

Estimating the Parameters of Infinite Scale Mixtures of Normals

Hasan Hamdan

and John Nolan

36th Symposium on Interface:
Computing Science and Statistics

May 26 -9

Baltimore, Maryland

An Outline of the Presentation

- Motivation, Definitions, General Problem
- Variance Mixtures of Normals (VMN)
 1. Examples of Variance Mixtures in \mathbb{R} and in \mathbb{R}^n
 2. Characterization Theorem
 3. Approximation Theorem
- Estimating the mixing measure
- Further Research

Motivation

- Identify and simplify infinite mixtures of normals, uniforms, and exponentials.
- Approximate infinite mixtures with finite mixtures.

Simpler forms.

Closed form.

Easier to study properties.

Variance Mixture of Normals

- A random variable X is a variance mixture of normals if $X \stackrel{d}{=} AZ$, where $Z \sim N(0, 1)$, A is a random scale, with A and Z independent. We assume $P(A = 0) = 0$.
- Equivalently, X has pdf $f(x) = \int_0^\infty g(x|\sigma)\pi(d\sigma)$, where $g(x|\sigma)$ is the $N(0, \sigma^2)$ density and the mixing measure π is the distribution of A .
- Equivalently, the characteristic function $\phi_X(t)$ of X can be written in the form

$$\phi_X(t) = \int_0^\infty \phi_{\sigma Z}(t)\pi(d\sigma),$$

where $\phi_{\sigma Z}(t)$ is the characteristic function of the random variable $\sigma Z \sim N(0, \sigma^2)$.

Examples in \mathbb{R} and \mathbb{R}^n

1. Symmetric stable distributions

A stable random variable X with index of stability $\alpha \in (0, 2]$, scale parameter $\sigma \in (0, \infty)$, skewness parameter $\beta \in [-1, 1]$ and location parameter $\mu \in (-\infty, \infty)$ is denoted by $S_\alpha(\sigma, \beta, \mu)$. The characteristic function

$$\phi_X(u) =$$

$$\begin{cases} \exp\left(-\sigma^\alpha |u|^\alpha \left[1 - i\beta \tan\left(\frac{\pi}{2}\alpha\right) s(u)\right] + i\mu u\right) & \alpha \neq 1 \\ \exp\left(-\sigma |u| \left[1 + i\beta \frac{2}{\pi} s(u) \ln(|\sigma u|)\right] + i\mu u\right) & \alpha = 1 \end{cases},$$

where $s(u) = \text{sign}(u)$.

Suppose that $X \sim N(0, 2\sigma^2)$, A is positive stable $S_{\alpha/2}((\cos(\pi\alpha/4))^{2/\alpha}, 1, 0)$, and A and X are independent. Then $W = A^{1/2}X$ is symmetric α -stable ($S_\alpha S$) with scale σ .

Sub-Gaussian random vectors

Choose $A \sim S_{\frac{\alpha}{2}} \left(\left(\cos \left(\frac{\pi\alpha}{4} \right) \right)^{\frac{2}{\alpha}}, 1, 0 \right)$ with $\alpha < 2$.

Let $\mathbf{G}' = (G_1, \dots, G_n) \sim N(0, \Sigma)$ independent of A .

Then, $\mathbf{X}' = (A^{\frac{1}{2}}G_1, \dots, A^{\frac{1}{2}}G_n)$ is $S\alpha S$ in \mathbb{R}^n with

$$\phi_n(\boldsymbol{\theta}) = \exp \left(- \left(\frac{|\boldsymbol{\theta}'\Sigma\boldsymbol{\theta}|}{2} \right)^{\frac{\alpha}{2}} \right).$$

For example, when $n = 2$, $\alpha = 1$, and \mathbf{G} iid $N(0, 2\sigma^2)$

$$\phi_2(\theta_1, \theta_2) = \exp(-\sigma (\theta_1^2 + \theta_2^2)^{1/2})$$

and $f(x_1, x_2)$ is the spherically symmetric Cauchy density in \mathbb{R}^2 .

2. Generalized t distributions

Suppose that $1/A^2$ has a $\text{Gamma}(\alpha, \beta)$ distribution.

Equivalently,

$$f_A(\sigma) = \frac{2}{\beta^\alpha \Gamma(\alpha)} \frac{1}{\sigma^{2\alpha+1}} \exp\left(-\frac{1}{\beta\sigma^2}\right).$$

Set the scale parameter $\beta = 2/c$. Then the density function of $X = AZ$ is given by

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

where $k = \frac{2^\alpha \Gamma(\alpha + \frac{1}{2})}{\pi^{1/2} \beta^\alpha \Gamma(\alpha)}$.

When $\alpha = n/2$ and $\beta = 2/n$, $f(x)$ is the t density with n degrees of freedom.

Multivariate t

If the mixing density is given by

$$f_A(\sigma) = \frac{2}{\beta^\alpha \Gamma(\alpha)} \frac{1}{\sigma^{2\alpha+1}} \exp\left(-\frac{1}{\beta\sigma^2}\right)$$

and $\mathbf{X} \sim N(0, I)$, then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{k_1}{(k_2 + \mathbf{x}'\mathbf{x})^{\alpha + \frac{n}{2}}},$$

where $k_1 = \left(\frac{2}{\beta}\right)^\alpha \frac{\Gamma(\alpha + \frac{n}{2})}{\pi^{\frac{n}{2}} \Gamma(\alpha)}$ and $k_2 = \frac{2}{\beta}$ are constants.

In particular, when $\alpha = \frac{n}{2}$, $f_{\mathbf{X}}(\mathbf{x})$ is the multivariate SS Cauchy density in \mathbb{R}^n .

Characterization Theorem

Definition

A function $h(x)$ on $(0, \infty)$ is *completely monotone* in x if it is infinitely differentiable and $(-1)^m h^{(m)}(x) \geq 0 \forall x$ and $\forall m = 1, 2, \dots$.

Examples are $\frac{1}{x}$, $\frac{1}{x+1}$, and $\exp(-x)$.

Theorem 1 (Schoenberg (1938))

X with density $f(x)$ is a *VMN* iff $h(x) = f(x^{1/2})$ is a completely monotone function.

Equivalently, X is a *VMN* iff ϕ_X is a real, even function such that $\phi_X(t^{1/2})$ is completely monotone on $(0, \infty)$.

Example

Exponential Power Family

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

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

See West (1987) and Box and Tiao (1973).

A random variable X with density $f(x)$ is a variance mixtures of normals iff $0 < b \leq 2$.

$h(x) = f(\text{sqrt}(x)) = k \exp(-x^{b/2})$ is completely monotonic iff $0 < b \leq 2$.

Approximating Scale Mixtures

Case 1: $A \in [a, b]$ where $0 < a < b < \infty$.

X with density $f(x)$ is a mixture of normals with known scale A having distribution π .

If $f(x)$ is difficult to compute, then we can approximate it by a finite mixture of the form

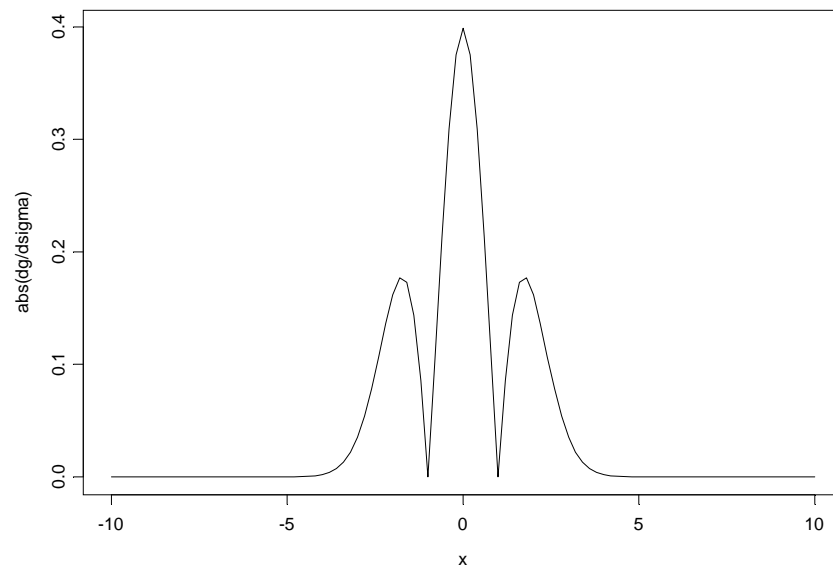
$$f^*(x) = \sum_{j=1}^M g(x|\sigma_j)\pi_j,$$

where π_1, \dots, π_M are point masses concentrated on $\sigma_1, \dots, \sigma_M$ in $[a, b]$.

Questions

- How many terms should we take to approximate $f(x)$ by $f^*(x)$ within ϵ ?
- What values of π_j and σ_j should we choose?

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



Lemma 1

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

$$|g(x|\sigma_1) - g(x|\sigma_2)| \leq \frac{1}{(2\pi) a^2} |\sigma_1 - \sigma_2| \quad \forall x \in \mathbb{R},$$

where $g(x|\sigma)$ is $N(0, \sigma^2)$.

Proof.

Fixing σ , $|\partial g(x|\sigma)/\partial \sigma| = \left| \frac{x^2 - \sigma}{\sigma^2} \right| g(x|\sigma)$ is maximized at $x = 0$, where it takes value

$$g(0|\sigma)/\sigma = 1/((2\pi)^{1/2} \sigma^2).$$

Hence,

$$\begin{aligned} |g(x|\sigma_1) - g(x|\sigma_2)| &\leq (\max |\partial g/\partial \sigma|) |\sigma_1 - \sigma_2| \\ &= |\sigma_1 - \sigma_2| / ((2\pi)^{1/2} a^2). \quad \blacksquare \end{aligned}$$

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

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

Proof. We adapted Lemma 1 from Byczkowski, Nolan, and Rajput (1993).

- Fix any $\epsilon > 0$, and $0 < a < b < \infty$.
- Define recursively.

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

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

- 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 $\sigma_j = \min(a_{2j-1}, b)$, $j = 1, \dots, M$.
- $g(x|\sigma_j)\pi_j = g(x|\sigma_j) \int_{I_j} \pi(d\sigma)$. Then,

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

■

Case 2: $A \in (0, \infty)$.

We can write $f(x)$ as a sum of three integrals.

$$\int_0^\infty g(x|\sigma)\pi(d\sigma) = \int_0^a (\cdot) + \int_a^b (\cdot) + \int_b^\infty (\cdot). \quad (3)$$

The following lemma shows that in all cases where $f(0)$ is bounded, there exists an a and b such that the first and last integrals can be made arbitrary small and the middle can be approximated using Theorem 3.

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

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

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

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

Proof.

$$f(x) = \int_0^\infty g(x|\sigma)\pi(d\sigma) \leq k \int_0^\infty \sigma^{-1}\pi(d\sigma) = f(0) < \infty.$$

Let $h(a) = \int_0^a g(x|\sigma)\pi(d\sigma)$. Then,

$$\begin{aligned} h(a) &\leq k \int_0^a \sigma^{-1}\pi(d\sigma) \\ &= k \int_0^\infty 1_{(0,a)}\sigma^{-1}\pi(d\sigma). \end{aligned}$$

Let a_n be any sequence that converges to 0.

Then

$1_{(0,a_n)}\sigma^{-1} \rightarrow 0$ pointwise on $(0, \infty)$ and $1_{(0,a_n)}\sigma^{-1} \leq \sigma^{-1} \in L^1(\pi)$. So, $h(a_n) \rightarrow 0$ by the Dominated Convergence Theorem.

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

Proof.

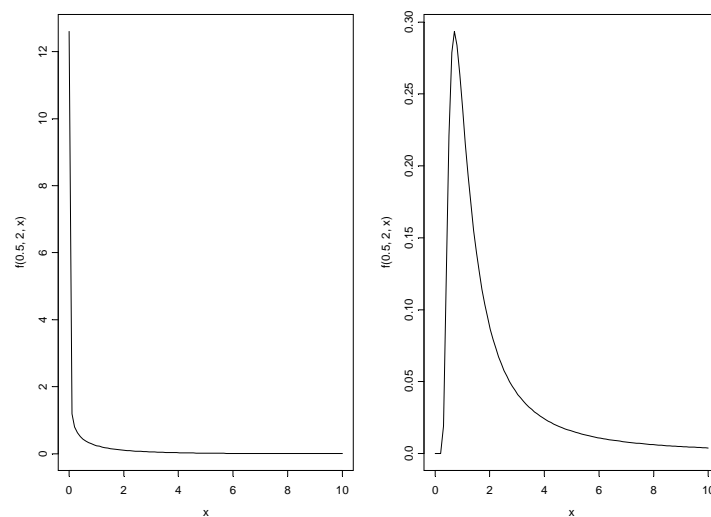
Let $h(b) = \int_b^\infty g(x|\sigma)\pi(d\sigma)$. Then,

$$\begin{aligned} h(b) &\leq k \int_b^\infty \sigma^{-1} \pi(d\sigma) \\ &\leq k \int_0^\infty \mathbf{1}_{(b, \infty)} \sigma^{-1} \pi(d\sigma). \end{aligned}$$

Let b_n be any sequence that converges to ∞ .

Then $\mathbf{1}_{(b_n, \infty)} \sigma^{-1} \rightarrow 0$ and since the last expression is dominated by $\frac{1}{b}$, the result holds by applying the Dominated Convergence Theorem.

Figure 2: Gamma and square root of Inverted Gamma with $\alpha = .5$ and $\beta = 2$.



Approximating the Cauchy Density

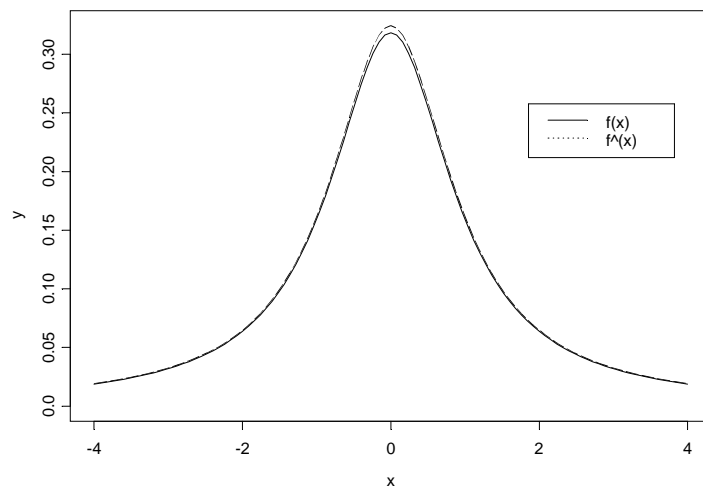
When $\alpha = \frac{1}{2}$ and $\beta = 2$, the generalized t distribution is the standard Cauchy. π is the square root of Inverted Gamma with parameters α and β . In this case, the corresponding Gamma has a vertical asymptote at 0 and it is decreasing on $\Theta = [a, b]$.

Example

A comparison between the finite and infinite mixture is made for different combinations of a , b , and ϵ .

The maximum difference between the actual density and the approximated density were found based on $a = .05$, $b = 50$, and $\epsilon = .03$ on a grid of 101 equally spaced points. The maximum value for the relative distance between f and f^* is around .028.

Figure 3: $a = .05, b = 50$ and $\epsilon = .03$.



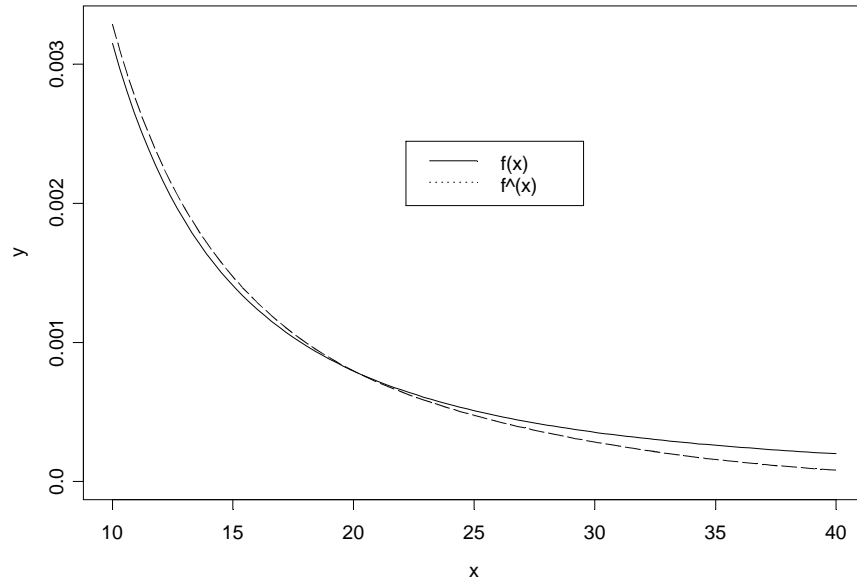


Figure 4: $a = .05$, $b = 50$, and $\epsilon = .03$.

Approximating the Cauchy Density

However, the approximation is not that good in the tails. The maximum relative distance is around .17.

The number of terms, M , used in the approximation process was found to be 31, which is considerably large.

Estimating the Mixing Measure

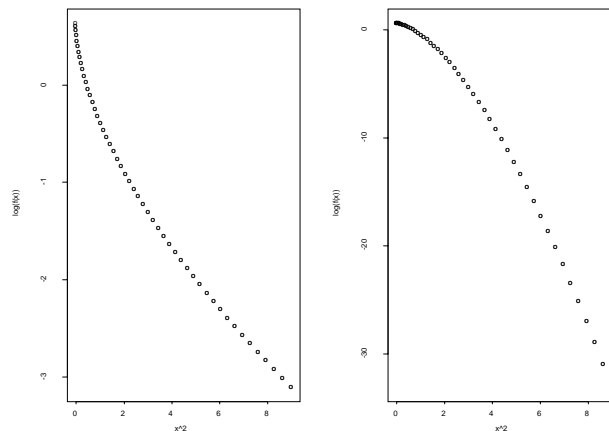
1. Diagnostics

How do we know that a given random sample can reasonably be assumed to come from some scale mixtures of normals? Here are some suggestions:

Check the unimodality

Check the symmetry

Check the log/square plot where $\log f(x)$ is plotted as a function of x^2 .



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

2. UNMIX method

We minimize the sum of the squared weighted distances between the estimated density of X and the corresponding density computed by discretizing the mixture over a pre-determined grid of R values, r_1, \dots, r_m and a grid of X values, x_1, \dots, x_k , where $k \geq m$.

For each x_i in the xgrid, $f(x_i)$ is estimated by $\hat{f}(x_i)$ using a kernel smoother. If we let $y_i = \hat{f}(x_i)$, then

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

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

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

subject to $\sum_{j=1}^m \pi_j = 1$ and $\pi_j \geq 0$. The quadratic programming routine, QPROG, from the International Mathematics and Statistics Library, IMSL, is employed and modified to fit the current problem.

r	$p(r)$
1	0.3568075038983
1.1	0.1751458056112
4.1	0.0692394259209
4.2	0.3988072645695

Table 1: Recovered Mixing measure using UNMIX .

Recommendations

One way to improve this approximation around 0 is to take a smaller a . But that will increase M because the a -sequence will have many terms near the origin. Similarly, to improve the approximation in the tails, one can truncate the mixing measure at a larger b .

To reduce the number of terms, one can try to eliminate the terms that have small weights. Although in most examples the same tolerance is maintained, it is not guaranteed that it will work for all cases.

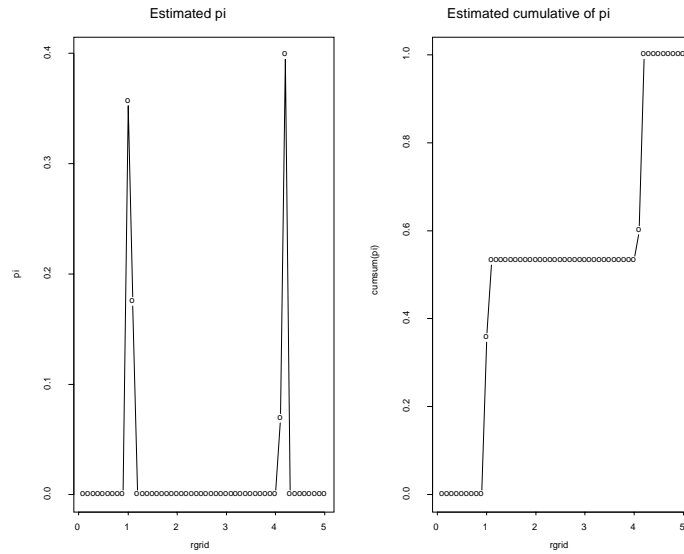


Figure 5: The estimated mixing measure using the UNMIX with $n=2000$. The exact mixing measure is concentrated on $r = 1$ and $r = 4$ with equal probability.

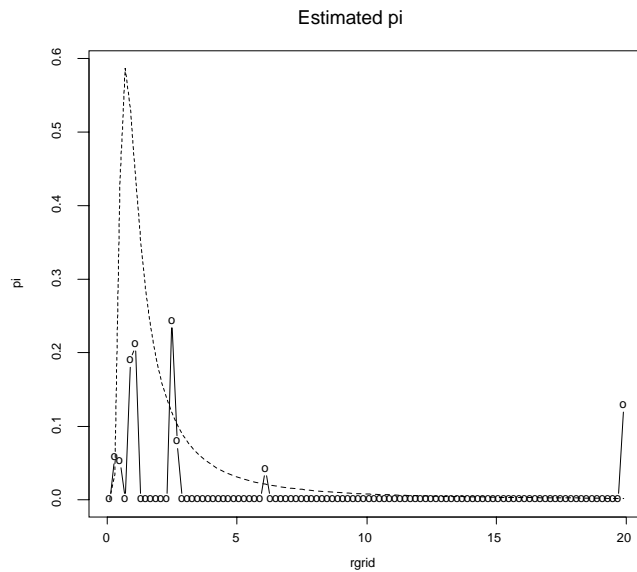


Figure 6: The estimated density of the mixing measure with $n=1000$ using the UNMIX (Cauchy example).

The solid line is the estimated discrete mixing measure and the dotted curve is the exact one.

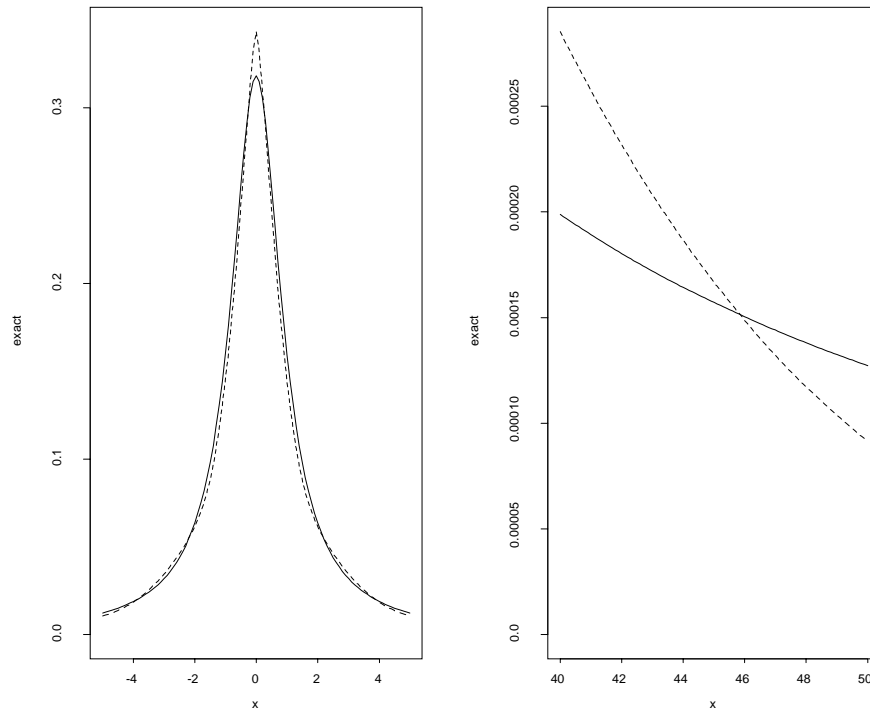


Figure 7: The recovered mixing measure is used to approximate the infinite mixture by a finite mixture. The solid line is the exact density and the dotted line is the estimated density based on the recovered weights.

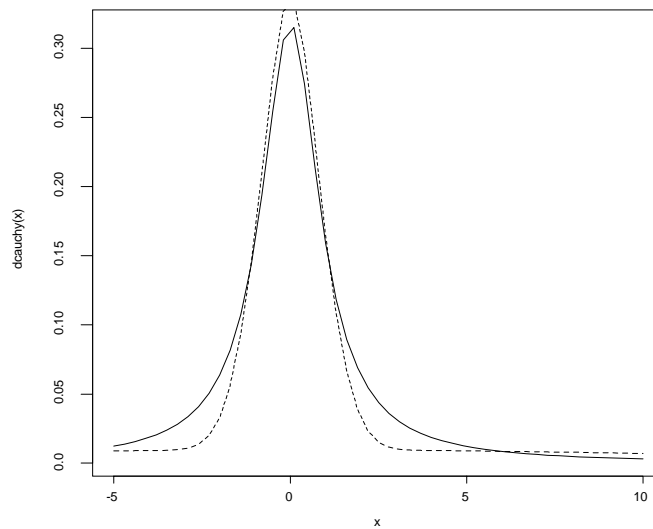


Figure 8: The exact density of one component of a bivariate Cauchy (Solid) and the corresponding estimated density using UNMIX.

Further Research

- Find ways to reduce the number of components.
- Approximate multivariate scale mixtures of uniforms and exponentials.
- Compare the estimated mixing measure by UNMIX with other existing methods such as the EM algorithm.
- Improve the estimated mixing measure by UNMIX using different density estimates in the tails.
- Extend the UNMIX to scale mixtures of uniforms and exponentials.