A Note on the Gao, An, and Bai (2019) Uniform Mixture Model in the Case of Regression

Mike G. Tsionas* Athanasios Andrikopoulos†

Abstract

We extend the Uniform Mixture Model of Gao, An and Bai (2019) to the case of linear regression. Gao, An and Bai (2019) proposed that to characterize the probability distributions of multimodal and irregular data observed in engineering, a Uniform Mixture Model can be used. This model is a weighted combination of multiple uniform distribution components. This case is of empirical interest since, in many instances, the distribution of the error term in a linear regression model cannot be assumed unimodal. Bayesian methods of inference organized around Markov Chain Monte Carlo are proposed. In a Monte Carlo experiment, significant efficiency gains are found in comparison to least squares justifying the use of the Uniform Mixture Model.

Key Words: Multimodal data · Uniform mixture model · Regression Models · Statistical Inference · Bayesian Analysis.

*Lancaster University Management School, LA1 4YX, U.K., m.tsionas@lancaster.ac.uk
†Faculty of Business, Law and Politics, University of Hull, UK, a.andrikopoulos@hull.ac.uk

1 Introduction

Gao, An and Bai (2019) proposed that to characterize the probability distributions of multimodal and irregular data observed from practical engineering, a Uniform Mixture Model (UMM) can be used, which is a weighted combination of multiple uniform distribution components. As these authors notice, because of noise in many data sets, "probability distributions of observed data can not be accurately characterized by typical unimodal distributions (such as normal, lognormal, and Weibull distributions), and the adequacy of typical unimodal distributions may be questioned".

The uniform distribution in the interval (a, b) has probability density:

$$f(u) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The UMM is defined by discretizing the support to points $\{a_1, a_2, ..., a_{N+1}\}$, where N is given, and using the following mixture density:

$$f_{UMM}(u) = \sum_{j=1}^{N} w_j \frac{1}{a_{j+1} - a_j} \mathbb{I}(a_j < u < a_{j+1}), \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function and the weights w_j satisfy

$$w_j \geq 0,\ j = 1, ..., N,\ \sum_{j=1}^{N} w_j = 1. \tag{3}$$

2 The case of linear regression

Consider now a regression model of the form

$$y_i = x_i'\beta + u_i, i = 1, ..., n, \tag{4}$$

where y_i is the dependent variable, $x_i \in \Re^k$ is a vector of explanatory variables, $\beta \in \Re^k$ is a vector of coefficients to be estimated, and $n > k$ is the number of observations. Suppose the first element of x_i is unity so that an intercept is always present in the model. Assuming the distribution of the error term, u_i, is unknown but can be approximated by a UMM, we must have $\mathbb{E}(u_i|\{x_t\}_{t=1}^n) = 0, i = 1, ..., n$, which implies the following constraint:

$$\mathbb{E}(u_i|x_i) = \sum_{j=1}^{N} w_j \frac{a_j + a_{j+1}}{2} = \Delta \sum_{j=1}^{N} j w_j + a_1 - \tfrac{\Delta}{2} = 0, \tag{5}$$

assuming $a_{j+1} - a_j = \Delta\, \forall j$. From (2) we have that:

$$f_{UMM}(u_i) = \sum_{j=1}^{N} w_j \frac{1}{\Delta} \mathbb{I}(a_j < y_i - x_i'\beta < a_{j+1}), i = 1, ..., n. \tag{6}$$

Since $a_j = a_1 + (j-1)\Delta$, we can write this equation as:

$$f_{UMM}(u_i) = \sum_{j=1}^{N} w_j \frac{1}{\Delta} \mathbb{I}(a_1 + (j-1)\Delta < y_i - x_i'\beta < a_1 + j\Delta), i = 1, ..., n, \tag{7}$$

which implies that

$$f_{UMM}(u_i) = \sum_{j=1}^{N} w_j \frac{1}{\Delta} \mathbb{I}(-\Delta < y_i - x_i'\beta - a_1 - j\Delta < 0), i = 1, ..., n. \tag{8}$$

3 Statistical inference

3.1 Markov Chain Monte Carlo (MCMC) in general

Very often, complicated posterior distributions arise in statistics, operations research, and related field. Given a parameter $\alpha \in \mathcal{A} \subseteq \Re^d$, and data $\mathcal{D}$, suppose that the likelihood function is $\mathscr{L}(\alpha; D)$. Suppose also we have a prior on the parameters, say $p(\alpha)$. By Bayes theorem we know that the posterior is:

$$p(\alpha|\mathcal{D}) \propto \mathscr{L}(\alpha; \mathcal{D}) p(\alpha). \tag{9}$$

In general, we are interested in the posterior means of certain functions of interest, say $f(\alpha)$. The posterior mean of this function of interest is:

$$\mathbb{E}_{\alpha|\mathcal{D}}[f(\alpha)] = \frac{\int_{\mathcal{A}} f(\alpha) p(\alpha|\mathcal{D})\, d\alpha}{\int_{\mathcal{A}} p(\alpha|\mathcal{D})\, d\alpha}, \tag{10}$$

where $\mathbb{E}_{\alpha|\mathcal{D}}[f(\alpha)]$ denotes posterior expectation, and the denominator is the normalizing constant of the posterior. Part of the problem could be to find marginal posterior densities. If we partition $\alpha = [\alpha_1', \alpha_2']$ then the marginal posterior density of α_1 would be

$$p(\alpha_1|\mathcal{D}) = \frac{\int_{\mathcal{A}} p(\alpha_1, \alpha_2|\mathcal{D})\, d\alpha_2}{\int_{\mathcal{A}} p(\alpha|\mathcal{D})\, d\alpha}. \tag{11}$$

These integrals are typically, not available in closed form unless the problem is very simple. The Gibbs sampler, a particular MCMC technique relies on the idea that we may be able to produce a sequence of parameter draws $\{\alpha^{(s)},\ s = 1, \ldots, S\}$, not necessarily iid, which converges (as $S \to \infty$) to the posterior whose unormalized density is given by (9). If such a sample were available, the posterior expectation in (10) could be accurately approximated as follows:

$$\mathbb{E}_{\alpha|\mathcal{D}}[f(\alpha)] \simeq S^{-1} \sum_{s=1}^{S} f(\alpha^{(s)}). \tag{12}$$

Therefore, a sampling approach would facilitate the tasks of Bayesian inference to a great degree. The Gibbs

sampler relies on the idea that the sequence $\left\{\alpha^{(s)},\, s=1,\ldots,S\right\}$ can be produced recursively by using the conditional posterior distribution of each element of α. Suppose for example $\alpha=[\alpha_1,\alpha_2]'$ where α_1,α_2 are two scalar parameters for simplicity (although clearly they can be vectors). The Gibbs sampler is as follows:

- Draw $\alpha_1^{(s)}$ from its conditional distribution $\alpha_1|\alpha_2^{(s-1)},\mathcal{D}$,
- Draw $\alpha_2^{(s)}$ from its conditional distribution $\alpha_2|\alpha_1^{(s)},\mathcal{D}$,

and so on, if there are additional parameters. We repeat for $s=1,\ldots,S$ and we assume $\alpha_2^{(0)}$ is available. Quite often, the conditional posterior distributions are univariate and amenable to random number generation by commonly available means.

3.2 MCMC in the UMM linear regression model

Suppose now there is an index $J_i\in\{1,...,N\}$ so that

$$-\Delta<y_i-x_i'\beta-a_1-J_i\Delta<0,\; i=1,...,n, \tag{13}$$

whose interpretation is that u_i is drawn from a uniform distribution in (a_{J_i},a_{J_i+1}) with probability w_j.

In turn, the posterior (augmented) distribution of the model is:

$$p(\theta,\{J_i\}_{i=1}^n|D)\propto\prod_{i=1}^n w_{J_i}\mathbb{I}\left(-\Delta<y_i-x_i'\beta-a_1-J_i\Delta<0\right)p(\theta). \tag{14}$$

Here, θ is the parameter vector which includes β and some other elements as we explain below, and D denotes the entire data set $\{y_i,x_i\}_{i=1}^n$. Therefore, we have:

$$p(\theta,\{J_i\}_{i=1}^n|D)\propto w_1^{n_1}...w_N^{n_N}\prod_{i=1}^n\mathbb{I}\left(-\Delta<y_i-x_i'\beta-a_1-J_i\Delta<0\right)p(\theta), \tag{15}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $n_j=\sum_{i=1}^n\mathbb{I}\left(-\Delta<y_i-x_i'\beta-a_1-j\Delta<0\right)$, and $\sum_{j=1}^N n_j=n$. So, n_j represents the number of observations in the jth sub-interval.

It turns out that given Δ and N the *endpoint* a_1 *can be estimated from the data.* Define the parameter vector as $\theta=[\beta',a_1,\{J_i\}_{i=1}^n,w']'$. Given the J_is we must have:

$$a_1+(J_i-1)\Delta<y_i-x_i'\beta<a_1+J_i\Delta,\; i=1,...,n. \tag{16}$$

Therefore, the conditional posterior of regression parameters, β, is:

$$\begin{gathered}p(\beta|\{J_i\},w,a_1)\propto\text{const.},\\ \text{s.t }\Psi\equiv(\min_{t=1,\ldots,n}y_t-a_{J_t})>x_i'\beta>(\max_{t=1,\ldots,n}y_t-a_{J_t})\equiv\psi,\; i=1,...,n.\end{gathered} \tag{17}$$

From (5) along with the posterior in (15) we have

$$\max_{t=1,...,n}(y_t - x_t'\beta) - N\Delta < a_1 < \min_{t=1,...,n}(y_t - x_t'\beta), \tag{18}$$

where the first inequality comes from the restriction: $a_{N+1} = a_1 + N\Delta > \max_{t=1,...,n}(y_t - x_t'\beta)$. Moreover, we have:

$$a_{N+1} = a_1 + N\Delta. \tag{19}$$

Therefore, the right endpoint can be expressed in terms of N, a_1 and a_{N+1}. If we wish to impose the constraint $a_1 = -a_{N+1}$ then we have $a_{N+1} = \frac{N\Delta}{2}$. In this case, $a_1 = -\frac{N\Delta}{2}$, and a_1 has to be treated as given. We follow this practice, throughout to simplify the analysis as treating a_1 adds a layer of technicalities, although it is straightforward to treat it as an unknown parameter. In practice, the support of the error can be accurately estimated using the standard error of LS residuals.

Given $\{w_j\}$, N and Δ, these equations determine the values of the endpoints. Suppose our prior is

$$p(\beta, w) \propto w_1^{-1}...w_N^{-1}p(\beta), \tag{20}$$

In turn, the conditional posterior of weights is:

$$p(w|\beta, \{J_i\}_{i=1}^n, D) \propto w_1^{n_1-1}...w_N^{n_N-1}, \tag{21}$$

subject to (3), which is a Dirichlet distribution.

From (17) we have that β has to be drawn from the prior $p(\beta)$ subject to the restrictions that $\Psi > x_i'\beta > \psi, i = 1, ..., n$, as in (17). A particular convenient prior is the flat prior, viz. $p(\beta) \propto$ const. All the above techniques can be implemented using straightforward Markov Chain Monte Carlo (MCMC) techniques organized around the Gibbs sampler (Gelfand and Smith, 1990) by drawing successively random numbers from the conditional posterior distributions in (17) and (21). In particular, for β we proceed as follows. The restrictions that $\Psi > x_i'\beta > \psi, i = 1, ..., n$, as in (17), can be written, in matrix notation as:

$$\Psi 1_n > X\beta > \psi 1_n, \tag{22}$$

where 1_n is an $n \times 1$ vector of ones, and X is the $n \times k$ matrix of regressors. In turn, the posterior conditional distribution of β is $p(\beta) \propto$ const. subject to these restrictions. Suppose $X = [\mathbf{x}_1, ..., \mathbf{x}_k]$ where $\mathbf{x}_j$ is the $j$the column of X, an $n \times 1$ vector. We can write (22) as follows:

$$\Psi 1_n > \beta_1\mathbf{x}_1 + ... + \beta_k\mathbf{x}_k > \psi 1_n. \tag{23}$$

Suppose we want to draw $\beta_1|\beta_2, ..., \beta_k, D$. Then the conditional posterior distribution of β_1 is uniform in $\Re$ subject

Table 1: Efficiency of regression-UMM versus LS

	case (a)	case (b)	case (c)	case (d)
$n = 25$	1.712	1.912	1.981	2.231
$n = 50$	1.515	1.832	1.872	1.945
$n = 500$	1.350	1.644	1.750	1.717
$n = 1,000$	1.210	1.355	1.515	1.422
$n = 10,000$	1.07	1.101	1.113	1.130

Notes: The results are based on 10,000 of Monte Carlo replications. The results refer to $b_{1,LS}$ and b_1. The efficiency of $b_{1,LS}$ and b_1 was quite similar to the results reported above. We use 10,000 Monte Carlo simulations to examine the efficiency of LS versus UMM-regression-based techniques. MCMC is implemented using 15,000 passes the first 5,000 of which are discarded during the "burn-in" phase. Initial conditions were obtained from LS and, in all cases, we have $N = 100$ points in the support of the error term.

to the restrictions:

$$\mathbf{\Psi}_1^* \equiv \Psi 1_n - \sum_{j \neq 1} \beta_j \mathbf{x}_j > \beta_1 \mathbf{x}_1 > \psi 1_n - \sum_{j \neq 1} \beta_j \mathbf{x}_j \equiv \boldsymbol{\psi}_1^*. \tag{24}$$

We can draw β_1 (conditional on all other βs) from a uniform distribution subject to the restrictions in (24) which are enforced via rejection sampling. Repeating for each $j = 1, ..., k$ we obtain draws from the posterior conditional distribution of $\beta_j | \beta_{(-j)}, D$, $j = 1, ..., k$. Finally, to obtain draws from the conditional distribution of $\{J_i\}_{i=1}^n$ we have:

$$p(J_i = j | \beta, w, D) \propto \sum_{t=1}^{n} \mathbb{I}\left(a_1 + (j-1)\Delta < y_t - x_t'\beta < a_1 + j\Delta\right),\ j = 1, ..., N. \tag{25}$$

In turn, we normalize $\pi_j = \frac{p(J_i = j | \beta, w, D)}{\sum_{j'=1}^{N} p(J_i = j' | \beta, w, D)}$, and we set $J_i = j$ with probability π_j, $j = 1, ..., N$. The Gibbs sampler yields a sample $\left\{\beta^{(s)}, w^{(s)}, J^{(s)}\right\}_{s=1}^{S}$ which converges to the posterior distribution whose non-normalized density is given in (15), as S increases.

4 Monte Carlo evidence

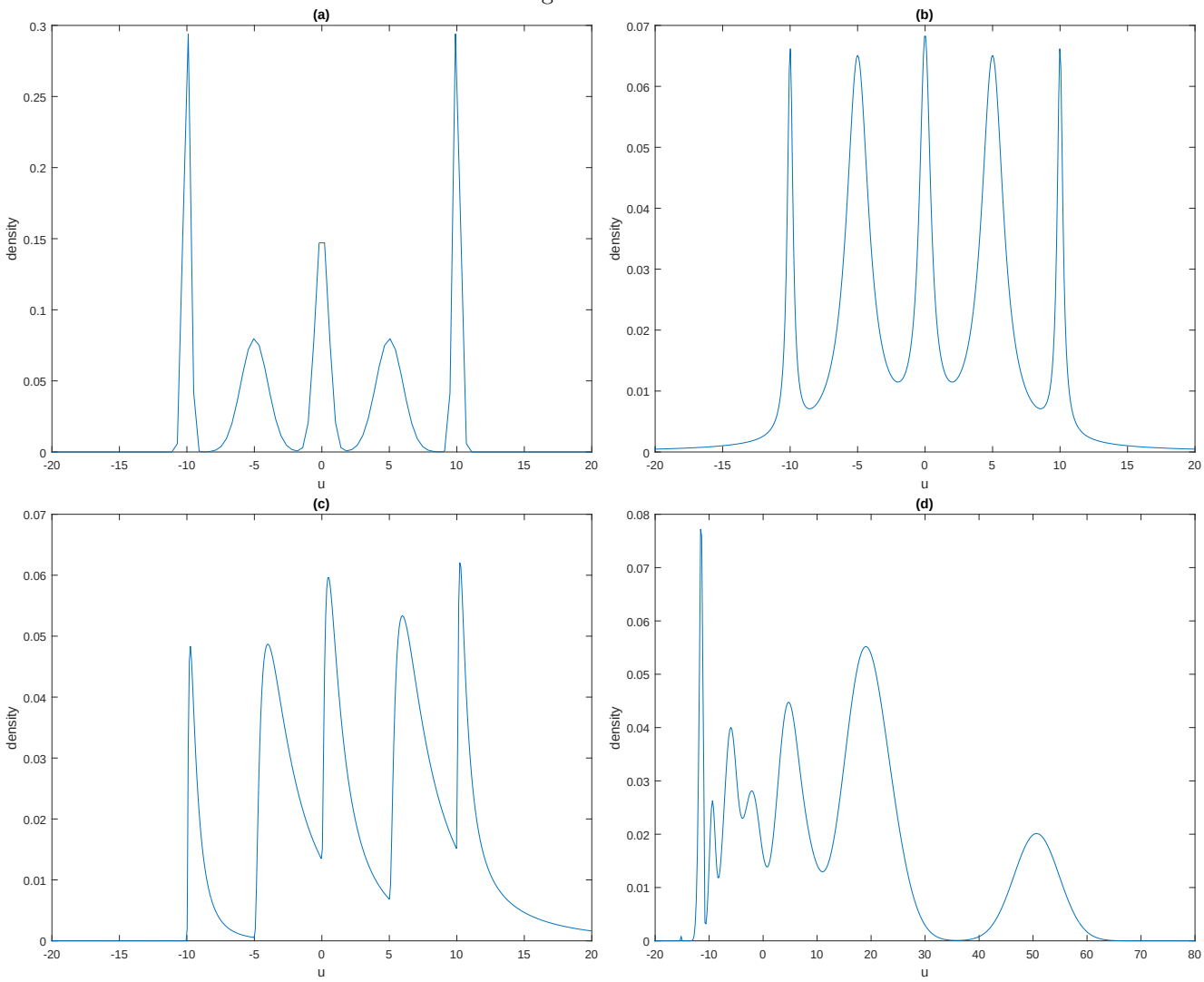
We consider four cases for the distribution of the error term as in Figure 1.

For each case we assume that the sample size is n =25, 50, 100, 500, 1,000 and 10,000. We have two correlated regressors: the first one, $x_{i1} \sim N(0, 1)$ and the second is $x_{i2} = x_{i1} + 0.1\varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$, $i = 1, ..., n$. The regression model is: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$, where u_i is generated according to cases (a) through (d). The true parameter values are: $\beta_0 = 10$, $\beta_1 = 1$, $\beta_2 = -1$.

Our interest focuses on comparing with least squares (LS) regression and the potential improvement in efficiency, which is defined as Eff $= \sqrt{\text{var}(b_{j,LS})/\text{var}(b_j)}$, where $j = 1, 2$, $b_{j,LS}$ is the Bayes posterior mean estimate of β_j from the UMM model, $b_{j,LS}$ is the LS estimator of β_j, and "var" denotes sampling variance. We use 10,000 Monte Carlo simulations to examine the efficiency of LS versus UMM-regression-based techniques. MCMC is implemented using 15,000 passes the first 5,000 of which are discarded during the "burn-in" phase. Initial conditions were obtained from LS and, in all cases, we have $N = 100$ points in the support of the error term.

From the results in Table 1, regression-UMM-based techniques are considerable more efficient compared to LS particularly for "small" samples (i.e. $n \leq 1,000$) although even at $n = 10,000$ the improvement in efficiency is quite evident. With $n = 10,000$ the efficiency is close to unity but still the efficiency of UMM is larger (notice that LS is

Figure 1: Cases



Case (a): A mixture of five normals, with means –10, –5, 0, 5, 10, standard deviations 0.25, 1, 0.5, 1, 0.25, and probabilities 0.2.
Case (b): A mixture of five Student-t densities with one degree of freedom and the same configurations as in Case (a).
Case (c): A mixture of five lognormal densities with one degree of freedom and the same configurations as in Case (a).
Case (d): A mixture of ten Student-t densities with randomly selected means using $N(0, 10^2)$, randomly selected standard deviations using $|N(0, 1)|$ and ten randomly selected probabilities in the interval (0,1) normalized so that they sum up to unity.

Table 2: Bias and efficiency of LS estimator of β_1 and UMM-regression

	$N = 10$	$N = 50$	$N = 100$
bias LS		0.014	
bias UMM	0.012	0.011	0.011
s.e. LS		0.011	
s.e. UMM	0.009	0.007	0.007

Notes: s.e. stands for standard error.

best linear unbiased, but the UMM-regression estimator is not linear so efficiency gains are possible even in quite large samples). Moreover, the regression-UMM-based estimator is, practically, unbiased as it mean squared error and variance are very similar (results available on request). Finally, efficiency gains are largest in cases (b) and (c) where the mixing components are far from normality (viz. Student-t with one degree of freedom and lognormal components).

Another interesting case is to consider $u_i \sim N(0, \sigma^2)$, $i = 1, ..., n$, where σ^2 is estimated using the LS estimator $s^2 = \frac{\sum_{i=1}^{n}(y_i - x_i' b_{LS})^2}{n-k}$, and $b_{LS} = (X'X)^{-1}X'y$. In turn, we know that the support of the error terms is, approximately, $(-3s, 3s)$ (perhaps too "generously"). Even a plot of LS residuals can inform us, at least in large samples, about the support as well as the form of the distribution of errors.

Using the same data generating process as in cases (a), (b), and (c), we examine the bias and efficiency of LS estimator of β_1 and UMM-regression with $n = 100$ but different number of points (N) in the support of UMM-regression. in Table 2.

For example the mean square error (MSE) of LS is $0.011^2 + 0.014^2 = 0.000317$ while the MSE of UMM-regression estimator with $N = 50$ is $0.007^2 + 0.011^2 = 0.00017$ so the ratio of MSEs is almost 1.86. The MSE is lower compared to LS even if we use only $N = 10$ points in the support of the error.

References

Gao, J., An, Z. & Bai, X. (2019). A new representation method for probability distributions of multimodal and irregular data based on uniform mixture model. Annals of Operations Research https://doi.org/10.1007/s10479-019-03236-9

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85 (410), 398–409.