

Estimating the Parameters of Infinite Scale Mixtures of Normals

Hasan Hamdan

*Department of Mathematics and Statistics
James Madison University, Harrisonburg, VA, 22807.*

and

John P. Nolan

*Department of Mathematics and Statistics
American University, Washington D. C., 20016*

June 17, 2004

Abstract

Conditions and classes of examples of variance mixture of normals are given, along with a constructive proof on how to guarantee that a finite variance mixtures of normals is uniformly close (up to a desired tolerance level) to a given infinite variance mixture distribution.

We wish to minimize the finite number of terms needed subject to a specified desired tolerance level. The method, which is based on discretizing the mixing measure is presented and illustrated through an example and the infinite and finite mixtures are displayed on the same graph. A new method for estimating the parameters of a variance mixture of normals is also introduced. The new method is based on minimizing the squared distance between the estimated density and the corresponding density computed by discretizing the mixture over a predetermined grid of R values and a grid of X values. This method looks promising especially for modeling data.

Key words: Discretize, recursive, point masses, mixing measure, least squares, unmix

1 Introduction

Necessary and sufficient conditions needed for a given distribution to be a variance mixture of normals are explored. These conditions and classes of examples are given, along with a constructive proof on how to guarantee that a finite mixture is uniformly close (up to a desired tolerance level) to a given infinite mixture distribution.

The aim is to minimize the finite number of terms needed subject to the

desired tolerance level. The number of terms needed for this approximation depends on the desired tolerance level and the mixing measure, π . The mixing measure may be continuous, however, a discrete version π^* of π is used in the approximation process as a means of simplification. Fujikoshi and Shimizu (1989) and Shimizu(1995) proposed a method in which the conditional distribution of the mixture given the mixing random variable (or the characteristic function) is expanded using Talyor series and error bounds between the exact mixture and the approximation were determined in terms of the L_1 -norm. Here, a new method which is based on discretizing the mixing measure is presented and illustrated through an example.

In section 2, we define Variance Mixture of Normals and give examples. In section 3, a characterization theorem is presented and applied and some results are derived. In section 4, we present and apply a new method for approximating infinite variance mixture by finite variance mixture up to a specified tolerance level. In section 5, we present a new method for estimating the mixing measure based on given data along with a simulated example.

2 Variance Mixtures of Normals

We will say that a random variable X is a (generalized or infinite) variance mixture of normals if

$$X \stackrel{d}{=} AZ, \quad \text{where } Z \sim N(0, 1), \quad A > 0, \quad A \text{ and } Z \text{ independent.}$$

We exclude the possibility that $A = 0$ with positive probability, which would make X have a point mass at the origin, so X would not have a density. Equivalently, X has pdf

$$f(x) = \int_0^\infty \phi(x|\sigma)\pi(d\sigma), \quad (1)$$

where $\phi(x|\sigma)$ is the Normal($0, \sigma^2$) density and the mixing measure π is the distribution of A .

The class of variance mixtures of normals is very large and contains some well known distributions. The following examples are used through out the paper.

Finite mixtures If A takes on a finite number of values, say $\sigma_1, \dots, \sigma_M$ with respective probabilities π_1, \dots, π_M , then the density of $X = AZ$ is

$$f(x) = \sum_{j=1}^M \phi(x|\sigma_j)\pi_j.$$

A common case is when A takes on two values, with $\sigma_1 < \sigma_2$ and $\pi_1 > \pi_2$, which is sometimes called a contaminated normal mixture.

Generalized t distributions Suppose that $1/A^2$ has a Gamma(α, β) distribution. If we set the scale parameter $\beta = 2/c$, then it is easy to see that the

density function of $X = AZ$ is

$$f(x) = \frac{k}{(x^2 + c)^{\alpha+1/2}} \quad -\infty < x < \infty. \quad (2)$$

In particular, when $\alpha = 1/2$ and $\beta = 2$ then $f(x)$ is the standard Cauchy. More generally, when $\alpha = n/2$ and $\beta = 2/n$, then $f(x)$ is the t density with n degrees of freedom.

Exponential Power Family The exponential power family consists of all distributions having densities of the form

$$f(x) = k \exp(-|x|^b) \quad x \in \mathbb{R} \text{ and } b > 0$$

If $1 < b \leq 2$, then f is a variance mixture of normals where the inverse of the scale variable $1/A$ is stable variable. Two important special cases are the normal ($b = 2$) and the Laplace or double-exponential ($b = 1$). See West (1987) and Box and Tiao (1973). The case $b > 2$ cannot be a variance mixture of normals as we will show in the next section.

There are many other classes of scale mixtures of normals. For example, it can be easily shown that the class of symmetrized gamma distributions presented by Feller (1971) and Rohatgi(1976) is scale mixtures of normals. Barndorff-Nielsen, J. Kent, and M. Sorensen (1982), Andrews and Mallows (1974) and Barndorff-Nielsen, J. Kent, and M. Sorensen(1982) showed that the logistic distribution is a scale mixture of normals with the Kolmogorove distant statistic as it's mixing random variable. Samorodnitsky and Taqqu (1994) showed that some of the symmetric stable distributions, are variance mixtures of normals with mixing measure being stable distribution.

3 Characterization of Variance Mixtures of Normals

The following known result gives necessary and sufficient conditions for a distribution to be a scale mixture of normals. A function $h(x)$ on $(0, \infty)$ is *completely monotone* in x if it is infinitely differentiable and $(-1)^m h^{(m)}(x) \geq 0$ for all x and all $m = 1, 2, \dots$. Feller (1971), pg. 441 shows that the product of two completely monotone functions is also completely monotone. Moreover, the composition of a completely monotone function with a positive function that has a completely monotone derivative is also completely monotone. Examples of completely monotone functions are $\frac{1}{x}$, $\frac{1}{x+1}$, and $\exp(-x^\alpha)$ for $0 < \alpha \leq 1$.

Theorem 1 (Schoenberg (1938)) X with density $f(x)$ is a variance mixture of normals if and only if $h(x) := f(\sqrt{x})$ is a completely monotone function. Equivalently, X is a variance mixture of normals if and only if its characteristic function φ_X is a real, even function such that $\varphi_X(\sqrt{t})$ is completely monotone on $(0, \infty)$.

Example The previous theorem can be applied to the Exponential Power Family to show $b \leq 2$ is necessary for variance mixture of normals. We need to find the range of b such that $h(x) := f(\sqrt{x}) = k \exp(-x^{\frac{b}{2}})$ is a completely monotonic function, and hence $f(x) = h(x^2)$ is a variance mixture of normals. When $0 < b \leq 2$, $h(x)$ is completely monotone by the second criterion of Feller (1971, pg. 441) because $\exp(-x)$ is completely monotone and $x^{\frac{b}{2}}$ has a completely monotone derivative.

Now we want to show that when $b > 2$, $h(x)$ cannot be completely monotone. Since $h(x)$ is an exponential function, it is infinitely differentiable on $(0, \infty)$ and positive. Therefore, all that is needed is to ensure that $(-1)^m h^{(m)}(x) \geq 0$ for $m = 1, 2, \dots$. For $m = 1$, $h'(x) = -\frac{b}{2} k x^{\frac{b-2}{2}} \exp(-x^{\frac{b}{2}})$. So, $-h'(x) \geq 0$ for $b > 0$. For $m = 2$,

$$\begin{aligned} h''(x) &= -\frac{b}{2} x^{\frac{b-2}{2}} h'(x) - \frac{b(b-2)}{2} x^{\frac{b-4}{2}} h(x) \\ &= \frac{b}{2} x^{\frac{b-4}{2}} \left(\frac{b}{2} x^{\frac{b}{2}} - \frac{b-2}{2} \right) h(x). \end{aligned}$$

For $h(x)$ to be completely monotonic, we need $(-1)^2 h''(x) \geq 0$, for all $x > 0$. Letting $x \rightarrow 0^+$, we must have $b \leq 2$. This shows that for $b > 2$, $h(x)$ cannot be completely monotone.

Theorem 1 can be applied to show that if Z_1 and Z_2 are independent variance mixtures of normals then $Z_1 + Z_2$, $Z_1 Z_2$ and $\frac{Z_1}{Z_2}$ are also variance mixtures of normals.

4 Approximating Variance Mixtures of Normals

As above, when A takes on a finite number of values

$$f(x) = \sum_{j=1}^M \phi(x|\sigma_j) \pi_j.$$

One can try to fit any data with such a mixture. Theorem 1 makes it clear that the distribution must be "bell shaped." ($f(\sqrt{x})$ is a completely monotone on $(0, \infty)$ i.e. infinitely differentiable and the derivatives have alternative signs and since $f(x)$ is symmetric, we can see that the density is "bell shaped.") When M , the number of terms, is known, the EM algorithm Redner and Walker (1984), and Lindsay (1995) can be used to estimate the parameters. If M is unknown, then one typically tries different values of M and selects one based on some selection criteria.

Here we consider a related question: suppose it is known that X is a mixture of normals with known scale A having distribution π . If the density of X is difficult to compute, then we might want to approximate it by a finite mixture. Two practical questions are how many terms to take and what values of π_j and σ_j should be chosen.

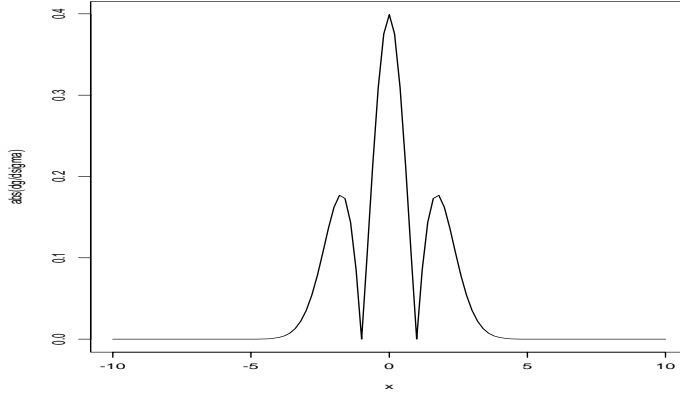


Figure 1: $\left| \frac{\partial \phi}{\partial \sigma} \right|$ at fixed σ as a function of x .

Although we can find the following results, Theorem 2, Lemma 1 and Lemma 2 in Hamdan and Nolan (2004), we present them and present the proof of Theorem 2 for the purpose of completeness. Theorem 2 gives a constructive way of determining these quantities (how many terms to take and what values of π_j and σ_j should be chosen) in a way that guarantees that the densities of X and the finite mixture are uniformly close. Although this may not be an optimal solution, it gives a concrete way of choosing the parameters in a setting that generalizes readily to a larger class of mixtures and multidimensional problems. First we consider the situation where the scale A is bounded away from zero and infinity.

Lemma 1 If $\sigma_1, \sigma_2 \in [a, \infty)$, then

$$|\phi(x|\sigma_1) - \phi(x|\sigma_2)| \leq \frac{1}{\sqrt{2\pi}a^2} |\sigma_1 - \sigma_2| \text{ for all } x \in \mathbb{R}.$$

Proof.

See Figure 1, fixing σ , $|\partial\phi(x|\sigma)/\partial\sigma| = \left| \frac{x^2 - \sigma}{\sigma^2} \right| \phi(x|\sigma)$ is maximized at $x = 0$, where it takes value $\phi(0|\sigma)/\sigma = 1/(\sqrt{2\pi}\sigma^2)$, Hence

$$|\phi(x|\sigma_1) - \phi(x|\sigma_2)| \leq (\max |\partial\phi/\partial\sigma|) |\sigma_1 - \sigma_2| = |\sigma_1 - \sigma_2|/(\sqrt{2\pi}a^2). \quad \blacksquare$$

Theorem 2 Suppose $X = AZ$ where A is a positive random variable with distribution π having support $[a, b]$. For any $\epsilon > 0$, there is a discrete distribution with at most $M = M(\epsilon, a, b)$ point masses π_1, \dots, π_M concentrated on $\sigma_1, \dots, \sigma_M$ in $[a, b]$ which satisfies

$$\sup_{x \in \mathbb{R}} \left| f(x) - \sum_{j=1}^M \text{phi}(x|\sigma_j) \pi_j \right| \leq \epsilon.$$

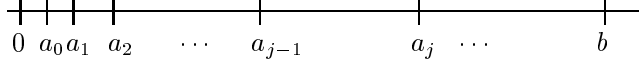


Figure 2: Illustration of the a_j sequence

Proof. We adapt the proof of Lemma 1 in Byczkowski, Nolan, and Rajput (1993). Fix any $\epsilon > 0$, any $0 < a < b < \infty$. Define recursively (see figure 2)

$$a_0 = a, \quad a_j = a_{j-1} + \sqrt{2\pi}a_{j-1}^2\epsilon. \quad (3)$$

The distances between the a_j 's are strictly increasing, so there exists an $M = M(\epsilon, a, b)$ such that $a_{2M} \geq b$.

The term $\sqrt{2\pi}a_{j-1}^2\epsilon$ has to do with the rate of change of $\phi(x|\sigma)$. Define a disjoint cover of $[a, b]$: $I_1 = [a_0, a_2]$, $I_2 = (a_2, a_4]$, \dots , $I_M = (a_{2M-2}, b]$. Set $\pi_j = \pi(I_j)$ and $\sigma_j = \min(a_{2j-1}, b)$, $j = 1, \dots, M$. Then $\phi(x|\sigma_j)\pi_j = \phi(x|\sigma_j) \int_{I_j} \pi(d\sigma)$, so

$$\begin{aligned} \left| f(x) - \sum_{j=1}^M \phi(x|\sigma_j)\pi_j \right| &= \left| \int_{[a, b]} \phi(x|\sigma)\pi(d\sigma) - \sum_{j=1}^M \int_{I_j} \phi(x|\sigma_j)\pi(d\sigma) \right| \\ &= \left| \sum_{j=1}^M \int_{I_j} (\phi(x|\sigma) - \phi(x|\sigma_j)) \pi(d\sigma) \right| \\ &\leq \sum_{j=1}^M \int_{I_j} |\phi(x|\sigma) - \phi(x|\sigma_j)| \pi(d\sigma). \end{aligned}$$

The definition of the a_j 's and Lemma 1 guarantee that if $\sigma \in I_j$, then $|\phi(x|\sigma) - \phi(x|\sigma_j)| \leq \epsilon$. Hence the last line above is $\leq \sum_{j=1}^M \int_{I_j} \epsilon \pi(d\sigma) = \epsilon \pi([a, b]) = \epsilon$.

■

Note: the value of M depends only on a, b , and ϵ through properties of the normal densities $\phi(x|\sigma)$. It can be found by solving equation (3) recursively.

In general, A can take arbitrarily small values and arbitrarily large values. In such a case, write

$$\int_0^\infty \phi(x|\sigma)\pi(d\sigma) = \int_0^a (\cdot) + \int_a^b (\cdot) + \int_b^\infty (\cdot). \quad (4)$$

The following lemma shows that in all cases where $f(0)$ is bounded, there exist a, b such that the first and last integrals are less than $\epsilon/3$, while the middle integral can be approximated using Theorem 2.

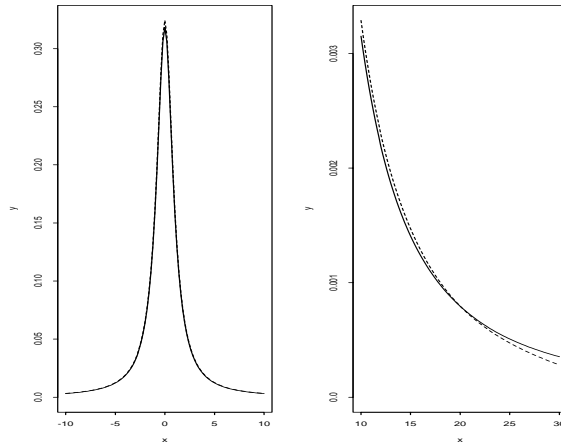


Figure 3: The exact Cauchy (solid) and the approximated Cauchy (dashed)

Lemma 2 Let $X = AZ$ be a scale mixture of normals, $\epsilon > 0$.

(a) If $f(0) < \infty$, then there exists an $a > 0$ such that $\int_0^a \phi(x|\sigma)\pi(d\sigma) < \epsilon$ for all $x \in \mathbb{R}$.

(b) There exists a $b > 0$ such that $\int_b^\infty \phi(x|\sigma)\pi(d\sigma) < \epsilon$ for all $x \in \mathbb{R}$.

Proof. See Hamdan and Nolan (2004). ■

The generalized t distribution example is revisited. Here we study the case $\alpha = \frac{1}{2}$ and $\beta = 2$, so that the marginal distribution is standard Cauchy. The weight function is the square root of Inverted Gamma with parameters a and β . In this case, the corresponding Gamma has a vertical asymptote at 0 and it is decreasing on $\Theta = [a, b]$. Therefore, to capture a reasonable weight of this gamma function, a should be close to zero. To illustrate this point, if $a = .01$ and $b = 6$, we miss 8% of the total density. However, if $a = .001$ and $b = 6$, we only miss 2%.

A comparison between the finite mixture density and the infinite mixture (theoretical mixture), given by equation (2) is made for different combinations of a , b , and ϵ . In the following example, the difference between the actual density and the approximated density was found based on the values of a , b , ϵ , on a grid of 101 equally spaced points. In particular, when $a = .05$, $b = 50$, and $\epsilon = .03$ then f^* is very close to f (the average value for the relative distance between f and f^* is 2.6%) as illustrated in Figure 3.

Although, this approximation is very good overall, the number of components needed ($M = 31$) is large. Moreover, the accuracy is not that good in the tails, as can be seen in Figure 4. For example, average value for the relative distance $\frac{|f-f^*|}{f}$ between f and f^* is around 5.8%.

Depending on the interval of interest, one can always improve this approximation. For example; starting the discretization at smaller a always improves

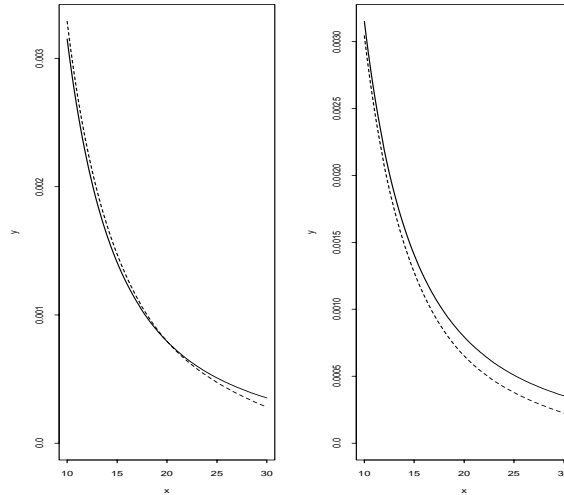


Figure 4: Before and after deleting the lowest 21 terms (Y is the density).

the approximation in the body or around $x = 0$ and ending the discretization at larger b always improves the approximation on the tails. The only problem that arises when we change the discretization limits, especially when a becomes close to zero, is that M can increase dramatically. In this case, we can ignore the terms that have small weights and normalize the remaining terms. For example, in this example, when we eliminate the lowest 21 terms, we still retain a reasonable approximation. See Figure 4.

5 Estimating Mixing Measures

Andrews and Mallows (1974) presented several examples on how to find the mixing measure π when the density of the infinite variance mixture is given.

The result worked for most examples of infinite variance mixtures of normals, but it is not clear how to apply it, and find the mixing measure, when the infinite variance mixture of normals has a symmetric stable density. Two practical questions also arise: Namely, how do we know that a given random sample can reasonably be assumed to come from some scale mixtures of normals? And if it does, how do we estimate the mixing distribution? We briefly discuss the first question; however our main focus is on the second question.

In Theorem 1, we presented necessary and sufficient conditions for a random variable X to be a scale mixtures of normals. In particular, $f(\sqrt{x})$ has to be completely monotonic. However, any empirical pdf cannot be easily tested for complete monotonic since it is not smooth. In particular, any kernel estimator of the pdf will not be completely monotone because it will have bumps at points

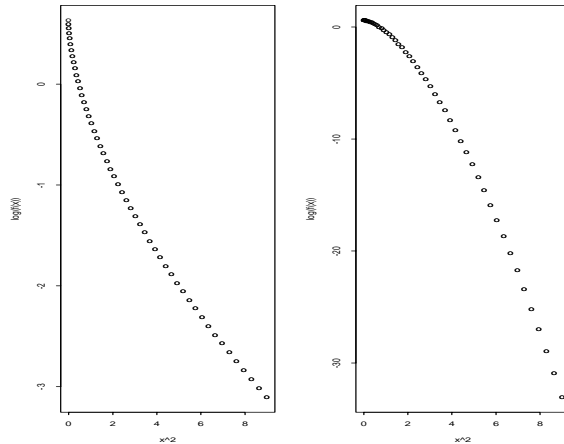


Figure 5: The log/square plot for the Exponential Power density with $b=1.2$ (left) and $b=3.2$ (right).

where the data values are spaced apart. Additionally, a second necessary and sufficient condition based on characteristic functions, which is not easy to apply, was presented by Beale and Mallows (1959). They showed that the kurtosis of a mixture is never less than the kurtosis of a normal. Additionally, they provided a necessary (but not sufficient) condition for X to be a scale mixtures of normals. Specifically, the log/square plot, in which $\log f(x)$ is plotted as a function of x^2 , must be convex. The tails of the infinite scale mixture are always heavier than any normal which can be utilized by modeling the data by a normal with variance equal to the sample variance. In practice, we suggest starting by looking at the unimodality of the empirical density of the sample. If the empirical density is unimodal, we proceed by checking the symmetry. One can use the test statistic suggested by Diks and Tong (1999) to test for symmetry or spherical symmetry. If the empirical density is significantly symmetrical, we proceed by looking at log/square plot.

EXAMPLE 1 *Exponential Power Family.* Recall that a random variable X has an exponential power density if $f(x) = k \exp(-|x|^b)$ and $b > 0$. It was shown in Section 2 that the exponential power family is a variance mixture of normals if and only if $b \leq 2$. To illustrate this result, $\log f(x)$ is plotted as a function of x^2 for $b = 1.2$ and $b = 3.2$. See Figure 5.

Since our main focus is on the second question, we start by reviewing briefly some of the major developments in this area.

5.1 Brief Literature Review

The problem of estimating the mixing measure has been the subject of a large diverse body of literature. Deely and Kruse (1968) outlined the construction of a mixing measure that converges weekly to a priori distribution π . Dempster, Larid, and Rubin (1977) used the EM algorithm for approximating the maximum likelihood estimates. They interpreted the finite mixture density estimation problem as an estimation problem involving incomplete data. They regard an unlabeled observation in the mixture as an observation which is missing a label indicating its component population of origin. Larid (1978) showed that under various conditions the nonparametric maximum likelihood estimate of the mixing distribution is a step function with a finite number of steps. A robust powerful approach based on minimum distance estimation is analyzed by Donoho and Liu (1988) and Beran (1977). Zhang (1990) used Fourier methods to derive kernel estimators and provided lower and upper bounds for the optimal rate of convergence. Priebe (1994) developed a nonparametric maximum likelihood technique from related methods of kernel estimation and finite mixtures.

There are many practical difficulties in estimating the mixing measure. Some of these are computationally difficult and intractable. For example, when we use the EM method to find the MLE of the mixing measure in the finite case, we might find a large local maxima that occurs as a consequence of a fitted component having a very small (but nonzero) variance. Moreover, it is not clear how to initialize the estimates, especially when the mixture is a scale mixture. The key problem in finite mixture models is the number of components in the mixture. Several criteria based on the penalized log-likelihood, such as Akaike Information Criterion, AIC, the Bayesian Information Criterion, BIC and the Information Complexity Criterion introduced by Bozdogan (1993), have been used. Finally, there are two good references in the field of finite mixtures, namely, Titterington, Smith and Makove (1985) and Lindsay (1995).

5.2 UNMIX Program

This method is based on minimizing the squared distance between the estimated density of X and the corresponding density computed by discretizing the mixture over a pre-determined grid of R values and a grid of X values. That is, given a sample of size n from the mixture, fix a grid of r_1, r_2, \dots, r_m values called rgrid and a grid of x_1, \dots, x_k values called xgrid, where $k \geq m$. We can use the rgrid to approximate $f(x)$ as described in section 4 as follows:

$$\begin{aligned} f(x) &= \int_0^\infty \frac{1}{r} \phi\left(\frac{x}{r}\right) \pi(r) dr \\ &\simeq \sum_{j=1}^m \frac{1}{r_j} \phi\left(\frac{x}{r_j}\right) \pi_j. \end{aligned}$$

For each x_i in the xgrid, $f(x_i)$ can be evaluated by $\hat{f}(x_i)$ using a kernel

smoother. If we let $y_i = \widehat{f}(x_i)$, then

$$y_i = \widehat{f}(x_i) = \sum_{j=1}^m \frac{1}{r_j} \phi\left(\frac{x_i}{r_j}\right) \pi_j + \varepsilon_i.$$

Assuming ε_i are independent with mean 0, we can solve for π_j by minimizing $S(\pi)$ where $\pi^T = (\pi_1, \dots, \pi_m)$,

$$S(\pi) = \sum_{i=1}^k \left(w_i \left(y_i - \sum_{j=1}^m \phi_{ij} \pi_j \right) \right)^2,$$

$\phi_{ij} = \frac{1}{r_j} \phi\left(\frac{x_i}{r_j}\right)$, and w_i are preassigned weights, in our case they are ones throughout. However, if the data are heavy-tailed then one can try different weights until he finds a good fit (in the heavy-tailed case, a good strategy might be weighting the points that are close to the mean of the xgrid less than those that are far from the mean of the xgrid). We initially tried to use regression approach to solve for π . Although ϕ is highly nonlinear, we found $\Phi^T \Phi$, where Φ is a k by m with entries ϕ_{ij} , to be singular especially when $m \geq 7$.

Instead of using standard regression techniques, we considered the problem as a quadratic programming problem with two constraints: $\sum \pi_j = 1$ and $\pi_j \geq 0$ for all j . In what follows, we expand $S(\pi)$ and reformulate the problem in a matrix environment:

$$\begin{aligned} S(\pi) &= \sum_{i=1}^k \left[w_i^2 y_i^2 - 2w_i y_i \sum_{j=1}^m \phi_{ij} \pi_j + w_i^2 \left(\sum_{j=1}^m \phi_{ij} \pi_j \right)^2 \right] \\ &= \sum_{i=1}^k w_i^2 y_i^2 - 2 \sum_{i=1}^k \left(w_i y_i \sum_{j=1}^m \phi_{ij} \pi_j \right) + \sum_{i=1}^k \left(\sum_{j=1}^m w_i \phi_{ij} \pi_j \right)^2 \\ &= \sum_{i=1}^k w_i^2 y_i^2 - 2 \sum_{j=1}^m \left(\sum_{i=1}^k w_i y_i \phi_{ij} \right) \pi_j + \sum_{i=1}^k \sum_{j=1}^m \sum_{l=1}^m (w_i \phi_{ij} \pi_j) (w_i \phi_{il} \pi_l). \end{aligned}$$

Since $\sum_{i=1}^k w_i^2 y_i^2$ is independent of π , it is a constant. Let g be the m by 1 vector defined as $g = \left(-\sum_{i=1}^k w_i y_i \phi_{i1}, \dots, -\sum_{i=1}^k w_i y_i \phi_{im} \right)^T$ and H be an m by m matrix defined as

$$H = \begin{bmatrix} \sum_{i=1}^k w_i^2 y_i \phi_{i1} \phi_{i1} & \sum_{i=1}^k w_i^2 y_i \phi_{i1} \phi_{i2} & \cdot & \cdot & \sum_{i=1}^k w_i^2 y_i \phi_{i1} \phi_{im} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^k w_i^2 y_i \phi_{im} \phi_{i1} & \sum_{i=1}^k w_i^2 y_i \phi_{im} \phi_{i2} & \cdot & \cdot & \sum_{i=1}^k w_i^2 y_i \phi_{im} \phi_{im} \end{bmatrix}.$$

Then $S(\pi) = 2 \left[c + g^T \pi + \frac{1}{2} \pi^T H \pi \right]$ where $c = 2 \sum_{i=1}^k w_i^2 y_i^2$ is constant. Hence, π can be found by minimizing $\left[g^T \pi + \frac{1}{2} \pi^T H \pi \right]$, subject to $\sum_{j=1}^m \pi_j = 1$ and

\mathbf{r}	$p(\mathbf{r})$
1	0.3568075038983
1.1	0.1751458056112
4.1	0.0692394259209
4.2	0.3988072645695

Table 1: Recovered Mixing measure using UNMIX . Note: The exact mixing measure is concentrated at 1 and 4.

$$A\pi \geq \mathbf{b}, \text{ where } A = \begin{bmatrix} 1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 1 \end{bmatrix} \text{ of order } m \times m \text{ and } \mathbf{b}^T = (0, \dots, 0)$$

order m . The quadratic programming routine, QPROG, from the International Mathematics and Statistics Library, IMSL, is employed and modified to fit the current problem. We have called this the new program UNMIX. The program requires a grid of x points, which is called `xgrid`, the estimated density of the `xgrid` using any estimate of the density of the sample X , which is named `yhat`, a grid of R points called `rgrid`, and a vector of weights with same length as `xgrid`. The default weights are a vector of ones. The output is the vector π that minimizes $S(\pi)$. The examples from the previous section are revisited, but the mixing measure is estimated using the UNMIX program.

EXAMPLE 2 *Let X be a mixture of a normal(0, 1) and a normal(0, 4) with weights $\pi_1 = P(R = 1) = .5$ and $\pi_2 = P(R = 4) = .5$. The estimated mixing measure using the UNMIX with $n = 2000$, $xgrid = (.1, .2, \dots, 10)$, $rgrid = (.1, .2, \dots, 5)$ and weights $= (1, \dots, 1)$. The recovered mixing measure is given in Table 1.*

EXAMPLE 3 *The generalized t distribution is a scale mixtures of normals where $1/R^2$ is Gamma(α, β). In particular, when $\alpha = \frac{1}{2}$ and $\beta = 2$, X is a standard Cauchy random variable. To estimate the distribution of R , a random sample of size $n = 1000$ is generated from a standard Cauchy. Then UNMIX is used with $rgrid = (.1, .3, \dots, 19.9)$, $xgrid = (.1, .3, .5, \dots, 29.9)$, the default weights and the estimated density at the `xgrid`. The normal kernel smoother is used to estimate the density of X . The estimated mixing measure is displayed in Table 2.*

In general, one needs to find the mixing measure that will provide a good fit of the data. In our case, the estimated mixing measure from UNMIX method will be used to model the data. Then the fitted density of X using the estimated mixing measure is compared with the exact density. In practice, the exact density of X is unknown, but one can use the estimated density of X via a

r	$p(r)$
.1	0.00333377176422
.3	0.00448111878394
.5	0.04177235490862
.9	0.29319314571840
1.1	0.24252281603712
3.3	0.25933886343956
3.5	0.25933886343956
19.9	0.11570533923886

Table 2: The recovered mixing measure using UNMIX for the Cauchy example.

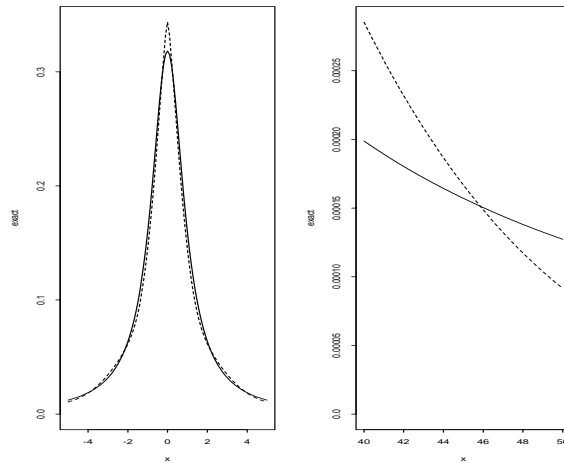


Figure 6: The exact density of X (solid) and the fitted density using the estimated mixing measure with the UNMIX method.

kernel smoother. Let $\hat{f}(x)$ be the estimated density of the sample and

$$\hat{f}(x|\hat{\pi}) = \sum_{i=1}^m \frac{1}{r_i} \phi\left(\frac{x_i}{r_i}\right) \hat{\pi}_i.$$

Then, we consider the fit a good one if $\hat{F}(x)$ is close to $\hat{F}(x|\hat{\pi})$, where

$$\hat{F}(x|\hat{\pi}) = \sum_{i=1}^m \frac{1}{r_i} \Phi\left(\frac{x_i}{r_i}\right) \hat{\pi}_i.$$

For the Cauchy example, the UNMIX method discretized the mixing measure using the provided rgrid. The estimated mixing measure is given in Table 2. The exact Cauchy and the fitted density are evaluated on equally spaced grids on the interval $[-5, 5]$ and on the interval $[40, 50]$. The results are shown in Figure 6. The reason why it does not do a good job in far the tails is because $X \stackrel{d}{=} RZ$; large values of x 's come from large values of R , and, since we prohibited large values of R , we cannot have a good estimate of the density of X in the tails. Of the recovered distribution of R 11% is at 19.9 and 0 beyond that, so the fitted probability of a large x is smaller than the exact probability. To improve the fit in the tail, one can take larger values of R in the rgrid. However, since we cannot estimate the density far in the tails using a kernel smoother with the high accuracy, we will never have a good fit far in the tails as illustrated in Figure 6. When we used the UNMIX with the exact density of X and the mixing measure is discrete with three point masses, the recovered weights are almost exact. However, for the same example, when the estimated density of X using a normal kernel smoother is used, the UNMIX mislocate the point masses and only was able to recover two rather than three.

6 Conclusion

In general, when infinite mixture doesn't have a closed form or hard to compute, it can be approximated with high accuracy. In particular, if the interest is in the body of the mixture one can use Theorem 2 with a small a value. However, if we are interested in the tails of the mixture, we can truncate the mixing measure at a larger b .

In practice, the necessary and sufficient conditions provided to verify whether a certain random variable is a scale mixture of normals are hard to apply, especially if we are looking at a sample. We would like to simplify these conditions or provide a better way based on a sample rather than the form of the density and be able to tell whether the sample came from a scale mixture.

A new method for estimating the mixing distribution was introduced. This method is called UNMIX and it is based on minimizing the squared weighted distance between the estimated density and the fitted density. The estimated density is found using a kernel smoother over a fixed grid of X values. The fitted density is found by discretizing the mixing measure over a fixed grid of R

values called rgrid. This method seems to be practical for fitting and modelling data and for recovering discrete mixing measures.

The new method has potential for improvement, one possibility is to improve the density estimate used by modelling the upper tail by a specific model, e.g. by a Pareto density. Also, one can expand r-grid using larger and larger spacings.

References

- [1] Andrews, D.R. and Mallows, C.L. (1974). Scale mixtures of normal distributions. *J.R Statist. Soc.*, **36**, 99-102.
- [2] Barndorff-Nielsen, O., Kent, J., and Sorensen, M. (1982). Normal Variance-Mean Mixtures and z Distributions. *Int. Statist. Rev.* , **50**, 145-159.
- [3] Beale, E. M. L. and Mallows (1959). Scale mixing of symmetric distributions with zero means. *Ann. Math. Statist.*, **40**, 1145-1151.
- [4] Box, G. E. P. and Tiao, G.C (1973). *Bayesian Inference in Statistical Analysis*, Reading, Mass: Addison-Wesley.
- [5] Bozdogan, H, Opitz, B.Lausen, and R.Klar (1993). Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix. *In Studies in Classification, Data Analysis, and Knowledge Organization*, Springer-Verlag, Heidelberg.
- [6] Byczkowski, T., Nolan, J. P., and Rajput, B. (1993). Approximation of multidimensional stable densities, *J. of Multivariate Anal.*, **46**, 13-31.
- [7] Deely, J.J., and Kruse, R. (1968). Construction of sequences estimating the mixing distribution. *The Annals of Statist.* **39**, 286-288.
- [8] Dempster, A. P., Larid, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **39**, 1-38.
- [9] Diks, C., Tong, Howell (1999). A test for symmetries of multivariate probability distributions. *Biometrika Trust* , **86**, 605-614.
- [10] Feller, W. J. (1971). *An Introduction to Probability and Its Applications, Vol. II, 2nd Edition*. Wiley, NY.
- [11] Fujikoshi, Y., and Shimizu, R. (1989). Error Bounds for Asymptotic Expansions of Scale Mixtures of Univariate and Multivariate Distributions. *J. Multivariate Anal.*, **30**, 279-291.
- [12] Hamdan, H and Nolan J. P. (2004). Approximating Scale Mixtures. *Stochastic Processes and Functional Analysis.*, **238**, 161-169.

- [13] Deely, J.J., and Kruse, R. (1968). Construction of sequences estimating the mixing distribution. *The Annals of Statist.*, **39**, 286-288.
- [14] Larid, Man. (1978). Nonparametric maximum likelihood estimate of a mixing distribution. *Journal of the American Statistical Ass.* **73**, 805-811.
- [15] Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry, and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5. IMS, Haywood, CA.
- [16] Pribe, E., Carey. (1994). Adaptive Mixtures. *Journal of the American Statistical Ass.*, **89**, 796-806.
- [17] Redner, R. A., and walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, **26**, 195-239.
- [18] Rohatgi, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. Wiley, NY
- [19] Romanowski, M. (1979). *Random Errors in Observations and the Influence of Modulation on Their Distribution*. Stuttgart: Verlag Konrad Witter.
- [20] Samorodnitsky, G., and Taqqu, M. (1994). *Stable Non-Gaussian Random Processes*, New York: Chapman and Hall.
- [21] Schoenberg, I. J. (1938). Metric spaces and completely monotonic functions. *Annals of Math.*, **39**, 811-841.
- [22] Shimizu, R. (1995). Expansion of the Scale Mixture of the Multivariate Normal Distribution with Error Bound Evaluated in the L_1 - Norm, *J. of Multivariate Anal.*, **53**, 126-138.
- [23] Springer, M. D. (1979). *The Algebra of Random Variables*. Wiley, NY.
- [24] Titterington, D. M., Smith, A. F. M., Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, NY.
- [25] West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, **74**, 646-648.
- [26] Zhang, Cun-Hui. (1990). Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics.*, **18**, 806-831.