Journal of Statistical Research *ISSN 0256 - 422 X*
2018, Vol. 52, No. 2, pp. 129-150

ESTIMATING VARIANCE-MEAN MIXTURES OF NORMALS

HASAN HAMDAN*
Department of Mathematics and Statistic
James Madison University, Harrisonburg, VA 22807, USA
Email: hamdanhx@jmu.edu

LING XU
Department of Mathematics and Statistic
James Madison University, Harrisonburg, VA 22807, USA
Email: xulx@jmu.edu

SUMMARY

A new semi-nonparametric method for modeling data with Normal variance-mean mixtures (NVMM) is presented. This new method is based on the least-squares programming routine, UNMIX. Density estimates based on random samples of size n from two, three, and four components of NVMM are found using UNMIX. Graphical comparisons of the UNMIX fit found by the EM Algorithm and the Bayesian approaches are done using three real life examples. A quantitative comparison using the AIC and Chi-Square is done for one of the most commonly used examples, the Galaxy Data. The results are promising and the method has great potential for improvement.

Keywords and phrases: Normal variance-mean mixture, scale mixtures of normals, UNMIX, weighted least squares minimization

AMS Classification: 62

1 Introduction

Normal variance-mean mixtures, NVMM, are used to study data in a variety of applications. Environmental applications include sedimentology, biometrics, ecology, and agriculture (Gomex, et al., 2007; Bhattacharya, 1967; Chen, et al., 2004; Ling, et al., 2010; and Holling, 1992). In economics, NVMM are often used in representing uncertain volatility (Alexander, 2004). These distributions are also used in image processing through remote sensing and signal detection (Permuter, et al., 2003; Titterington, 1985).

Estimating the NVMM involves three sets of parameters (means, variances, and weights) besides the number of components. Determining and assessing the number of components in a mixture is very important but it is a difficult problem; some of the recent relevant works can be found in Kasahara and Shimotsu (2013). They developed a procedure to estimate a lower bound on the number of components of a k-variate mixture. The bound is based on the rank of a matrix constructed from

[*] Corresponding author

the distribution function of the observed variable. Kasahara and Shimotsu (2014) extended the EM approach of Li et al. (2009) to test the null hypothesis of m_0 components against the alternative of m_0+1 components. Belomestny and Panov (2018) provided a method for estimating the mean of the underlying normal density and then non-parametrically recovering the density of the corresponding mixing distribution.

A more general overview of the available literature and published results can be found in Oliveira-Brochado and Martins (2005). They divided the published criteria into five groups depending on their theoretical backgrounds. The five groups are criteria based on hypothesis test, information criteria, classification criteria, minimum-information ratio, and other criteria. The other criteria group covers approaches such as graphical diagnosis tools and cross-validation.

McLachlan and Peel (2000) discussed the number of components and the progress made up to that time. Although the number of modes does not necessarily determine the number of mixture components, it can be a good start when they are present. See Titterington et al. (1985), for a thorough discussion of the modality and the number of components of a mixture.

The motivation for our approach is providing an alternative, easy to use, fitting tool or fitting algorithm similar to the FMM SAS procedure, that relies on the EM (Expectation-Maximization) algorithm. There are two main approaches for estimating the NVMM densities and each has its own limitations. The first and most commonly used approach is the EM. In spite of all the research that has been done in this area, there are still two issues. The first one is that when a few parameter values are close (e.g two means are close to each other or two variances are close to each other), the algorithm may identify a local maximum as a global maximum and the likelihood function will converge to the wrong value. The second problem is that the initialization step is still somewhat arbitrary. A far as we know, there is still no clear answer yet on how to initialize the values. The second approach is the Bayesian approach which also has some limitations. The prior selection is still somewhat arbitrary (See the Eyes example below), the predictive density is not easy to use, and the popularity of this approach is still limited, which sometimes makes it difficult to communicate and disseminate.

The proposed approach estimates the weights of the components assuming that the parameters of the individual components are known. These known parameters are called the μ-grid and the σ-grid. Therefore, if the size of the μ-grid is q and that of the σ-grid is m, the total number of unknown parameters is $qm - 1$ since the total weight must be 1. More details will be shown on Section 2.

This paper focuses only on the univariate variance-mean mixtures of normals assuming that the mean and variance are independent. The multivariate variance-mean mixtures is more involved and should be treated separately, even with a simple dependent structure such as the one presented in Barndorff-Nielsen, et al. (1982). The latter example shows how quickly the size of the parameter space can grow and how fast it can become computationally difficult.

The rest of this paper is organized as follows: The next section introduces normal variance-mean mixtures and presents a few examples and classes. Then, a new method for estimating the NVMM density is introduced, along with a brief overview of the two existing estimation methods. In Section 4, the new method of estimating the NVMM is applied to a few simulated examples with varying sample sizes. A comparison and analysis of the new method for estimating NVMM versus the

Bayesian estimates and the EM estimates with real life data are found in Section 5. Section 6 shows how to estimate the standard error using the bootstrap samples. Finally, in Section 7, the conclusion and some recommendations are provided. All programs in this study are written in R.

2 Normal Variance-Mean Mixtures

The following is a general definition of normal variance-mean mixtures.

Definition 2.1. A random variable X is said to be a normal variance-mean mixture or NVMM when its probability density function is given by $f_X(x)$, where

$$f_X(x) = \int_{-\infty}^{\infty} \int_{0}^{\infty} \phi\,(x, \mu, \sigma)\, \pi_1(d\sigma)\pi_2(d\mu), \tag{2.1}$$

where $\phi()$ is a normal density with mean $\mu \in \mathbb{R}$, standard deviation $\sigma > 0$, and π_1 and π_2 are independent mixing measures over σ and μ respectively. Equivalently, $X \overset{d}{=} AZ + \mu$, where $A > 0$ is a positive random variable with distribution π_1, Z is the standard normal random variable, and μ is a continuous random variable with distribution π_2 defined on the real line. Furthermore, A, Z, and μ are independent.

The following are general classes of this type of mixture.

Example 2.1. (The Z Distribution). The Z distribution has four parameters α, β, σ and μ. Its density with respect to the Lebesgue measure is given by

$$f_X(x) = \frac{\left(\exp^{(x-\mu)/\sigma}\right)^{\alpha}}{\sigma B(\alpha, \beta)\left(1 + \left(\exp^{(x-\mu)/\sigma}\right)\right)^{\alpha+\beta}},$$

where x, $\mu \in \mathbb{R}$, $\alpha, \beta, \sigma > 0$, and B is the Beta constant. The Z distribution appeared for the first time in Fisher 1921 and Fisher 1935. Much later, Barndorff-Nielsen et al. (1982) showed that the Z distributions are normal variance-mean mixtures and specified their mixing distributions. If X is Beta with parameters α and β then $\log(Z/(1 - X))$ is $Z(\alpha, \beta, 1, 0)$. Therefore, if X is F with degrees of freedom f_1 and f_2, then $\log(X)$ is $Z(\frac{1}{2}f_1, \frac{1}{2}f_2, 1, \log(f_2/f_1))$.

Example 2.2. (Pearson Type VII Distribution). In equation (2.1), if the scale parameter has a square root of inverse gamma distribution with parameters α and β, and μ is a constant, that is if π_1 is the square root of a reciprocal of a gamma r.v. and π_2 is degenerate at one point, say $\mu = \mu_0$, then X is a Pearson Type VII distribution.

Pearson distributions are a class of 12 curves generated by the solutions of the differential equation $\frac{d\ln(f(x))}{dx} = \frac{x+a}{b+cx+dx^2}$, where $a, b, c,$and d are constants. Pearson Type VII contains all variance mixtures of normals with possible shifts such as the Cauchy and the t distribution. See Johnson, Kotz, and Balakrishnan(1994, pp.15-25) and Johnson, Kotz, and Balakrishnan(1995, p.396). Because of the size and the importance of this class of mixtures, we will present it in detail in the following example.

Example 2.3. (Scale Mixtures of Normals). This is a special case of Example 2.2 i.e the scale parameter has a square root of inverse gamma distribution with parameters α and β, and μ is known to be 0. In this case, the general formula for an infinite normal scale mixture is

$$\begin{aligned} f(x) &= \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\big(- x^2/2\sigma^2 \big)\pi(d\sigma) \\ &= \int_0^\infty \phi(x, 0, \sigma)\pi(d\sigma) \\ &= \int_0^\infty \frac{1}{\sigma}\phi(x/\sigma, 0, 1)\pi(d\sigma), \end{aligned} \tag{2.2}$$

where $\phi(x)$ is the standard normal density and π is the mixing distribution (or the weighting distribution).

There are many common examples of scale mixtures of normals. Some of them are finite, such as the contaminated normal; others are infinite, like the t distribution and the Laplace distribution. For more examples, see Hamdan and Nolan (2004). Yao and Taimre (2016) presented an algorithm for estimating tail probabilities of these distributions. Their algorithm combines the importance sampling and conditional Monto Carlo methods.

When the mean is a constant, as mentioned in Example 2.2, the scale mixture is a generalized t distribution. The discretized (or approximated) version of (2.2) is given by

$$f(x) = \sum_{i=1}^{m} \pi_i \phi(x, 0, \sigma_i). \tag{2.3}$$

Scale mixtures can be used to model heavy-tailed data. They can be used as good alternatives to symmetric stable distributions which are often used in finance. A commonly used example of (2.3) is the contaminated normal, with $m = 2$ and $\sigma_2 > \sigma_1$. Similarly, the discretized version of (2.1) is denoted by $f^\star_X(x)$, where

$$f^\star_X(x) = \sum_{a=1}^{q} \sum_{b=1}^{m} \pi_{ab} \phi\,(x, \mu_a, \sigma_b). \tag{2.4}$$

Also in the general case, $f^\star_X(x)$ is more practically useful compared to $f_X(x)$. In modeling data with $f^\star_X(x)$, like in any other modeling question, a random sample of size n is available and the goal is to estimate the parameters: μ_a, σ_b, and π_{ab}, where $a = 1, \dots, q$, and $b = 1, \dots, m$. That means a total of $q + m + qm - 1$ parameters are to be estimated (assuming that q and m are known, which is not usually the case) and it is one less than the total number of parameters because $\sum_{a=1}^{q} \sum_{b=1}^{m} \pi_{ab} = 1$. Therefore and since q and m are also unknown, there are $q + m + qm + 1$ parameters that need to be estimated. However, in our case the μ_a, σ_b, q, and m are assumed and only the π_{ab} need to be estimated. That means the total of unknown parameters is $(qm - 1)$.

3 Estimating the NVMM Model using UNMIX

In Section 1, it was mentioned that there are many useful methods for estimating the parameters and the most common one is the EM (Expectation-Maximization) algorithm. Another method used here is the Bayesian method, which basically treats the parameters as random variables with a certain probability distribution called the prior distribution. The prior is updated based on the observed sample. The new updated prior is called the posterior, and is used to find the predictive distribution. The predictive distribution summarizes the information concerning the likely value of a new observation given the likelihood, the prior, and the data that have been observed. For more details on this method, refer to Carlin and Louis (2009).

The EM algorithm is a maximum-likelihood based algorithm. It started with the initial work of Larid and Dempester (1977). Their work was based on the idea that each data point is missing a label that indicates which component it belongs to. Since then, the EM has become the subject of research for many years because of its solid theoretical basis and optimal convergence properties (in most cases). However, there are still unresolved computational issues. In particular, the initialization of the parameters is still somewhat arbitrary, occasionally the maximization step might be stuck in a local maximum. Most importantly, it does not do well when the parameters are close in value. Obviously, all three issues are related. Also, deciding the number of components is still a difficult problem.

The Bayesian approach has gained some momentum in recent years because of advances made in computational power and the availability of software programs such as WinBUGS, the CRAN packages like RBugs and Dpdensity, and many others that use the MCMC efficiently. However, the prior selection is somewhat arbitrary (see the Eyes example in section 5), the predictive density is not easy to use for predicting new values, and the popularity of this approach is still limited, which sometimes makes it difficult to communicate and disseminate.

In this paper, estimating $f_X^\star(x)$ is based on finding a least-squares estimate of the density, as will be shown later. The approximated density is found by discretizing the mixing measures π_1 and π_2. Then, the sample is used to estimate the approximated mixture density using kernel smoothing techniques. The first step should be treated separately because it requires building a μ-sequence and a σ-sequence similar to that in Hamdan and Nolan (2004) and Hamdan (2006) (i.e. approximating the infinite mixture with a finite mixture based on discretizing the corresponding mixing measures). That is, how to make $f_X^\star(x)$ close to $f_X(x)$. Here, we will show a new approach for estimating $f_X^\star(x)$ with $\hat{f}_X(x)$ over specified grids of x values, μ values and σ values over the range of the data. Then, we will use this method to fit simulated data from two, three, and four NVMM respectively.

We will compare the empirical density found by using this approach with that found using the EM and the Bayesian approaches in two ways: graphically by eye-balling the density fits, and quantitatively in one of the examples. The inputs of the UNMIX program (Hamdan et al., 2004),used for estimating scale mixtures, are manipulated to be applied to the NVMM case.

The proposed way of estimating NVMM is based on the least-squares minimization method, which gives it the advantage of accessibility (since least squares methods are well developed, well known to most interdisciplinary researchers, and available in all statistics software). Most importantly, this method has great research potential as will be discussed later. Finally, this method can be used to improve the performance of the EM by selecting good initial points.

3.1 Developing UNMIX for Scale Mixtures

Given a sample of size n from the mixture, fix a grid of $r_1, r_2, \ldots, r_m$ of potential σ values called r-grid, as stated in Hamdan and Nolan in (2004), and a grid of $x_1, \ldots, x_k$ values called x-grid, where $k \geq m$ (the number of knowns should be more than the number of unknowns). Then, we can use the r-grid to approximate $f(x)$

$$f(x) = \int_0^\infty \frac{1}{r}\phi(\frac{x}{r}, 0, 1)\pi(r)dr \ \simeq \ \sum_{j=1}^{m} \frac{1}{r_j}\phi(\frac{x}{r_j}, 0, 1)\pi_j.$$

For each x_i in the x-grid, $f(x_i)$ can be evaluated by $\hat{f}(x_i)$ using a kernel smoother. In our case, we use the default empirical density with the Gaussian kernel provided by the R- software package. If $y_i = \hat{f}(x_i)$, then

$$y_i = \hat{f}(x_i) = \sum_{j=1}^{m} \frac{1}{r_j}\phi(\frac{x_i}{r_j}, 0, 1)\pi_j + \varepsilon_i.$$

Assuming the ε_i's are independent with mean 0, we can solve for π_j by minimizing $S(\pi)$, where $\pi^T = (\pi_1, \ldots, \pi_m)$,

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

$\phi_{ij} = (1/r_j)\,\phi(x_i/r_j, 0, 1)$, and w_i are preassigned weights. In our case, $w_i = 1$ for $1 \leq i \leq k$. However, if the data are heavy-tailed, one can try different weights until a good fit is found (a good strategy might be to give less weight to the points that are close to the mean of the x-grid and more weight to those that are far from the mean of the x-grid). We initially tried to use a regression approach to solve for π by writing the expansion of $S(\pi)$ in matrix format and defining $H = C^T C$, where C is a k by m matrix with entries $w_i\phi_{ij}$. Unfortunately, C is found to be singular although ϕ is highly nonlinear, especially when $m \geq 7$.

Therefore, instead of using the 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 particular, if we let g be the m by 1 vector defined as

$$g = \Big(-\sum_{i=1}^{k} w_i y_i \phi_{i1}, \ldots, -\sum_{i=1}^{k} w_i y_i \phi_{im} \Big)^T,$$

then $S(\pi) = 2\big[c + g^T\pi + (1/2)\pi^T H\pi\big]$, where $c = 2\sum_{i=1}^{k} w_i^2 y_i^2$ is constant . Hence, π can be found by minimizing $\big[g^T\pi + \frac{1}{2}\pi^T H\pi\big]$, subject to $\sum_{j=1}^{m} \pi_j = 1$ and $A\pi \geq \mathbf{b}$, where $A = I_m$ is an identity matrix of order m and $\mathbf{b}^T = (0, \ldots, 0)$ of order m. The quadratic programming routine, QPROG, from the International Mathematics and Statistics Library, IMSL, is employed and modified to fit the current problem. This program is called UNMIX. The program requires a grid of x points, which is called the x-grid, an estimate of the density at each element of the x-grid, a grid of R points called σ-grid (or r-grid), and a vector of weights with the same length as the length of the x-grid. The density estimate is found using a Gaussian kernel smoother. The default weights, $w_i = 1$ are all made equal to 1, where $1 \leq i \leq k$. The output is the vector π that minimizes $S(\pi)$.

3.2 The Sum of Squares Function in the NVMM Model

The following terms are needed for introducing the NVMM, the normal variance-mean mixtures model. Let the x-grid be $x = (x_1, x_2, \ldots, x_k)^T$, where $k \geq q \cdot m$, σ-grid be $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_m)^T$, and μ-grid be $\mu = (\mu_1, \mu_2, \ldots, \mu_q)^T$. The NVMM model is defined as

$$y_i = \sum_{a=1}^{q}\sum_{b=1}^{m} \phi(x_i, \mu_a, \sigma_b)\pi_{ab} + \epsilon_i,$$

where $\phi(x_i, \mu_a, \sigma_b)$ now is the normal p.d.f. with mean μ_a and standard deviation σ_b, evaluated at x_i in the x-grid, and ϵ_i is an error term assumed to be independent with mean 0 and a constant variance.

The estimated mixture at x_i is

$$y_i = \sum_{a=1}^{q}\sum_{b=1}^{m} \phi(x_i, \mu_a, \sigma_b)\hat{\pi}_{ab},$$

where $\hat{\boldsymbol{\pi}}$ minimizes

$$S(\boldsymbol{\pi}) = \sum_{i=1}^{k}(w_i(y_i - \sum_{a=1}^{q}\sum_{b=1}^{m} \phi(x_i, \mu_a, \sigma_b)\pi_{ab}))^2. \tag{3.1}$$

It is assumed that all terms in (3.2), (3.2), and (3.1) are known except $\boldsymbol{\pi}$, where

$$\pi = (\pi_{11}, \pi_{12}, \ldots, \pi_{1m}, \pi_{21}, \ldots, \pi_{2m}, \ldots, \pi_{q1}, \ldots, \pi_{qm})^T,$$

y_i is the empirical estimate at x_i or simply $\hat{f}(x_i)$, and w_i is the weight at x_i. In this study, we use the default $w_i = 1$ for all i, but one can use any weighted square distance based on the presence or absence of extremes (i.e. one can select a set of weights that minimize the effect of outliers on the empirical density). One way of simplifying (3.1) is rewriting the two-vector array as a one-vector array. That is, the subtracted term in (3.1) can be expressed as follows

$$\sum_{a=1}^{q}\sum_{b=1}^{m} \phi(x_i, \mu_a, \sigma_b)\pi_{ab} = \sum_{j=1}^{q \cdot m} \phi(x_i, z_j)\pi_j, \tag{3.2}$$

where $0 < j \leq q \cdot m$ and $\phi(x_i, z_j)\pi_j$ is the jth term of $\sum_{a=1}^{q}\sum_{b=1}^{m} \phi(x_i, \mu_a, \sigma_b)\pi_{ab}$. Hence,

$$z = \{(\mu_1, \sigma_1), \ldots, (\mu_1, \sigma_m), (\mu_2, \sigma_1), \ldots, (\mu_2, \sigma_m), \ldots, (\mu_q, \sigma_1), \ldots, (\mu_q, \sigma_m)\}.$$

By substituting (3.2) into (3.1),

$$S(\boldsymbol{\pi}) = \sum_{i=1}^{k}(w_i(y_i - \sum_{j=1}^{q \cdot m} \phi(x_i, z_j)\pi_j))^2. \tag{3.3}$$

Therefore, equation (3.3) is now in the form that can be minimized using one of the least-square scale mixture routines, and specifically we will use a programming routine called UNMIX. That is when equation (3.6) is used instead of equation (3.5). In what follows, the details of the UNIMX scale-mixture of normals routine is shown.

In our case, the C matrix is extended to account for μ grid and can be easily seen to be

$$\begin{pmatrix} \phi(x_1,\mu_1,\sigma_1) & \cdots & \phi(x_1,\mu_1,\sigma_m) & \phi(x_1,\mu_2,\sigma_1) & \cdots & \phi(x_1,\mu_2,\sigma_m) & \cdots & \phi(x_1,\mu_q,\sigma_m) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi(x_k,\mu_1,\sigma_1) & \cdots & \phi(x_k,\mu_1,\sigma_m) & \phi(x_k,\mu_2,\sigma_1) & \cdots & \phi(x_k,\mu_2,\sigma_m) & \cdots & \phi(x_k,\mu_q,\sigma_m) \end{pmatrix}$$

or equivalently,

$$C = \begin{pmatrix} \phi(x_1,z_1) & \phi(x_1,z_2) & \cdots & \phi(x_1,z_{qm}) \\ \vdots & \vdots & \vdots & \vdots \\ \phi(x_k,z_1) & \phi(x_k,z_2) & \cdots & \phi(x_k,z_{qm}) \end{pmatrix}$$

3.3 Identifiability Issues

In theory, since the recovered or estimated density depends on the grids chosen, the solution will not be unique. However, the Triangular Inequality can make the estimation fairly accurate and minimize the impact identifiable issue. To illustrate this, let the true density be denoted by $f(x)$, the approximated density be denoted by $f^*(x)$, and the empirical density be denoted by $\hat{f}(x)$. Hamdan and Nolan (2006) showed theoretically that one can find a σ-grid such that $|f^*(x) - f(x)| < \epsilon$ for all x and $\epsilon > 0$. Now, since

$$\left|f^*(x) - \hat{f}(x)\right| \leq \left|f^*(x) - f(x)\right| + \left|f(x) - \hat{f}(x)\right|,$$

and the empirical density under certain conditions can be made very close to the true density (i.e.$|f(x) - \hat{f}(x)| \leq \epsilon$), the approximated density can be found with high accuracy.

Therefore, the job in UNMIX becomes modeling the approximated density, $f^*(x)$, using the empirical density, $\hat{f}(x)$. A logical question that arises is about the real gain of using this method when modeling with it compared to, or in relative to, using the empirical estimate provided by any software, or when comparing it with estimates that are based on kernel smoothing techniques or any other techniques. With this method, the answer is that we get the functional form of the model. And that is in general, why parameters are estimated, whether this method is used or the EM or any estimation method.

To estimate the scale mixture of normals using UNMIX, three inputs are required: a pre-specified x-grid of size k, a y-grid that has the same size as that of the x-grid, and a σ-grid of size r. The points in the y-grid are the empirical density evaluated at each point in the x-grid values i.e. $y_i = \hat{f}(x_i)$, where $\hat{f}$ is the empirical density. The variance-mean extension of UNMIX, which will be called NVMM, requires one additional input which is a μ-grid of size m. These inputs are all required for the least-squares minimization routine described. The output vector, denoted by the π vector, consists of all, $r \times m$, weights that correspond to all possible points in the cross product of the σ-grid and μ-grid. Therefore, the resulting matrix of all possible combinations of σ's, μ's and the estimated weights or π's form an approximated mixture with r×m components over the selected *x*-grid. In the next subsection, a description of how to select these grids is provided.

To limit the number of terms in the estimate, components with small weights are ignored and the remaining are normalized. In most cases, the majority of the weights are less than .01. The choice of eliminating the terms with $\pi_i \leq .01$ is subjective and based on simulation and observations. Further implications of this limit are discussed later in Section 7.

3.4 *x*-Grid, σ-Grid, and μ-Grid

x-Grid In general, capturing the tails of the probability density requires having an *x*-grid that extends a little beyond the range of the data. Just like the traditional modeling is a delicate balance between capturing the details of the density in regions with high data concentration and capturing the extreme values. A basic and natural way to form an *x*-grid is to construct an equally-spaced sequence of *k* *x*-values starting at a point that is a little bit below the minimum sample value and ending at a value that is a little bit over the maximum sample value. If we go far beyond the range of the data, we may miss some major features of the main density. So far, based on empirical evidence, extending the range of the *x*-grid by about 15% of the range of the data when we have *k*$= 100$ seems to work well.

σ-Grid The σ-grid is selected to capture the tails of the mixture density. It can be an equally-spaced grid starting at a small positive point, which is equal to a small fraction of the sample standard deviation, and ending at a large point, which is equal to a constant times the sample standard deviation. For example, if S is the sample standard deviation then the σ-grid could be a set of equally-spaced k points between .01S and 5S (assuming the case where the scale has square root of inverse gamma, the range of possible standard deviations is almost 100% covered by this interval). We can also select these points interactively until we get a suitable or satisfactory fit of the empirical density.

μ-Grid This is a sequence of potential estimates of the μ values. We only use equally-spaced grids that include the peaks of the empirical density of the sample.

4 Simulation Study

A set of simulated examples are used to evaluate the method. In each example, the mixture was estimated using the default empirical density provided by the R software. Another estimate was based on the approximated density and using UNMIX.

4.1 Examples

4.1.1 Two-Component Mixture

This is a simulated example of two variance-mean normal components with parameter values $\mu_1 = -2$, $\sigma_1 = 10$ and $\mu_2 = 5$, $\sigma_2 = 15$ with weights of 0.833 and 0.167, respectively.

The μ-grid in this case was 21 equally spaced points between -10 and 10. The σ-grid consists of 11 points starting at .1 and ending at 20 i.e. $\{$ 0.10, 2.09, 4.08, 6.07, 8.06, 10.05, 12.04, 14.03, 16.02, 18.01, 20.00 $\}$. Therefore, the initial number of terms is 231 (11×21). The density estimate used here is based on the 17 top-weighted components with a total weight of 0.995. The remaining 214 components contribute only to 0.005 of the total weight in the mixture density. Therefore, ignoring these 214 terms, then using the top normalized 17 terms (i.e. divide the 17 remaining weights by .995, so that the density is a valid one) provided an excellent fit to the true mixture. Alternatively, we can update the μ-grid and the σ-grid to include only the terms that have significant weights. However, 17 terms is still a large number to work with. Fortunately, retaining 11 terms after deleting the terms that have weights of 2% or less still does a good job in fitting the density as shown in Figure 1. Further cuts can still be made: when the terms with weights of 3% or less are deleted, only 4 terms remain and the new density fits the data reasonably well (in terms of capturing the overall shape). What is even more promising is that when the size of the σ-grid is increased from 11 to 15 and when the terms with weights of 3% or less are ignored, only three terms remain. The estimated density with these three terms is really close to the exact density.

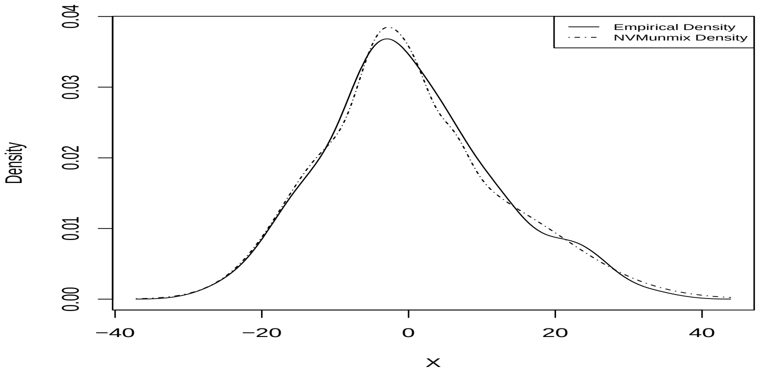

Figure 1: Graph of $f(x) = 0.833\phi(x, -2, 10) + 0.167\phi(x, 5, 15)$ and the estimated density. The fit is based on $n = 120$.

Of course, fitting a two-component variance-mean mixture with 11 components is not a real gain no matter how good the fit is. On the other hand, since the search was done with only one μ-grid and one σ-grid, it can be improved with larger grids. In fact, the selected grid provided one of the worst possible scenario. In most cases, only three to five terms contribute to the top 95% of the weights.

The same pattern continued when the sample size was reduced from 120 to 60. The shape of the density was captured reasonably well with four terms when the weights of 4% or less are deleted.

4.1.2 Three-Component Mixture

In this simulated example, a three-component variance-mean mixtures of normals is fitted with two different samples. The first sample has 250 observations and the second has 125 observations. The parameters of this mixture are $\mu_1 = -2$, $\sigma_1 = 10$, $\mu_2 = 20$, $\sigma_2 = 20$, and $\mu_3 = 10$, $\sigma_3 = 15$ with weights of 0.4, 0.2, and 0.4, respectively.

The μ-grid consists of the following 12 equally-spaced points {-2, 0 ,2, 4 , 6, 8, 10, 12, 14, 16, 18, 20} and the σ-grid consists of the following 11 equally-spaced points {0.10, 2.09, 4.08, 6.07, 8.06, 10.05, 12.04, 14.03, 16.02, 18.01, 20.00 }. The selection was merely based on programming convenience. However, as discussed later in this paper, the selection should be done interactively by investigating the empirical density found using kernel smoothing techniques.

The first observation made, which is not surprising, is that when all 132 terms are used, the density fit as seen in Figure 2 is captured almost perfectly. It has the same shape as the original density and does a great job in terms of capturing the details of the empirical density. Now, to reduce the number of terms to few, two scenarios are tried. In the first scenario, the terms that make up the smallest 1% of the weights are ignored, and only the seven terms that make up the top 99 % of the weights are retained. The fit, as shown in Figure 3, is somewhat compromised around the peak of the density. The same thing happens with the second try when 8 terms are retained with a total weight of 0.9669317. In particular, the estimated density is shifted to the left but, as seen in Figure 3, the general shape is captured very well. The trade off from reducing the number of terms from 132 to 8 is definitely worth it. Of course that doesn't say anything about the parameter estimates.

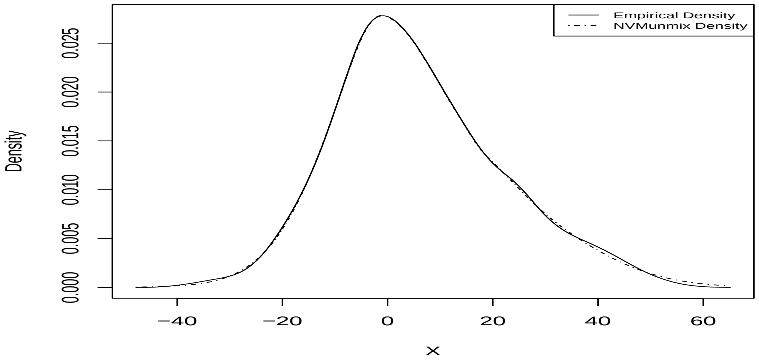

Figure 2: Graph of $f(x) = .4\phi(x, -2, 10) + .2\phi(x, 20, 20) + .4\phi(x, 10, 15)$ and the estimated density. The fit is based on $n = 250$.

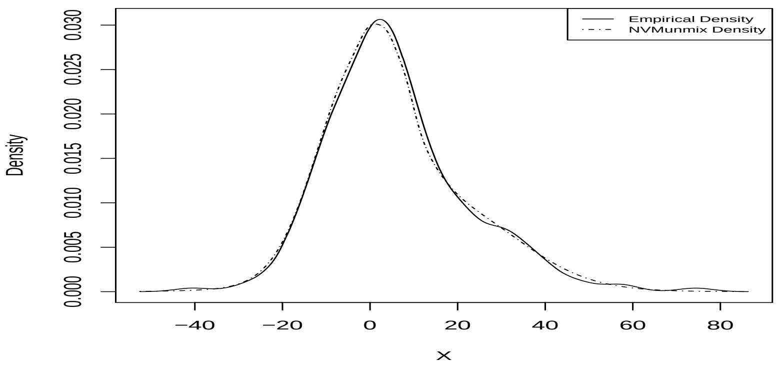

Figure 3: Graph of $f(x) = .4\phi(x, -2, 10) + .2\phi(x, 20, 20) + .4\phi(x, 10, 15)$ and the estimated density without the terms with weights of 1% or less. The fit is based on $n = 250$.

Figure 4 shows the fitted density with 5 terms only. That is when 126 terms are ignored since these terms have weights of 3% or less. The sample size in this case is 125.

Table 1 shows the terms of the selected mixture for the fit with 5 components (when weights with 3% or less are ignored). Notice that in terms of estimating the parameters, the third component is totally missed but the fit is very good.

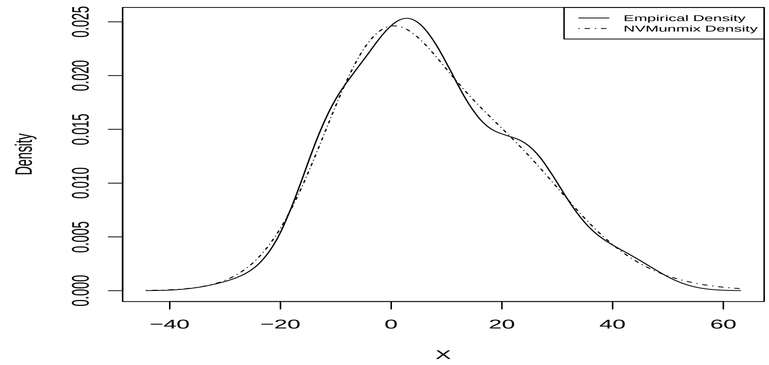

Figure 4: Graph of $f(x) = .4\phi(x, -2, 10) + .2\phi(x, 20, 20) + .4\phi(x, 10, 15)$ and the estimated density without the terms with weights of 3% or less. The fit is based on $n = 125$.

Table 1: Three component mixture fit with five components

term	μ	σ	π
006	-2	10.05	0.25611511
007	-2	12.04	0.31738999
127	20	10.05	0.05702863
128	20	12.04	0.08807759
130	20	16.02	0.28138868

4.1.3 Four-Component Mixture

In this simulated example, the density of a four-component normal mixture is estimated. The weights of this mixture are .4, .25, .15, and .2; the means are -2, 8, 12, and 20; and the standard deviations are 10, 25, 15, and 20, respectively.

The σ-grid used is a sequence of 16 equally spaced points between 0.1 and 30, and the μ-grid is a sequence of 18 equally spaced points between -4 and 30. Therefore, the grid size used in the estimation is 288 (16×18). When the terms with weights of .02 or less are ignored, only 6 terms remain. These 6 terms make up about 94% of the total weights. The density fit with the 6 normalized terms looks great in terms of capturing the overall shape. It is also worthwhile to mention that finding this estimate is done with only a few tries of different combinations of μ-grids and σ-grids. See Figure 5 for details.

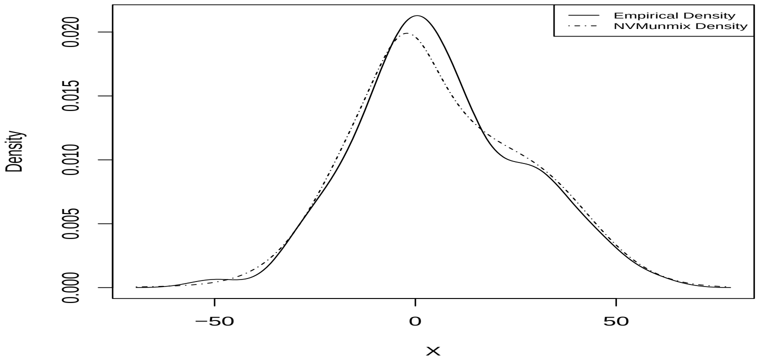

Figure 5: Graph of $f(x) = .4\phi(x, -2, 10) + .25\phi(x, 8, 25) + .15\phi(x, 12, 15) + .2\phi(x, 20, 20)$ and the estimated density without the terms with weights of 2% or less. The fit is based on $n = 200$.

5 Comparing the NVMM, EM, and Bayesian Fits

In this section, the UNMIX fit of the NVMM and the fit based on the EM are compared over a grid of equally spaced x-values. The comparison is done using three real life examples and two simulated examples. The criteria used for comparison in the first example are the chi-square goodness of fit and the Akaike Information Criterion (AIC) while the remaining examples of the comparison are done graphically. The goal of the UNMIX fit of the NVMM is to fit a model containing a small number of parameters that captures the main shape. That is, the new NVMM fit is intended to be similar to the SAS FMM, Finite Mixture Model, procedure. The FMM procedure is built to fit the density based on either the Bayesian approach or the EM.

5.1 Galaxy Data

Roeder (1990) and Escobar and West (1995) studied data from the Corona Borealis sky survey with the velocities of 82 galaxies in a narrow slice of the sky. Cosmological theory suggests that the observed velocity of each galaxy is proportional to its distance from the observer. Thus, the presence of multiple modes in the density of these velocities could indicate a clustering of the galaxies at different distances. The computed variable v represents the measured velocity in thousands of kilometers per second. Roeder(1990) modeled the density with a five-component normal mixture after exploring all models with a number of components ranging from 3 to 7 and restricting the scale parameter at 0.9025 mixture. Lunn et al. (2013) fit a Dirichlet process mixture of normals. In Figure 6, the new NVMM fit is compared with the Bayesian fit while the next table compares the SAS experimental mixture modeling procedure, known as FMM, NVMM, and Roeder's fit using the AIC criteria and the Chi-square goodness of fit.

Table 2: Comparing the NVMM fit using UNMIX with EM and Roeder's method for modeling the Galxy data

Method	AIC	Pearson
Reoder's	430.2	82.5549
EM (Using SAS FMM)	422.96	82.00
NVMM	851.72	8.374

It is very clear that the NVMM fit using UNMIX has a much better fit than the FMM and Roeder, but the AIC is more than double now. So, a better balance between the number of components and the fit can be reached by trying for a new combination of μ-grids and σ-grids.

5.2 Eyes Data

The data set consists of 48 measurements on the peak sensitivity wavelengths for individual microspectrophot records on the eyes of monkeys. Modeling these can also be found in Lunn et al.

ht]

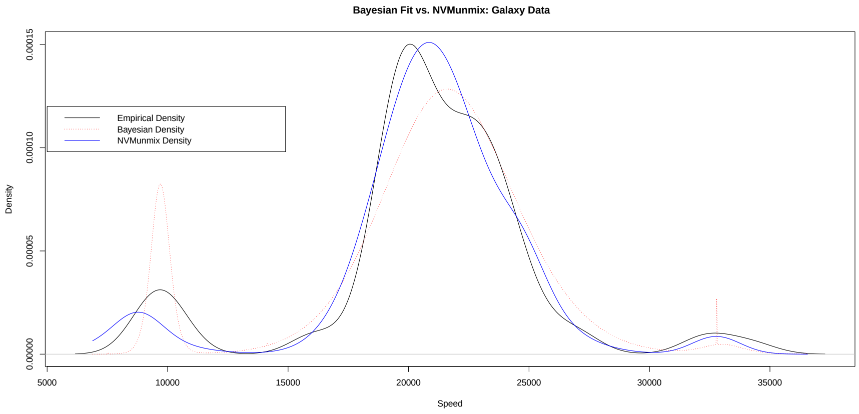

Figure 6: Comparing the estimated density of the Galaxy data using the Bayesian approach and UNMIX without the terms with weights of 2% or less.

(2013) and Carlin and Louis (2009). The density of the data is modeled using a two-component mixture of normals and compared to the kernel density estimate. In particular, they both use

$$y_i \sim \phi\left(\lambda_{T_i}, \tau\right), \quad T_i \sim Categorical(p),$$

where $i = 1, 2$ and the distribution of T_i is Bernoulli(p). Further, to better identify the components of the mixture, they both assumed that $\lambda_2 = \lambda_1 + \theta$ and $\theta > 0$. However, Carlin and Louis (2009) assumed a truncated normal prior for θ and a normal prior for λ_1 while Lunn et al. (2013) assumed a uniform prior for θ between 0 and 1000 and a uniform prior for λ_1 between -1000 and $+1000$. See Figure 7.

5.3 Boreal Forest Birds

Holling (1992) studied the body mass of 101 Boreal forest birds found east of the Manitoba-Ontario border in pure or mixed conifer stands. Holling's results suggest that there are four or more body size clumps for the body mass distribution. Xu et al. (2010) showed that there are 2 or 3 components for the body mass distribution. See Figure 8.

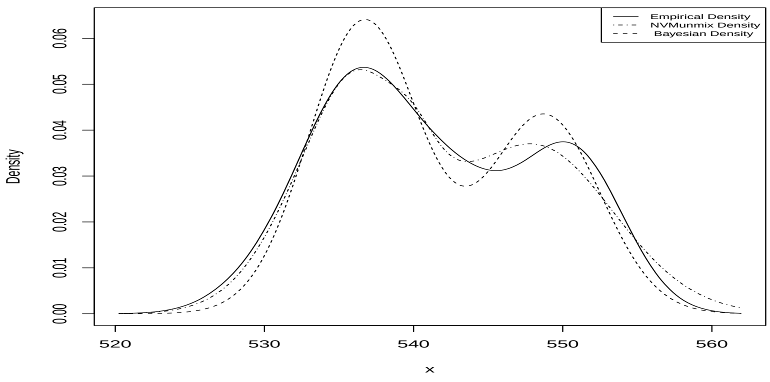

Figure 7: Comparing the estimated density of the Eyes data using the Bayesian approach and UNMIX without the terms with weights of 2% or less.

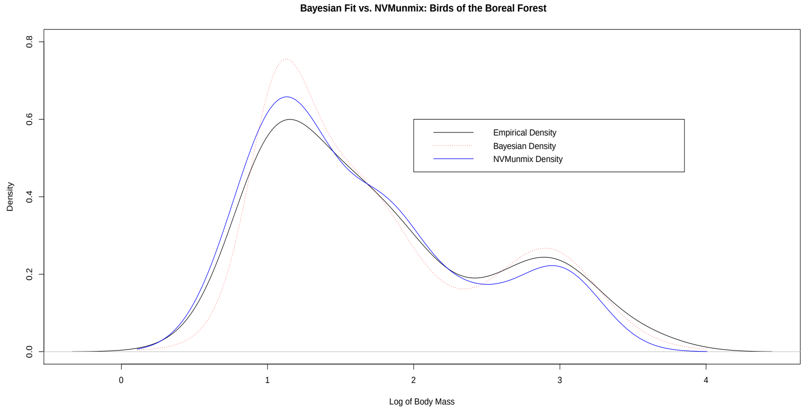

Figure 8: Comparing the estimated density of the birds of Boreal Forest data using the Bayesian approach and UNMIX without the terms with weights of 2% or less.

5.4 Simulated Data with Two Components

In this example, 400 observations are simulated from a two-component mixture with means $\mu_1 = 10$ and $\mu_2 = 20$, standard deviations $\sigma_1 = 15$ and $\sigma_2 = 6$, and a mixing weight of $\pi = 0.25$. The estimated values of $(\pi, \mu_1, \mu_2, \sigma_1, \sigma_2)$ using the EM with 500 iterations are (0.339, 8.465, 21.723,14.914, 6.695).

On the other hand, when UNMIX with a μ-grid of (0,4, 8,12,16,20) and a σ-grid of 20 equally spaced points between .1 and 20 is used, after ignoring the terms with weights of 10% or less, there are only three terms remain. The three terms are given in table 3.

Table 3: The retained terms for the simulated 2-component normal mixture

μ	σ	π
0	12.04	0.3388810
20	8.06	0.2603132
20	10.05	0.4008058

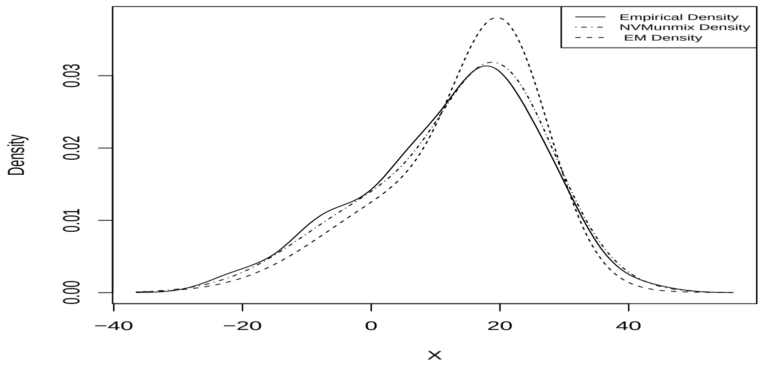

Figure 9: Comparing the estimated density of $.25N(20,6) + .75N(10,15)$ using the EM and the NVMM without the terms with weights of 10% or less.

The estimated density using the UNMIX method and the fit based on the EM estimates are compared with the empirical density in Figure 9. Overall, both fits seem to be reasonably good but the UNMIX fit outperforms the EM near the main peak.

It is very clear that for recovering or estimating the parameters the EM performs better than the UNMIX even when the μ-grid and σ-grid contain the true values. These methods work on different principles: one is based on finding the best fit and the other is based on updating the parameter estimates until convergence.

6 Finding the Standard Error

In this section, we talk about the variability in estimating the weights or the $\pi's$. There are three sources of variations. The first and most significant source is the sampling variability. The second is due to the selection of the μ-grid and the σ-grid. The third is the variability due to the density estimation.

First, recall that the point of using the UNMIX program is to find the $\pi's$ that minimize the squared distance between the empirical density and the approximated one over a fixed μ-grid and a fixed σ-grid. In the case of scale mixture UNMIX, this was done only for a fixed σ-grid which was found by approximating the true infinite mixture or density, usually unknown, by a finite mixture in the sense of Hamdan and Nolan (2006). The good news is that there is so much literature on empirical densities. The variability associated with using empirical densities is either due to sampling variability or due to choices of the smoothing parameters (size of the window, kernel smoother, etc.). Here, we will estimate the sampling variability using bootstrap samples, i.e. by fixing the σ-grid and the μ-grid then bootstrapping the sample.

Example 6.1. The three-component variance-mean mixture of normals is revisited. The sample size n =250, and the parameters of this mixture are $\mu_1 = -2$ and $\sigma_1 = 10$, $\mu_2 = 20$ and $\sigma_2 = 20$, and $\mu_3 = 10$ and $\sigma_3 = 15$ with weights of 0.4, 0.2, and 0.4, respectively.

In this case, the μ-grid and σ-grid are modified slightly. In particular, the μ-grid consists of the following 6 equally spaced points $\{-2, 2, 6, 10, 14, 18\}$ and the σ-grid consists of the following 11 equally spaced points between .01 and 20.00:
$\{0.10, 2.09, 4.08, 6.07, 8.06, 10.05, 12.04, 14.03, 16.02, 18.01, 20.00\}$.

Table 4: Bootstrapping the standard error of the π's

Term	μ	σ	π	$\hat{\sigma}\,(\pi)$
6	-2	10.05	0.25195790	0.17532474
7	-2	12.04	0.24164040	0.18291332
9	-2	16.02	0.08757814	0.07734820
59	18	6.07	0.05082034	0.03306025
66	18	20.00	0.36800322	0.12943057

Figure 10 and the previous table show the size of the sampling variability relative to the actual weights (or estimated π's). There are many possible causes of this huge sampling variability. The main cause of this is that the weight concentration shifts quickly at nearby points in the μ-grid and nearby points in the σ-grid. That happens because the objective function varies with the sample, i.e. the squared distance between the empirical density and the approximated one changes slightly each time a new bootstrap sample is selected. Of course this variability can always be controlled by starting with a large μ-grid and a large σ-grid. Also, one factor found to be very significant in terms of controlling this variability is including the modes of the empirical density in the μ-grid.

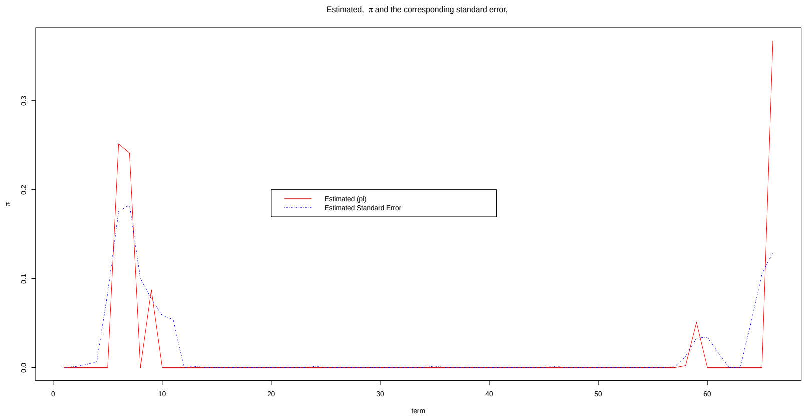

Figure 10: The standard error is large near and at the selected terms

7 Conclusion

A new approach for estimating the density of finite variance-mean mixture of normals is introduced. The main idea is based on estimating the weights of the components of the mixture over specified grids of x-values, σ-values, and μ-values.

As shown in many simulated and real life examples, when the objective of the study is fitting the data, UNMIX fits the empirical density and the true density very well. In fact, in all studied cases UNMIX outperforms the fits provided by the EM, SAS-FMM procedure, and Bayesian approach. Besides its ease of use, initially, there is no need to worry about the number of components. One can start with the number of components as large as possible and then can reduce that number significantly by accelerating the terms that have small weights.

A good modeling strategy is to start with the largest possible μ-grid and σ-grid and reduce them systematically and interactively until a satisfactory fit with a reasonable number of components is found. Based on the examples we have done, the best working strategy seems to be making the x-grid larger than the μ-grid, equally spaced and covering the range of the data. If the empirical density is multimodal, including points that are close to these modes in the μ-grid can speed up the search for a grid that provides good fit. Usually, the interactive search takes at most 4 or 5 tries.

The only drawback to this method of fitting mixtures is that the solution is not identifiable and not unique. Changing the grid will change the estimated parameter values. Also, sometimes reducing the number of components by eliminating the terms that make up the lowest 1 or 2 percent of the weights doesn't reduce the number of terms significantly, which might cause some over-fitting problems.

Our future plans are to study this method with more examples and find out the most effective way of selecting the grids as well as most effective way of weighting the sample. We plan to explore how one can work with the EM and other existing techniques to improve their performance close under the conditions that don't perform well. For example, when the parameter values are close.

Acknowledgments

The authors would like to thank the Department of Mathematics and Statistics at James Madison University for their support and the referees in advance for their feedback.

References

Alexander, C. (2004), "Normal mixture diffusion with uncertain volatility: Modelling short- and long-term smile effects", *Journal of Banking & Finance*, 28, 2957–2980.

Barndorff-Nielsen, O., Kent, J., and Sorensen, M. (1982), "Normal variance-mean mixtures and z distributions", *International Statistical Review*, 50, 145–159.

Belomestny D. and Panov V. (2018), "Semiparametric estimation in the normal variance-mean mixture model", *Statistics*, 52(3), 571–589.

Bhattacharya, C. T. (1967), "A simple method of resolution of a distribution into Gaussian components", *Biometrics*, 23, 115–137.

Carlin, B P. and Louis, T. A. (2009), "Bayesian Methods for Data Analysis, 3rd edition", CRC Press.

Chen, H., Chen, J. and Kalbfleisch, J. (2004). "Testing for a finite mixture model with two components", *Journal of Royal Statistical Society, Series B*, 66, 95–115.

Dempster, A. P., Larid, N. M., and Rubin D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of Royal Statistical Society*, 39(1), 11–38.

Escobar, M. D. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures", *Journal of the American Statistical Association*, 90, 577–588.

Gomez, H., Venegas, O., and Bolfarine, H. (2007), "Skew-symmetric distributions generated by the distribution function of the normal distribution", *Environmetrics*, 18, 395–407.

Hamdan, H. (2006), "Characterizing and approximating infinite scale mixtures of normals", *Communications in Statistics - Theory and Methods*, 35, 407–413.

Hamdan, H. and Nolan, J. (2004), "Approximating Scale Mixtures" in A. C. Krinik and R. J. Swift, Stochastic Processes and Functional Analysis, A Dekker Series of Lecture Notes in Pure and Applied Mathematics, 161–169.

Hamdan, H., Nolan, J.,Wilson, M., and Dardia, K. (2005). "Using scale mixtures of normals to model continuously compounded returns", *Journal of Modern Applied Statistical Models*, 4(1), 214–226.

Holling, C. S. (1992), “Cross-scale morphology, geometry, and dynamics of ecosystems”, *Ecological Monographs*, 62(4), 447–502.

James, L., Priebe, C., and Marchette, D. (2001), “Consistent estimation of mixture complexity”, *The Annals of Statistics*, 29(5), 1281–1296.

Jones, M., Marron, J. S., and Sheather, S. (1996), “A brief survey of bandwidth selection for density estimation”, *Journal of the American Statistical Association*, 91(433), 401–407.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1992), “Progress in Data-Based Bandwidth Selection for Kernel Density Estimation”, University of New South Wales, Australian Graduate School of Management, Working Paper Series 92–014.

Kasahara, H. and Shimotsu, K. (2014), “Testing the Number of Components in Normal Mixture Regression Models”, *Journal of the American Statistical Association*, 103, 1674—1683.

Kasahara, H. and Shimotsu, K. (2014), “Non-parametric identification and estimation of the number of components in multivariate mixtures”, *Journal of Royal Statistical Society, Series B*, 76(1), 97–111.

Kessler, D. and McDowell A. (2012), “Introducing the FMM Procedure for Finite Mixture Models. SAS Global Forum: Statistics and Data Analysis”, SAS Institue Inc., Paper 328.

Lunn, D., Jackson, Ch., Best, N., Thomas, A., and Spiegelhalter, D. (2012), “The BUGS BOOK, A Practical Introduction to Bayesian Statistics”, Chapman and Hall/CRC.

Oliviera-Brochado, A. and Vitorino M. F. (2005), “Assessing the Number of Components in Mixture Models: A Review”, FEP Working Paper 194.

McLachlan, G. and Peaal, D. (2000), “Finite Mixture Models”, Wiley.

Parzen, E. (1962), “On the estimation of probability density function and mode”, *The Annals of Mathematical Statistics*, 33, 1065–1076.

Park, B. and Marron, J. S. (1990), “Comparison of data-driven bandwidth selectors”, *Journal of the American Statistical Association*, 85(409), 66–72.

Permuter, H., Francos, H., and Jermyn, I. (2003), “Gaussian mixture models of texture and colour for image database retrieval”, In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3, III–569.

Redner, R. and Walker, H. (1984), “Mixture Densities, Maximum Likelihood and the Em Algorithm”, *SIAM Review*, 26(2), 195–239.

Roeder K. (1990), “Density estimation with confidence sets exemplified by superclusters and voids in the galaxies”, *Journal of the American Statistical Association*, 85, 617–624.

Sheather, S. (2004), “Density estimation”, *Statistical Science*, 19(4), 588–597.

Sheather, S. and Jones, M. (1991), “A reliable data-based bandwidth selection method for Kernel Density Estimation”, *Journal of Royal Statistical Society, Series B*, 53(3), 683–690.

Silverman, B. W. (1986), “Density Estimation for Statistics and Data Analysis”, Chapman and Hall: New York.

Titterington, D., Smith, A., and Makov, U. (1985), “Statistical Analysis of Finite Mixture Distribution”, John Wiley & Sons, Chichester, U. K.

R Core Team, (2018). “R: A Language and Environment for Statistical Computing”, *R Foundation for Statistical Computing*, Vienna, Austria.

Yao, H. and Taimre T. (2016), “Estimating Tail Probabilities of Random Sums of Infinite Mixtures of Phase-Type Distributions”, In *IEEE, Proceeding of the Winter Simulation Conference*, 347–358.

Xu, L., Hansan, T., Bedrick, E., and Restrepo, C. (2010), “ Hypothesis tests on mixture model components with applications in Ecology and Agriculture”, *Journal of Agricultural, Biological, and Environmental Statistics*, 3, 308–326.

Received: June 14, 2018

Accepted: January 30, 2019