.95	1.895	2.649	3.163	3.427	3.563	3.637	3.682	3.711	3.730	3.744	3.753	3.797
	0.975	2.365	3.363	4.041	4.389	4.567	4.665	4.724	4.762	4.787	4.805	4.818	4.875
	0.99	2.998	4.371	5.301	5.775	6.018	6.152	6.232	6.284	6.318	6.342	6.360	6.438
8	0.01	-2.897	-1.971	-1.573	-1.417	-1.346	-1.310	-1.289	-1.276	-1.267	-1.261	-1.257	-1.237
	0.025	-2.306	-1.623	-1.324	-1.204	-1.149	-1.120	-1.104	-1.094	-1.087	-1.082	-1.079	-1.063
	0.05	-1.860	-1.336	-1.102	-1.006	-0.962	-0.939	-0.925	-0.917	-0.911	-0.907	-0.905	-0.892
	0.5	0.000	0.217	0.344	0.407	0.438	0.455	0.465	0.471	0.476	0.479	0.481	0.491
	0.95	1.860	2.540	3.000	3.236	3.357	3.424	3.464	3.489	3.506	3.519	3.527	3.566
	0.975	2.306	3.199	3.802	4.110	4.269	4.356	4.408	4.442	4.464	4.480	4.491	4.542
	0.99	2.897	4.113	4.931	5.348	5.562	5.680	5.751	5.796	5.826	5.847	5.863	5.931
9	0.01	-2.821	-1.972	-1.599	-1.450	-1.382	-1.346	-1.326	-1.313	-1.305	-1.299	-1.295	-1.276
	0.025	-2.262	-1.631	-1.348	-1.233	-1.180	-1.152	-1.136	-1.126	-1.120	-1.115	-1.112	-1.097
	0.05	-1.833	-1.346	-1.124	-1.032	-0.989	-0.967	-0.954	-0.946	-0.940	-0.936	-0.934	-0.922
	0.5	0.000	0.204	0.324	0.382	0.411	0.427	0.437	0.443	0.447	0.449	0.452	0.461
	0.95	1.833	2.456	2.875	3.089	3.199	3.260	3.296	3.320	3.335	3.346	3.354	3.389
	0.975	2.262	3.075	3.621	3.900	4.043	4.122	4.169	4.199	4.219	4.234	4.244	4.290
	0.99	2.821	3.920	4.655	5.030	5.222	5.328	5.391	5.432	5.459	5.478	5.492	5.553
10	0.01	-2.764	-1.975	-1.622	-1.479	-1.413	-1.379	-1.359	-1.347	-1.338	-1.333	-1.328	-1.310
	0.025	-2.228	-1.638	-1.369	-1.258	-1.207	-1.180	-1.165	-1.155	-1.148	-1.144	-1.141	-1.126
	0.05	-1.813	-1.355	-1.143	-1.054	-1.013	-0.991	-0.979	-0.971	-0.965	-0.962	-0.959	-0.947
	0.5	0.000	0.193	0.306	0.361	0.389	0.404	0.413	0.418	0.422	0.425	0.427	0.435
	0.95	1.813	2.390	2.776	2.973	3.074	3.130	3.164	3.185	3.199	3.209	3.217	3.249
	0.975	2.228	2.978	3.479	3.734	3.865	3.938	3.981	4.009	4.027	4.040	4.050	4.092
	0.99	2.764	3.771	4.442	4.783	4.958	5.055	5.112	5.149	5.174	5.191	5.204	5.260
11	0.01	-2.718	-1.979	-1.642	-1.504	-1.441	-1.408	-1.388	-1.376	-1.368	-1.363	-1.359	-1.341
	0.025	-2.201	-1.645	-1.388	-1.281	-1.231	-1.205	-1.190	-1.180	-1.174	-1.170	-1.166	-1.152
	0.05	-1.796	-1.363	-1.160	-1.074	-1.034	-1.013	-1.001	-0.993	-0.988	-0.984	-0.982	-0.970
	0.5	0.000	0.183	0.291	0.343	0.370	0.384	0.392	0.398	0.401	0.404	0.406	0.414
	0.95	1.796	2.336	2.695	2.879	2.973	3.024	3.055	3.075	3.089	3.098	3.105	3.135
	0.975	2.201	2.899	3.364	3.601	3.722	3.789	3.829	3.854	3.872	3.884	3.892	3.931
	0.99	2.718	3.651	4.271	4.586	4.747	4.836	4.889	4.923	4.946	4.962	4.973	5.025

Quantiles for Skew-t Distributions

Note: Skew-t Distributions have been shifted to the left δ and rescaled by $\sqrt{\nu/(1+\delta^2)}$

	δ:	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	∞
ν:	α:												
12	0.01	-2.681	-1.983	-1.661	-1.528	-1.466	-1.434	-1.415	-1.403	-1.395	-1.390	-1.386	-1.369
	0.025	-2.179	-1.651	-1.405	-1.301	-1.253	-1.227	-1.213	-1.203	-1.197	-1.193	-1.190	-1.176
	0.05	-1.782	-1.370	-1.175	-1.092	-1.053	-1.033	-1.021	-1.013	-1.008	-1.004	-1.002	-0.991
	0.5	0.000	0.175	0.278	0.328	0.353	0.367	0.375	0.380	0.383	0.386	0.388	0.395
	0.95	1.782	2.291	2.628	2.800	2.888	2.936	2.965	2.984	2.996	3.005	3.011	3.040
	0.975	2.179	2.835	3.269	3.490	3.603	3.666	3.703	3.727	3.743	3.754	3.762	3.799
	0.99	2.681	3.554	4.130	4.424	4.573	4.656	4.706	4.737	4.758	4.773	4.784	4.832
13	0.01	-2.650	-1.987	-1.678	-1.549	-1.489	-1.457	-1.439	-1.427	-1.420	-1.414	-1.411	-1.394
	0.025	-2.160	-1.658	-1.420	-1.319	-1.272	-1.248	-1.233	-1.224	-1.218	-1.214	-1.211	-1.197
	0.05	-1.771	-1.377	-1.189	-1.108	-1.070	-1.050	-1.038	-1.031	-1.026	-1.023	-1.020	-1.009
	0.5	0.000	0.168	0.267	0.315	0.338	0.352	0.359	0.364	0.368	0.370	0.372	0.379
	0.95	1.771	2.253	2.571	2.733	2.816	2.862	2.889	2.906	2.918	2.926	2.932	2.959
	0.975	2.160	2.780	3.189	3.397	3.503	3.562	3.597	3.619	3.634	3.645	3.653	3.687
	0.99	2.650	3.472	4.013	4.288	4.428	4.506	4.552	4.582	4.602	4.616	4.626	4.671
14	0.01	-2.625	-1.992	-1.694	-1.568	-1.510	-1.479	-1.461	-1.450	-1.442	-1.437	-1.433	-1.417
	0.025	-2.145	-1.663	-1.434	-1.336	-1.290	-1.266	-1.252	-1.243	-1.237	-1.233	-1.230	-1.217
	0.05	-1.761	-1.384	-1.201	-1.123	-1.086	-1.066	-1.055	-1.047	-1.043	-1.039	-1.037	-1.026
	0.5	0.000	0.162	0.257	0.303	0.326	0.338	0.346	0.350	0.354	0.356	0.357	0.365
	0.95	1.761	2.220	2.522	2.675	2.754	2.797	2.823	2.840	2.851	2.859	2.864	2.889
	0.975	2.145	2.733	3.120	3.317	3.418	3.473	3.506	3.527	3.542	3.552	3.559	3.591
	0.99	2.625	3.402	3.913	4.173	4.305	4.378	4.422	4.450	4.468	4.482	4.491	4.533
15	0.01	-2.603	-1.996	-1.708	-1.586	-1.529	-1.499	-1.481	-1.470	-1.463	-1.458	-1.454	-1.438
	0.025	-2.131	-1.669	-1.447	-1.352	-1.307	-1.283	-1.269	-1.261	-1.255	-1.251	-1.248	-1.235
	0.05	-1.753	-1.389	-1.213	-1.136	-1.100	-1.081	-1.070	-1.063	-1.058	-1.055	-1.052	-1.042
	0.5	0.000	0.156	0.248	0.292	0.314	0.326	0.333	0.338	0.341	0.343	0.345	0.352
	0.95	1.753	2.192	2.479	2.625	2.700	2.741	2.766	2.782	2.792	2.800	2.805	2.829
	0.975	2.131	2.693	3.061	3.248	3.343	3.396	3.428	3.448	3.461	3.471	3.478	3.508
	0.99	2.603	3.343	3.827	4.073	4.199	4.268	4.309	4.336	4.353	4.366	4.375	4.415
16	0.01	-2.584	-2.000	-1.721	-1.602	-1.546	-1.517	-1.500	-1.489	-1.482	-1.477	-1.473	-1.457
	0.025	-2.120	-1.674	-1.458	-1.366	-1.322	-1.299	-1.285	-1.277	-1.271	-1.267	-1.264	-1.251
	0.05	-1.746	-1.395	-1.223	-1.148	-1.113	-1.094	-1.083	-1.076	-1.072	-1.069	-1.066	-1.056
	0.5	0.000	0.151	0.240	0.282	0.304	0.315	0.322	0.327	0.330	0.332	0.333	0.340
	0.95	1.746	2.167	2.442	2.581	2.653	2.692	2.715	2.730	2.741	2.748	2.753	2.776
	0.975	2.120	2.658	3.009	3.187	3.278	3.329	3.359	3.378	3.391	3.400	3.406	3.435
	0.99	2.584	3.291	3.752	3.986	4.106	4.172	4.211	4.236	4.253	4.265	4.274	4.312

Quantiles for Skew-t Distributions

Note: Skew-t Distributions have been shifted to the left δ and rescaled by $\sqrt{\nu/(1+\delta^2)}$

	δ:	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	∞
ν:	α:												
17	0.01	-2.567	-2.005	-1.734	-1.617	-1.563	-1.534	-1.517	-1.507	-1.5	-1.495	-1.491	-1.475
	0.025	-2.110	-1.679	-1.469	-1.379	-1.336	-1.313	-1.300	-1.291	-1.286	-1.282	-1.279	-1.267
	0.05	-1.740	-1.400	-1.233	-1.160	-1.125	-1.107	-1.096	-1.089	-1.085	-1.082	-1.079	-1.069
	0.5	0.000	0.147	0.232	0.274	0.294	0.306	0.312	0.317	0.320	0.322	0.323	0.330
	0.95	1.740	2.145	2.408	2.542	2.610	2.648	2.671	2.685	2.695	2.702	2.706	2.728
	0.975	2.110	2.626	2.963	3.134	3.221	3.269	3.298	3.316	3.328	3.337	3.343	3.371
	0.99	2.567	3.245	3.687	3.910	4.024	4.087	4.125	4.149	4.165	4.176	4.184	4.221
18	0.01	-2.552	-2.009	-1.745	-1.632	-1.578	-1.550	-1.533	-1.523	-1.516	-1.511	-1.508	-1.492
	0.025	-2.101	-1.684	-1.480	-1.391	-1.349	-1.327	-1.314	-1.305	-1.300	-1.296	-1.293	-1.281
	0.05	-1.734	-1.405	-1.242	-1.170	-1.136	-1.118	-1.108	-1.101	-1.097	-1.094	-1.091	-1.081
	0.5	0.000	0.142	0.226	0.266	0.286	0.297	0.303	0.308	0.310	0.312	0.314	0.320
	0.95	1.734	2.125	2.379	2.507	2.573	2.609	2.631	2.644	2.654	2.660	2.665	2.686
	0.975	2.101	2.599	2.922	3.086	3.169	3.216	3.243	3.261	3.273	3.281	3.287	3.314
	0.99	2.552	3.204	3.628	3.842	3.951	4.012	4.048	4.071	4.086	4.097	4.105	4.140
19	0.01	-2.540	-2.013	-1.756	-1.645	-1.592	-1.565	-1.549	-1.538	-1.532	-1.527	-1.523	-1.508
	0.025	-2.093	-1.689	-1.489	-1.402	-1.361	-1.339	-1.326	-1.318	-1.313	-1.309	-1.306	-1.294
	0.05	-1.729	-1.409	-1.250	-1.180	-1.147	-1.129	-1.119	-1.112	-1.108	-1.105	-1.103	-1.093
	0.5	0.000	0.138	0.219	0.258	0.278	0.289	0.295	0.299	0.302	0.304	0.305	0.311
	0.95	1.729	2.107	2.352	2.476	2.539	2.574	2.595	2.608	2.617	2.623	2.628	2.648
	0.975	2.093	2.574	2.885	3.043	3.123	3.168	3.194	3.211	3.223	3.231	3.236	3.262
	0.99	2.540	3.168	3.575	3.781	3.886	3.944	3.979	4.001	4.016	4.026	4.034	4.068
20	0.01	-2.528	-2.017	-1.766	-1.657	-1.606	-1.579	-1.563	-1.553	-1.546	-1.541	-1.538	-1.523
	0.025	-2.086	-1.693	-1.498	-1.413	-1.372	-1.351	-1.338	-1.330	-1.325	-1.321	-1.319	-1.307
	0.05	-1.725	-1.414	-1.258	-1.189	-1.157	-1.139	-1.129	-1.123	-1.118	-1.115	-1.113	-1.104
	0.5	0.000	0.135	0.214	0.252	0.271	0.281	0.287	0.291	0.294	0.296	0.297	0.303
	0.95	1.725	2.091	2.327	2.447	2.508	2.542	2.562	2.575	2.583	2.589	2.594	2.613
	0.975	2.086	2.551	2.852	3.004	3.082	3.124	3.150	3.166	3.177	3.185	3.191	3.215
	0.99	2.528	3.135	3.528	3.727	3.828	3.884	3.917	3.938	3.953	3.963	3.970	4.002
21	0.01	-2.518	-2.021	-1.776	-1.669	-1.619	-1.592	-1.576	-1.566	-1.560	-1.555	-1.552	-1.537
	0.025	-2.080	-1.697	-1.506	-1.423	-1.383	-1.362	-1.350	-1.342	-1.337	-1.333	-1.330	-1.319
	0.05	-1.721	-1.417	-1.265	-1.198	-1.166	-1.149	-1.139	-1.132	-1.128	-1.125	-1.123	-1.114
	0.5	0.000	0.132	0.208	0.245	0.264	0.274	0.280	0.284	0.287	0.288	0.290	0.295
	0.95	1.721	2.076	2.305	2.421	2.480	2.513	2.532	2.544	2.553	2.559	2.563	2.582
	0.975	2.080	2.530	2.822	2.969	3.044	3.085	3.1100	3.126	3.136	3.144	3.149	3.173
	0.99	2.518	3.106	3.485	3.677	3.775	3.829	3.861	3.881	3.895	3.905	3.912	3.943
22	0.01	-2.508	-2.024	-1.785	-1.680	-1.630	-1.604	-1.589	-1.579	-1.573	-1.568	-1.565	-1.550
	0.025	-2.074	-1.701	-1.514	-1.432	-1.393	-1.372	-1.360	-1.353	-1.347	-1.344	-1.341	-1.330
	0.05	-1.717	-1.421	-1.272	-1.206	-1.174	-1.158	-1.148	-1.142	-1.137	-1.134	-1.132	-1.123
	0.5	0.000	0.128	0.203	0.240	0.258	0.268	0.274	0.277	0.280	0.282	0.283	0.288
	0.95	1.717	2.062	2.285	2.397	2.454	2.486	2.505	2.517	2.525	2.530	2.535	2.553
	0.975	2.074	2.512	2.794	2.936	3.009	3.049	3.073	3.088	3.098	3.106	3.111	3.134
	0.99	2.508	3.079	3.446	3.632	3.726	3.778	3.810	3.829	3.843	3.852	3.859	3.889

Quantiles for Skew-t Distributions

Note: Skew-t Distributions have been shifted to the left δ and rescaled by $\sqrt{\nu/(1+\delta^2)}$

	δ:	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	∞
ν:	α:												
23	0.01	-2.500	-2.028	-1.793	-1.691	-1.642	-1.616	-1.601	-1.591	-1.585	-1.580	-1.577	-1.563
	0.025	-2.069	-1.704	-1.522	-1.441	-1.403	-1.382	-1.370	-1.363	-1.358	-1.354	-1.352	-1.340
	0.05	-1.714	-1.425	-1.279	-1.214	-1.183	-1.166	-1.156	-1.150	-1.146	-1.143	-1.141	-1.132
	0.5	0.000	0.126	0.199	0.234	0.252	0.262	0.267	0.271	0.273	0.275	0.276	0.282
	0.95	1.714	2.050	2.266	2.375	2.431	2.461	2.479	2.491	2.499	2.505	2.509	2.526
	0.975	2.069	2.495	2.768	2.907	2.977	3.016	3.039	3.054	3.064	3.071	3.076	3.099
	0.99	2.500	3.054	3.411	3.590	3.682	3.732	3.763	3.782	3.795	3.804	3.810	3.840
24	0.01	-2.492	-2.032	-1.801	-1.700	-1.652	-1.627	-1.612	-1.603	-1.596	-1.592	-1.589	-1.575
	0.025	-2.064	-1.708	-1.529	-1.450	-1.412	-1.392	-1.380	-1.372	-1.367	-1.364	-1.361	-1.350
	0.05	-1.711	-1.428	-1.285	-1.221	-1.190	-1.174	-1.164	-1.158	-1.154	-1.152	-1.149	-1.140
	0.5	0.000	0.123	0.195	0.229	0.247	0.256	0.262	0.265	0.268	0.269	0.270	0.276
	0.95	1.711	2.039	2.249	2.355	2.409	2.438	2.456	2.468	2.475	2.481	2.484	2.502
	0.975	2.064	2.479	2.745	2.879	2.948	2.986	3.008	3.022	3.032	3.039	3.044	3.066
	0.99	2.492	3.031	3.378	3.552	3.641	3.690	3.719	3.738	3.751	3.759	3.766	3.794
25	0.01	-2.485	-2.035	-1.809	-1.710	-1.663	-1.637	-1.623	-1.613	-1.607	-1.603	-1.600	-1.586
	0.025	-2.060	-1.711	-1.535	-1.457	-1.420	-1.400	-1.389	-1.381	-1.376	-1.373	-1.371	-1.360
	0.05	-1.708	-1.431	-1.290	-1.228	-1.198	-1.182	-1.172	-1.166	-1.162	-1.159	-1.157	-1.148
	0.5	0.000	0.120	0.191	0.224	0.241	0.251	0.256	0.260	0.262	0.264	0.265	0.270
	0.95	1.708	2.028	2.233	2.336	2.388	2.417	2.435	2.446	2.453	2.458	2.462	2.479
	0.975	2.060	2.464	2.723	2.854	2.921	2.957	2.979	2.993	3.003	3.009	3.014	3.035
	0.99	2.485	3.010	3.347	3.517	3.603	3.651	3.680	3.698	3.710	3.718	3.725	3.752
26	0.01	-2.479	-2.038	-1.816	-1.719	-1.672	-1.647	-1.633	-1.624	-1.618	-1.613	-1.610	-1.597
	0.025	-2.056	-1.715	-1.542	-1.465	-1.428	-1.409	-1.397	-1.390	-1.385	-1.382	-1.379	-1.369
	0.05	-1.706	-1.435	-1.296	-1.234	-1.205	-1.189	-1.179	-1.174	-1.170	-1.167	-1.165	-1.156
	0.5	0.000	0.118	0.187	0.220	0.237	0.246	0.251	0.255	0.257	0.258	0.260	0.265
	0.95	1.706	2.018	2.218	2.318	2.370	2.398	2.415	2.425	2.433	2.438	2.441	2.458
	0.975	2.056	2.450	2.703	2.830	2.895	2.931	2.952	2.966	2.975	2.982	2.986	3.007
	0.99	2.479	2.991	3.319	3.484	3.568	3.615	3.643	3.660	3.672	3.680	3.686	3.713
27	0.01	-2.473	-2.041	-1.823	-1.727	-1.681	-1.657	-1.643	-1.634	-1.628	-1.623	-1.620	-1.607
	0.025	-2.052	-1.718	-1.548	-1.472	-1.436	-1.417	-1.405	-1.398	-1.394	-1.390	-1.388	-1.377
	0.05	-1.703	-1.437	-1.301	-1.240	-1.211	-1.196	-1.186	-1.181	-1.177	-1.174	-1.172	-1.163
	0.5	0.000	0.116	0.183	0.216	0.232	0.241	0.246	0.250	0.252	0.253	0.255	0.260
	0.95	1.703	2.009	2.204	2.302	2.352	2.379	2.396	2.406	2.413	2.418	2.422	2.438
	0.975	2.052	2.438	2.684	2.809	2.872	2.907	2.928	2.941	2.950	2.956	2.961	2.981
	0.99	2.473	2.973	3.293	3.454	3.536	3.581	3.608	3.625	3.637	3.645	3.651	3.677

Quantiles for Skew-t Distributions

Note: Skew-t Distributions have been shifted to the left δ and rescaled by $\sqrt{\nu/(1+\delta^2)}$

	δ:	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	∞
ν:	α:												
28	0.01	-2.467	-2.044	-1.830	-1.736	-1.690	-1.666	-1.652	-1.643	-1.637	-1.633	-1.630	-1.617
	0.025	-2.048	-1.721	-1.554	-1.479	-1.443	-1.424	-1.413	-1.406	-1.401	-1.398	-1.396	-1.385
	0.05	-1.701	-1.440	-1.306	-1.246	-1.217	-1.202	-1.193	-1.187	-1.183	-1.181	-1.179	-1.170
	0.5	0.000	0.114	0.180	0.212	0.228	0.237	0.242	0.245	0.247	0.249	0.250	0.255
	0.95	1.701	2.000	2.191	2.287	2.335	2.362	2.378	2.389	2.396	2.400	2.404	2.419
	0.975	2.048	2.426	2.667	2.788	2.850	2.884	2.904	2.917	2.926	2.932	2.936	2.956
	0.99	2.467	2.956	3.269	3.426	3.506	3.550	3.576	3.593	3.604	3.612	3.618	3.643
29	0.01	-2.462	-2.047	-1.837	-1.743	-1.699	-1.675	-1.661	-1.652	-1.646	-1.642	-1.639	-1.626
	0.025	-2.045	-1.724	-1.559	-1.486	-1.451	-1.432	-1.421	-1.414	-1.409	-1.406	-1.403	-1.393
	0.05	-1.699	-1.443	-1.311	-1.252	-1.223	-1.208	-1.199	-1.194	-1.190	-1.187	-1.185	-1.177
	0.5	0.000	0.112	0.177	0.208	0.224	0.232	0.237	0.241	0.243	0.244	0.245	0.250
	0.95	1.699	1.992	2.179	2.272	2.320	2.346	2.362	2.372	2.379	2.383	2.387	2.402
	0.975	2.045	2.415	2.651	2.769	2.829	2.863	2.882	2.895	2.904	2.910	2.914	2.933
	0.99	2.462	2.941	3.246	3.399	3.477	3.520	3.546	3.562	3.573	3.581	3.587	3.612
30	0.01	-2.457	-2.050	-1.843	-1.751	-1.707	-1.683	-1.669	-1.661	-1.655	-1.651	-1.648	-1.635
	0.025	-2.042	-1.726	-1.564	-1.492	-1.457	-1.439	-1.428	-1.421	-1.416	-1.413	-1.411	-1.400
	0.05	-1.697	-1.446	-1.316	-1.257	-1.229	-1.214	-1.205	-1.200	-1.196	-1.193	-1.191	-1.183
	0.5	0.000	0.110	0.174	0.205	0.220	0.228	0.233	0.237	0.239	0.240	0.241	0.246
	0.95	1.697	1.985	2.167	2.259	2.305	2.331	2.347	2.356	2.363	2.368	2.371	2.386
	0.975	2.042	2.405	2.635	2.751	2.810	2.842	2.862	2.874	2.883	2.888	2.893	2.911
	0.99	2.457	2.926	3.225	3.374	3.451	3.493	3.518	3.534	3.545	3.552	3.558	3.582
40	0.01	-2.423	-2.074	-1.893	-1.812	-1.772	-1.751	-1.739	-1.731	-1.726	-1.722	-1.719	-1.708
	0.025	-2.021	-1.749	-1.607	-1.543	-1.512	-1.495	-1.485	-1.479	-1.475	-1.472	-1.470	-1.461
	0.05	-1.684	-1.467	-1.352	-1.301	-1.275	-1.262	-1.254	-1.249	-1.246	-1.243	-1.242	-1.234
	0.5	0.000	0.095	0.150	0.177	0.190	0.197	0.202	0.204	0.206	0.207	0.208	0.213
	0.95	1.684	1.927	2.081	2.157	2.196	2.218	2.230	2.238	2.244	2.248	2.251	2.263
	0.975	2.021	2.327	2.520	2.616	2.665	2.692	2.708	2.719	2.726	2.730	2.734	2.750
	0.99	2.423	2.817	3.066	3.190	3.253	3.288	3.309	3.322	3.331	3.337	3.342	3.362
50	0.01	-2.403	-2.093	-1.929	-1.855	-1.819	-1.800	-1.789	-1.782	-1.777	-1.773	-1.771	-1.760
	0.025	-2.009	-1.766	-1.638	-1.579	-1.551	-1.536	-1.527	-1.521	-1.517	-1.515	-1.513	-1.504
	0.05	-1.676	-1.482	-1.379	-1.331	-1.309	-1.296	-1.289	-1.284	-1.281	-1.279	-1.277	-1.271
	0.5	0.000	0.085	0.134	0.158	0.170	0.176	0.180	0.182	0.184	0.185	0.186	0.190
	0.95	1.676	1.891	2.025	2.092	2.126	2.145	2.156	2.163	2.168	2.171	2.173	2.184
	0.975	2.009	2.278	2.447	2.531	2.573	2.597	2.611	2.620	2.626	2.630	2.633	2.646
	0.99	2.403	2.749	2.966	3.074	3.129	3.159	3.177	3.188	3.196	3.202	3.206	3.223
60	0.01	-2.390	-2.108	-1.958	-1.889	-1.856	-1.838	-1.827	-1.821	-1.816	-1.813	-1.811	-1.801
	0.025	-2.000	-1.779	-1.661	-1.607	-1.581	-1.567	-1.558	-1.553	-1.550	-1.547	-1.545	-1.537
	0.05	-1.671	-1.494	-1.399	-1.355	-1.334	-1.322	-1.315	-1.311	-1.308	-1.306	-1.305	-1.298
	0.5	0.000	0.077	0.122	0.144	0.155	0.161	0.164	0.166	0.168	0.169	0.170	0.173
	0.95	1.671	1.865	1.986	2.046	2.076	2.093	2.103	2.109	2.113	2.116	2.119	2.128
	0.975	2.000	2.244	2.395	2.470	2.508	2.529	2.542	2.550	2.555	2.559	2.562	2.574
	0.99	2.390	2.702	2.896	2.993	3.042	3.069	3.085	3.095	3.102	3.107	3.110	3.126

Quantiles for Skew-t Distributions

Note: Skew-t Distributions have been shifted to the left δ and rescaled by $\sqrt{\nu/(1+\delta^2)}$

	δ:	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	∞
ν:	α:												
70	0.01	-2.381	-2.120	-1.980	-1.916	-1.885	-1.868	-1.858	-1.852	-1.848	-1.845	-1.843	-1.833
	0.025	-1.994	-1.790	-1.680	-1.630	-1.605	-1.592	-1.584	-1.579	-1.576	-1.573	-1.572	-1.564
	0.05	-1.667	-1.503	-1.414	-1.374	-1.354	-1.343	-1.337	-1.333	-1.330	-1.328	-1.327	-1.321
	0.5	0.000	0.072	0.113	0.133	0.143	0.149	0.152	0.154	0.155	0.156	0.157	0.160
	0.95	1.667	1.846	1.956	2.011	2.039	2.054	2.063	2.069	2.073	2.075	2.077	2.086
	0.975	1.994	2.218	2.356	2.425	2.460	2.479	2.490	2.498	2.502	2.506	2.508	2.519
	0.99	2.381	2.667	2.844	2.932	2.977	3.001	3.016	3.025	3.032	3.036	3.039	3.053
80	0.01	-2.374	-2.130	-1.999	-1.938	-1.909	-1.893	-1.884	-1.878	-1.874	-1.871	-1.869	-1.860
	0.025	-1.990	-1.799	-1.696	-1.648	-1.625	-1.612	-1.605	-1.600	-1.597	-1.595	-1.593	-1.586
	0.05	-1.664	-1.511	-1.427	-1.389	-1.370	-1.360	-1.354	-1.350	-1.348	-1.346	-1.345	-1.339
	0.5	0.000	0.067	0.106	0.125	0.134	0.139	0.142	0.144	0.145	0.146	0.147	0.150
	0.95	1.664	1.830	1.933	1.984	2.009	2.023	2.032	2.037	2.041	2.043	2.045	2.053
	0.975	1.990	2.198	2.326	2.389	2.422	2.439	2.450	2.457	2.461	2.464	2.466	2.477
	0.99	2.374	2.640	2.804	2.885	2.926	2.949	2.962	2.971	2.977	2.981	2.984	2.997
90	0.01	-2.369	-2.139	-2.015	-1.957	-1.929	-1.914	-1.905	-1.899	-1.896	-1.893	-1.891	-1.883
	0.025	-1.987	-1.807	-1.709	-1.663	-1.641	-1.629	-1.622	-1.618	-1.615	-1.613	-1.611	-1.604
	0.05	-1.662	-1.517	-1.438	-1.402	-1.384	-1.374	-1.369	-1.365	-1.363	-1.361	-1.360	-1.354
	0.5	0.000	0.063	0.100	0.117	0.126	0.131	0.134	0.136	0.137	0.138	0.138	0.141
	0.95	1.662	1.818	1.914	1.961	1.985	1.998	2.006	2.011	2.015	2.017	2.019	2.026
	0.975	1.987	2.182	2.301	2.361	2.391	2.407	2.417	2.423	2.428	2.430	2.433	2.442
	0.99	2.369	2.618	2.771	2.847	2.885	2.906	2.919	2.927	2.932	2.936	2.939	2.951
100	0.01	-2.364	-2.147	-2.028	-1.973	-1.947	-1.932	-1.924	-1.918	-1.915	-1.912	-1.910	-1.902
	0.025	-1.984	-1.813	-1.720	-1.677	-1.656	-1.644	-1.637	-1.633	-1.630	-1.628	-1.627	-1.620
	0.05	-1.660	-1.523	-1.448	-1.413	-1.396	-1.387	-1.381	-1.378	-1.375	-1.374	-1.373	-1.367
	0.5	0.000	0.060	0.095	0.111	0.120	0.124	0.127	0.129	0.130	0.131	0.131	0.134
	0.95	1.660	1.808	1.898	1.943	1.965	1.978	1.985	1.990	1.993	1.995	1.997	2.004
	0.975	1.984	2.168	2.281	2.337	2.365	2.381	2.390	2.396	2.400	2.402	2.404	2.413
	0.99	2.364	2.599	2.744	2.815	2.851	2.871	2.883	2.890	2.895	2.899	2.902	2.913
∞	0.01	-2.326	-2.326	-2.326	-2.326	-2.326	-2.326	-2.326	-2.326	-2.326	-2.326	-2.326	-2.326
	0.025	-1.960	-1.960	-1.900	-1.960	-1.960	-1.960	-1.960	-1.960	-1.960	-1.960	-1.960	-1.960
	0.05	-1.645	-1.645	-1.645	-1.645	-1.645	-1.645	-1.645	-1.645	-1.645	-1.645	-1.645	-1.645
	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.95	1.645	1.645	1.645	1.645	1.645	1.645	1.645	1.645	1.645	1.645	1.645	1.645
	0.975	1.960	1.960	1.960	1.960	1.960	1.960	1.960	1.960	1.960	1.960	1.960	1.960
	0.99	2.326	2.326	2.326	2.326	2.326	2.326	2.326	2.326	2.326	2.326	2.326	2.326

References

Abramowitz, M. and Stegun, I. A., eds. (1970). *Handbook of Mathematical Functions*. Dover, New York.

Baten, W. D. (1934). The probability law for the sum of n independent variables, each subject to the law (1/2h) sech($\pi x/2h$). *Bulletin of the American Mathematical Society* **40**, 284–290.

Coifman, R. R. and Donoho, D. L. (1995). Translation-invariant de-noising. In *Wavelets and Statistics* (Editors: A. Antoniadis and G. Oppenheim), 125–150. Springer–Verlag, New York.

Devroye, L. (1993). On random variate generation for the generalized hyperbolic secant distributions. *Statistics and Computing* **3**, 125–134.

Division, D. A. P. (1999). *Splus 2000 Guide to Statistics, Volume 1*. MathSoft, Seattle, Washington.

Ebeling, H., White, D., and Rangarajan, F. (2000). ASMOOTH: A simple and efficient algorithm for adaptive kernel smoothing of two-dimensional imaging data. *Monthly Notices of the Royal Astronomical Society* .

Freeman, P. E., Kashyap, V., Rosner, R., and Lamb, D. Q. (2002). A wavelet-based algorithm for the spatial analysis of poisson data. *The Astrophysical Journal Supplement* **138**, 185–218.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulations using multiple sequences (with discussion). *Statistical Science* **7**, 457–472.

Grinold, R. C. and Kahn, R. N. (1995). *Active Portfolio Management*. McGraw Hill.

Harkness, W. L. and Harkness, M. L. (1968). Generalized hyperbolic secant distributions. *Journal of the American Statistical Association* **63**, 329–337.

Hobler, G. and Selberherr, S. (1988). Two-dimensional modeling of ion implantation induced point defects. *IEEE Transactions on CAD* **7**, 174.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions. Volume 2 (Second Edition) (ISBN 0471584940)*. Wiley.

Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistics. Vol. I: Distribution Theory (4th Ed); Vol. 2: Inference and Relationship (3rd Ed)*. Charles Griffin.

Lucy, L. B. (1974). An iterative technique for the rectification of observed distributions. *The Astronomical Journal* **79**, 745–754.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (Second edition)*. Chapman & Hall, London.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics* **10**, 65–80.

Morris, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics* **11**, 515–529.

Morris, C. N. (1988). Approximating posterior distributions and posterior moments. In *Bayesian Statistics 3*, 327–344. Clarendon Press (Oxford).

Nagahara, Y. (1999). The Pdf and Cf of Pearson type Iv distributions and the Ml estimation of the parameters. *Statistics & Probability Letters* **43**, 251–264.

Nowak, R. D. and Kolaczyk, E. D. (2000). A bayesian multiscale framework for poisson inverse problems. *IEEE Transactions on Information Theory* **46**, 1811–1825.

Richardson, W. H. (1972). Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America* **62**, 55–59.

Slater, L. J. (1960). *Confluent Hypergeometric Functions*. Cambridge University Press, Cambridge.

van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *The Astrophysical Journal* **548**, 224–243.