# Robust Bayesian model selection for heavy-tailed linear regression using finite mixtures

Marcos O. Prates*
Departamento de Estatística, Universidade Federal de Minas Gerais, Brazil -
email:marcosop@est.ufmg.br

Flávio B. Gonçalves
Departamento de Estatística, Universidade Federal de Minas Gerais, Brazil -
email:fbgoncalves@est.ufmg.br

Victor H. Lachos
Departamento de Estatística, Universidade Estadual de Campinas, Brazil -
email:hlachos@ime.unicamp.br

## Abstract

In this paper we present a novel methodology to perform Bayesian model selection in linear models with heavy-tailed distributions. The new method considers a finite mixture of distributions to model a latent variable where each component of the mixture corresponds to one possible model within the symmetrical class of normal independent distributions. Naturally, the Gaussian model is one of the possibilities. This allows a simultaneous analysis based on the posterior probability of each model. Inference is performed via Markov chain Monte Carlo - a Gibbs sampler with Metropolis–Hastings steps for a class of parameters. Simulated studies highlight the advantages of this approach compared to a segregated analysis based on arbitrary model selection criteria. Examples with real data are presented and an extension to censored linear regression is introduced and discussed.

**Keywords**: Finite mixture, heavy-tailed errors, linear models, model selection, MCMC.

## 1 Introduction

Statistical practitioners are generally using model selection criteria in order to select a best Bayesian model in different applications. However, Bayesian model selection has been shown not to be an easy task and that each criterion performs better under different situations. For more complex models, it is not clear which criterion is preferable. Recently Gelman et al. (2014) studied and compared different model criteria and concluded that “The current state of the art of measurement of predictive model fit remains unsatisfying.”. From their study it is clear that the criteria fail in selecting the most adequate model under a variety of circumstances. We focus on the problem of considering different approaches to model the error in linear regression models. Previous works have shown the importance of considering more general structures than the Gaussian distribution for this component such as heavy-tailed distributions. This gives rise to the model selection problem for which existent solutions use arbitrary model selection criteria and, therefore, motivates the development of more robust methods.

An interesting way to define a class of heavy-tailed linear regression models, which we will consider, is by using the scale mixtures of normal (SMN) distributions. Andrews and Mallows (1974) use the Laplace transform technique to prove that a continuous random variable $Y$ has a SMN distribution if it can be expressed as follows

$$Y = \mu + \kappa^{1/2}(U)Z,$$

where $\mu$ is a location parameter, $Z$ is a normal random variable with zero mean and variance $\sigma^2$, $\kappa(U)$ is a positive weight function, $U$ is a mixing positive random variable with density $h(. \mid \boldsymbol{\nu})$ and $\boldsymbol{\nu}$ is a scalar or parameter vector indexing the distribution of $U$. As in Lange and Sinsheimer (1993) and Chow and Chan (2008), we restrict our attention to the case where $\kappa(U) = 1/U$, that is, the normal independent (NI) class of distributions. Thus, given $U$, $Y \mid U = u \sim \mathcal{N}(\mu, u^{-1}\sigma^2)$ and the pdf of $Y$ is given by

$$f(y \mid \mu, \sigma^2, \boldsymbol{\nu}) = \int_0^\infty \phi((y-\mu)/\sqrt{u^{-1}\sigma^2})h(u \mid \boldsymbol{\nu})du. \tag{1}$$

From a suitable choice of the mixing density $h(. \mid \boldsymbol{\nu})$, a rich class of continuous symmetric distributions can be described by the density given in (1) to accommodate thicker-tails than the normal distribution. Note that when $U = 1$ (a degenerate random variable), we retrieve the normal distribution. Apart from the normal model, we explore two different types of

The aim of this paper is to propose a general formulation to perform Bayesian model selection for heavy-tailed linear regression models in a simultaneous setup. That is achieved by specifying a full model which includes the space of all individual models under consideration, which are specified using the SMN approach described above. This way, the model selection criterion can be based on the posterior probability of each model. A mixture distribution is adopted to one of the full model's variable, with each component of the mixture referring to one of the individual models. This approach has two main advantages when compared to an ordinary analysis where each model is fitted separately and some model selection criterion is used. Firstly, there is a significant gain in the computational cost. Secondly, the model selection criterion is fully based on the Bayesian Paradigm and, therefore, is more robust for different choices of individual models then some other arbitrary model selection criteria such as DIC, EAIC, EBIC (Spiegelhalter et al., 2002), CPO (Geisser and Eddy, 1979) WAIC (Watanabe, 2010) and others. The posterior distribution of the unknown quantities has a significant level of complexity which motivates the derivation of a MCMC algorithm to obtain a sample from this distribution.

This paper is organised as follows: Section 2 presents the general model for simultaneous analysis; a real data set is shown in Section 3. Finally, Section 4 discusses some extension of the proposed methodology.

## 2 Linear regression model with heavy-tailed mixture structured errors

Model selection is an important and complex problem in statistical analysis and the Bayesian approach is particularly appealing to solve it. In particular, the use of mixtures is a nice way to pose and solve the problem, whenever possible. It allows for an analysis where all models are considered and compared in a simultaneous setup without the need of complicated reversible jump MCMC algorithms. Note that, from (1), each model is determined by the distribution of the scale factor $u_i$, $\forall i$, which suggests that a mixture distribution could be used for this latent variable. We present a general finite mixture model framework capable of capturing different behavior of the response and indicate which individual distribution is preferred, if any.

### 2.1 The model

Define the $n$-dimensional response vector $\mathbf{Y}$, the $n \times q$ design matrix $\mathbf{X}$, the $q$-dimensional coefficient vector $\boldsymbol{\beta}$ and two 3-dimensional vectors $\boldsymbol{\gamma} = (\gamma_1 \ \dots \ \gamma_3)'$ and $\mathbf{p} = (p_1 \ \dots \ p_3)'$. Finally, let $diag(\mathbf{u}^{-1})$ be a $n$-dimensional diagonal matrix with $i$-th diagonal $u_i^{-1}$, $i = 1, \dots, n$. We derive the algorithm considering the three most common choices in the NI family - Normal, t-Student, Slash. We propose the following general model:

$$\begin{aligned}
(\mathbf{Y}|Z_j = 1) &\sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\gamma_j diag(\mathbf{u}^{-1})\right) &(2)\\
\mathbf{Z} &\sim Mult(1, p_1, p_2, p_3) &(3)\\
U_i &\overset{iid}{\sim} \begin{cases} \delta_1, & \text{if } Z_1 = 1\\ \mathcal{G}\left(\nu_t/2, \nu_t/2\right), & \text{if } Z_2 = 1,\ i = 1, \ldots, n,\\ \mathcal{B}(\nu_s, 1), & \text{if } Z_3 = 1 \end{cases} &(4)\\
\gamma_j &= \begin{cases} 1, & \text{if } Z_1 = 1\\ (\nu_t - 2)/\nu_t, & \text{if } Z_2 = 1\\ (\nu_s - 1)/\nu_s, & \text{if } Z_3 = 1, \end{cases} &(5)
\end{aligned}$$

This way, marginally, $\mathbf{Y}$ has a distribution from the NI family with heavy tail behavior. Specifically, $\nu_t$ and $\nu_s$ are degrees of freedom for the Student-t and Slash distributions, respectively.

The particular structure chosen for the variance in (2) was thought of in order to avoid identifiability issues. The function $\gamma_j$ has to be specified in a way such that, for each $j$, the variance of the model is the same - $\sigma^2$, which makes this parameter interpretable and allows us to treat it as a common parameter to all of the individual models. Model selection is also more efficient in the sense that it is focuses on the tail behavior of the observations. Finally, this also contributes to speed the convergence of the MCMC algorithm.

Note that each component from the mixture distribution of $u_i$ corresponds to one of the models being considered. Model selection is made through the posterior distribution of $\mathbf{Z}$. A subtle but important point here is the fact that there is no $i$ index for $Z$. This means that we assume that all the observations come from the same model, which poses the inference problem in the model selection framework.

Another advantage of the simultaneous approach is that it allows the use of Bayesian model averaging (see Raftery et al., 1996). This is particularly useful in cases where more than one model have a significant posterior probability, which is a typical case for the class of models under consideration. Note that the models can be quite similar in some situations - specially for higher values of the degrees of freedom (df) parameters.

# 3 Application

## 3.1 AIS

In this section we introduce a biomedical study realised by the Australian Institute of Sports (AIS) in 202 athletes (Cook and Weisberg, 1994). To exemplify our modeling we consider the body mass index (BMI) as our response and the the percentage of body fat (Bfat) as our covariate. This way, we have the fitting model (2)-(5) with $\mathbf{X}_{i\cdot} = (1, \text{Bfat}_i)$ for $i = 1, \ldots, 202$.

Initially, each model of our mixture, Normal, Student-t and Slash was fitted separately. A Markov Chain of $110k$ iterations was ran for each one with a burn-in period of $10k$. After that, we used the model selection criterion to determine which was the preferred one. Table 1 shows the model selection results.Notice, that in Table 1 we present $-$LPML, this way, for all the criteria, smaller means better fit. From Table 1 we can see that the Slash model is the preferred one. Although the difference between the criteria from the Slash and Student-t models and the Normal models are significantly large, the difference between the Student-t and Slash models are much smaller, specially for the $-$LMPL and WAIC.

Table 2 summarises the posterior results for the Slash and mixture models. The percentage of body fat has a significant positive impact in the BMI as expected. The posterior mean of the degrees of freedom for both Student-t and Slash distribution are estimated to be small, presenting a divergence to the traditional Normal assumption. More interestingly, the mean posterior estimates for $\mathbf{p}$ is $\hat{\mathbf{p}} = (0.001, 0.304, 0.695)$ for the Normal, Student-t and Slash, respectively. Clearly, we see that the posterior distribution of the

Table 1: Model selection criterion for the fitting of the Normal, Student-t and Slash regression models.

| Models | −LPML | DIC | EAIC | EBIC | WAIC |
|---|---|---|---|---|---|
| Normal | 498.497 | 2976.407 | 994.142 | 1000.758 | 996.971 |
| Student-t | 491.623 | 2935.009 | 982.059 | 991.984 | 983.210 |
| Slash | 491.033 | 2931.636 | 980.633 | 990.558 | 982.049 |

Table 2: Posterior results for the BMI analysis with Bfat as covariate for the robust mixture model. The posterior mean, median a standard deviation (Sd) are presented as well as the 95% high posterior density (HPD) interval.

| Model | Parameters | Mean | Median | Sd | 95% HPD interval |
|---|---|---|---|---|---|
| Slash Model | $\beta_0$ | 21.810 | 21.810 | 0.419 | (20.980, 22.620) |
| | $\beta_1$ | 0.070 | 0.070 | 0.028 | (0.015, 0.126) |
| | $\sigma^2$ | 10.093 | 8.989 | 3.587 | (5.702, 17.940) |
| | $\nu_s$ | 1.705 | 1.612 | 0.442 | (1.110, 2.569) |
| Slash Selected Model | $\beta_0$ | 21.794 | 21.799 | 0.418 | (21.022, 22.667) |
| | $\beta_1$ | 0.071 | 0.071 | 0.028 | (0.016, 0.128) |
| | $\sigma^2$ | 9.200 | 8.462 | 2.954 | (5.543, 14.765) |
| | $\nu_s$ | 1.716 | 1.628 | 0.434 | (1.111, 2.549) |

mixture model have 30.4% chance of the data be better adjusted by the Student-t residuals while a 69.5% chance of a better fit from the Slash distribution. Therefore, from the mixture model the Slash distribution is the preferred one. As expected, $\nu_s$ is closely estimated in both proposals.

# 4 Conclusions and some extensions

Our proposed methodology has shown considerable flexibility to perform model selection over heavy-tailed data explained by covariates under a regression framework. From theoretical arguments, simulation studies and application to real datasets, it is clear that the methodology provides a robust alternative to select the best model instead of relying on model selection criteria which can be unstable (Gelman et al., 2014). Also, we argue that fitting a more complete model is more effective and computationally efficient then fitting 3 separate models. In addition, extension to include more distributions in the finite mixture is almost direct. It is clear from our results that this finite mixture idea can be used in a variety of problems where a common parametrization exists for a family of distributions.

Besides the computational advantage of fitting one general model instead of 3 separated models. We also emphasize that our robust model selection framework automatically perform multiple comparison between the 3 models, which gives an advantage if one, instead, prefer to use the Bayes factor performing 2 by 2 comparisons in each individual model.

Although the presented methodology enriches the class of the traditional cenrosed regression models, we conjecture that the methodology presented in this paper may not provide satisfactory result when the response exhibit assimetry besides the non-normal behavior. To overcome this limitation extending the work to account for skewness behavior is also a possibility, for example by using the scale mixtures of skew-normal (SMSN) distributions proposed in Lachos et al. (2010). Nevertheless, a deeper investigation of those modifications in the parameterisation and implementations is beyond the scope of the present paper, but provides stimulating topics for further research. Another possibility of future research is to generalise these modeling framework to linear mixed model, e.g., clustered, temporal or spatial dependence. These extensions are being studied in a different manuscript.

## Acknowledgements

The second author would like to thank FAPEMIG-Brazil and CNPq-Brazil for partial financial support. Lachos was supported by grants 305054/2011-2 from CNPq-Brazil, and 2011/17400-6 from FAPESP-Brazil.

## References

Andrews, D. F. and S. L. Mallows (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B 36*, 99–102.

Chow, S. T. B. and J. S. K. Chan (2008). Scale mixtures distributions in statistical modelling. *Australian & New Zeland Journal of Statistics 50*, 135–146.

Cook, R. D. and S. Weisberg (1994). *An Introduction to Regression Graphics*. New York: Wiley.

Geisser, S. and W. F. Eddy (1979). A predictive approach to model selection (Corr: V75 p765). *Journal of the American Statistical Association 74*, 153–160.

Gelman, A., J. Hwang, and A. Vehtari (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing 24*, 997–1016.

Lachos, V., P. Ghosh, and R. Arellano-Valle (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica 20*, 303–322.

Lange, K. L. and J. S. Sinsheimer (1993). Normal/independent distributions and their applications in robust regression. *J. Comput. Graph. Stat 2*, 175–198.

Raftery, A., D. Madigan, and C. Volinsky (1996). Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In *Bayesian Statistics 5*. Oxford University Press.

Spiegelhalter, D., N. Best, B. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B 64*, 583–639.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research 11*, 3571–3594.