ATLANTIS PRESS
Journal of Statistical Theory and Applications, Vol. 16, No. 1 (March 2017) 53–64

Bayesian analysis of hierarchical heteroscedastic linear models using Dirichlet-Laplace priors

S.K. Ghoreishi
Department of Statistics, Faculty of Sciences, University of Qom,
Qom, I. R. of Iran

Received 13 December 2015

Accepted 8 May 2016

From practical point of view, in a two-level hierarchical model, the variance of second-level usually has a tendency to change through sub-populations. The existence of this kind of local (or intrinsic) heteroscedasticity is a major concern in the application of statistical modeling. The main purpose of this study is to construct a Bayesian methodology via shrinkage priors in order to estimate the interesting parameters under local heteroscedasticity. The suggested methodology for this issue is to use of a class of the local-global shrinkage priors, called Dirichlet-Laplace priors. The optimal posterior concentration and straightforward posterior computation are the appealing properties of these priors. Two real data sets are analyzed to illustrate the proposed methodology.

Keywords: Global-local priors, Heteroscedasticity, Hierarchical models, SURE estimators.

2000 Mathematics Subject Classification: 62F15; 62F30

1. introduction

The main objective of this paper is to develop a Bayesian approach for local heteroscedastic hierarchical linear models using Dirichlet-Laplace priors. Hierarchical modeling provides an efficient approach which combines partial information to achieve accurate and stable inference about interesting parameters. In this context, risk properties of shrinkage estimators, admissible minimax estimators, and many other estimators of hierarchical models have been noted by many authors under different loss functions. The literature covers both homoscedastic and heteroscedastic models. It seems heteroscedastic hierarchical modeling, which assumes unequal subpopulation variances, is more popular in real world than homoscedastic hierarchical modeling. For more detail on the subject, see Berger and Strawderman(1996) and Brown and Greenshtein (2009).
One of serious challenge in Bayesian analysis of this kind of modeling is how to estimate the hyperparameters which are initiated from assumed prior distributions. The situation gets worse if one encounters with some locally unknown sources(or hyperparameters) which cause the heteroscedasticty. This phenomenon occurs in regression modeling over sub-populations and covariance analysis especially for high-dimensional data. For big data, rapid computation of point estimates of parameters(hyperparameters) of interest along with the uncertainty associated with them has been the main focus of investigation, Efron et al. (2004).

Xie et al. (2012) proposed a class of shrinkage estimators that can be readily applied in heterpscedastic hierarchial normal models. Indeed, their motivation was to illustrate the comparability of various shrinkage estimators and their 'optimal' properties. Their assumed model for subpopulations $i = 1, 2, \cdots, n$ was

$$\begin{aligned} y_i &\sim N(\theta_i, A_i) \\ \theta_i &\sim N(\mu, \sigma^2), \end{aligned}$$

where A_is are some known and possibly distinct real values and σ^2 and μ are hyperparameters.
Ghoreishi and Meshkani (2014) proposed and developed a class of shrinkage estimators that can be readily applied in a two-level heteroscedastic hierarchical normal models. In this framework, they assumed one or several explanatory variables produce the heteroscedasticty. They showed that negligence in considering this fundamental assumption can lead to substantial bias in the estimates. Their assumed model was

$$\begin{aligned} y_i &\sim N(\theta_i, g(z_i)), \\ \theta_i &\sim N(\mu, \sigma^2 h(z_i)), \end{aligned} \tag{1.1}$$

where z_i is an explanatory variable(or a vector of explanatory variables of order k in $\mathscr{D}$), global quantities $\sigma^2 > 0$ and μ are hyperparameters, $h : \mathscr{D} \subseteq \mathbb{R}^k \to \mathbb{R}^+$ is known function and finally $g : \mathscr{D} \subseteq \mathbb{R}^k \to \mathbb{R}^+$ is either completely known or will be known by some plug-in robust estimators.
We believe assuming that the heteroscedasticty arises only from some explanatory variables in the second level of model (1.1) is an exception to the rule. Because, in the real data, there are some local unknown source of disturbance in the variances of hierarchical models which move over subpopulations. Moreover, a model misspecification may lead to this case. However, it is important to note that this actual assumption involves several parameters(hyperparameters) in the model which the analyst has to find a way to estimate them and investigate their properties. So, we are interested in extending the model (1.1) to

$$\begin{aligned} y_i &\sim N(\theta_i, g(z_i)) \\ \theta_i &\sim (\mu, \sigma_i^2 h(z_i)), \end{aligned} \tag{1.2}$$

for subpopulations $i = 1, 2, \cdots, n$, where μ and local unknown quantities $\sigma_i^2 > 0$ are hyperparameters. The functions g and h are defined as in model (1.1). As we illustrated before, the model (1.2) is popular in regression modeling over subpopulations and covariance analysis with usually many factor levels where one can not naturally control the homoscedasticity for each factor level over experimental duration.
As one can see, in addition to the fact that the variances of both levels of model (1.2) depend on explanatory variable z_i, some other unknown local quantities $\sigma_i^2 > 0$ also have essential role in disturbing the variances.
Our main purpose in this work is to construct a statistical inference methodology via shrinkage priors for hyperparameters μ and σ_i^2s. In this line, we hope to introduce an alternative Bayesian variable selection strategy for choosing those variables which have effective role in altering the homoscedasticity. Equivalently, this reduces to the choice of the variables that should be included in the model, Miller (2002) and Broman and Speed (2002). Of course, those variable selection methods that can be implemented easily in the MCMC framework are needed, Dellaportas et al.(2000) and Robert and Casella(2004).

The rest of the article is organized as follows. Section 2 gives the general notations and definitions. In Section 3, we develop our theoretical results. Section 4 illustrates the performance of our methodology by analyzing two real datasets.

2. Preliminaries

2.1. *Bayes Shrinkage estimators*

Let $y_i \sim N(\theta_i, g(z_i)),\ (i = 1, \cdots, n)$, be n independent normal observations for given $z_i \in \mathscr{D} \subseteq \mathbb{R}^k$. Again, consider the model (1.2)

$$\begin{aligned} y_i &\sim N(\theta_i, g(z_i)) \\ \theta_i &\sim (\mu, \sigma_i^2 h(z_i)), \end{aligned}$$

where $g,\ h : \mathscr{D} \subseteq \mathbb{R}^k \to \mathbb{R}^+$ are known functions and μ and $\sigma_i^2 > 0$ are hyperparameters for subpopulations $i = 1, 2, \cdots, n$.
From Bayes' theorem, the posterior distribution of θ_i is

$$N(\frac{\sigma_i^2 h(z_i)}{\sigma_i^2 h(z_i) + g(z_i)} y_i + \frac{g(z_i)}{\sigma_i^2 h(z_i) + g(z_i)} \mu, \frac{\sigma_i^2 h(z_i) g(z_i)}{\sigma_i^2 h(z_i) + g(z_i)}).$$

Assuming the weighted squared loss function,

$$l_q(\theta_i, \hat{\theta}_i) = \frac{1}{\sum q_i} \sum q_i (\theta_i - \hat{\theta}_i)^2, \tag{2.1}$$

where $q_i = \frac{1}{g(z_i)}$, the Bayes shrinkage estimators are of the form

$$\hat{\theta}_i = \frac{\sigma_i^2 h(z_i)}{\sigma_i^2 h(z_i) + g(z_i)} y_i + \frac{g(z_i)}{\sigma_i^2 h(z_i) + g(z_i)} \mu. \tag{2.2}$$

In practice, one may prefer the shrinkage estimators (2.2) to previous ones since the local nuisance parameters σ_i^2 and the explanatory variable z_i, which have important roles in disturbing the variances of model (1.2), are involved to build the shrinkage factors $\frac{\sigma_i^2 h(z_i)}{\sigma_i^2 h(z_i) + g(z_i)}$ and $\frac{g(z_i)}{\sigma_i^2 h(z_i) + g(z_i)}$.
Using the Bayes shrinkage estimators (2.2) in the loss function (2.1), then respectively minimizing this loss function, $l_q(\theta_i, \hat{\theta}_i^{\sigma_1^2, \cdots, \sigma_n^2, \mu})$, and its unbiased estimator,

$$MGS(\sigma_1^2, \cdots, \sigma_n^2, \mu) = \frac{1}{\sum q_i} \sum \{ \frac{g(z_i)(y_i - \mu)^2}{(\sigma_i^2 h(z_i) + g(z_i))^2} + \frac{\sigma_i^2 h(z_i) - g(z_i)}{\sigma_i^2 h(z_i) + g(z_i)} \},$$

with respect to μ and σ_i^2s leads to the 'oracle' estimators $(\mu^{OL}, \sigma_1^{2(OL)}, \cdots, \sigma_n^{2(OL)})$ and SURE estimators $(\mu^{SURE}, \sigma_1^{2(SURE)}, \cdots, \sigma_n^{2(SURE)})$. The SURE estimators are the solution of the following equations.

$$\begin{aligned} \sigma_i^2 &= \{ \frac{(y_i - \mu)^2 - g(z_i)}{h(z_i)} \}^+, \ \ i = 1, \cdots, n, \\ \mu &= \frac{\sum \frac{g(z_i) y_i}{(\sigma_i^2 h(z_i) + g(z_i))^2}}{\sum \frac{g(z_i)}{(\sigma_i^2 h(z_i) + g(z_i))^2}}. \end{aligned} \tag{2.3}$$

In addition to the SURE estimators (2.3), the empirical Bayes estimators including moment and maximum likelihood estimators are also obtained from the marginal distribution of y_is. For more details on this subject when σ_i^2s are assumed equal, see Ghoreishi and Meshkani (2014).

2.2. *Local-global shrinkage priors*

Although the SURE estimators $\hat{\mu}^{SURE}$ and $\hat{\sigma}_i^{2(SURE)}$, (2.3), have acceptable asymptotic performance as n goes to infinity, solving $n+1$ non-linear equations often produce zero values for many σ_i^2s and therefore it may misleads us to some biased estimates. So, we have to find an approach to deal with this problem. Our suggested Bayesian approach for this issue is using of a class of priors which are constructed based on the local-global shrinkage priors, called Dirichlet-Laplace priors, Bhattacharya et al. (2014). These priors possess optimal posterior concentration and lead to efficient and straightforward posterior computation, Gibbs sampler, exploiting results from normalized random measure theory. A brief discussion on these priors is given below.
Bhattacharya et al. (2014) investigated the theoretical properties of a whole class of local-global shrinkage priors. Indeed, they derived posterior concentration rates for a global-local shrinkage prior. Their assumed model was a single observation corrupted with i.i.d. standard normal noise:

$$\begin{aligned} y_i &= \theta_i + \varepsilon_i; \quad 1 \le i \le n \\ \theta_i &\sim N(0, \sigma_i^2), \end{aligned} \tag{2.4}$$

where $\varepsilon_i \sim N(0,1)$. In their approach, Bhattacharya et al. (2014) assumed $\sigma_i^2 = \psi_i \phi_i^2 \tau^2$, where τ controls global variation while the local scales ϕ_i allow deviations in the degree of variations and finally, the volatility variables ψ_i are i.i.d. samples from an exponential distribution with scale parameter 1/2. Moreover, for given constant a, they assumed a Dirichlet distribution for $(\phi_1, \cdots, \phi_n)$ and a gamma distribution for hyper-parameter τ. That is,

$$\begin{aligned} (\phi_1, \cdots, \phi_n) &\sim Dir(a, \cdots, a) \\ \tau &\sim G(na, 1/2). \end{aligned} \tag{2.5}$$

It is easy to see that the model (2.4) and its hierarchical priors (2.5), can be rewritten as

$$\begin{aligned} y_i &\sim N(\theta_i, 1) \\ \theta_i | \phi_i, \tau, \psi_i &\sim N(0, \psi_i \phi_i^2 \tau^2) \\ \psi_i &\sim Exp(1/2) \\ \phi_i &\sim Beta(a, (n-1)a) \\ \tau &\sim G(na, 1/2), \end{aligned} \tag{2.6}$$

or equivalently,

$$\begin{aligned} y_i &\sim N(\theta_i, 1) \\ \theta_i | \phi_i, \tau &\sim DE(0, \tau \phi_i) \\ \phi_i &\sim Beta(a, (n-1)a) \\ \tau &\sim G(na, 1/2), \end{aligned} \tag{2.7}$$

where $DE(\mu, \theta)$ denotes a zero mean Laplace double exponential distribution with density function $f(x|\mu, \theta) = (2\theta)^{-1} e)^{-|x-\mu|/\theta}$.

Although, this model may find some applications for real data, usually one primary objective in statistical modeling is to assess certain regressors that play essential role in predicting response variable. Furthermore, selecting a small set of "covariates" is an important task too. So, the aim of this work is to adapt model (2.6) to the case of model (1.2), hoping to remove the challenges (mentioned before) that may arise in estimating of the SURE estimators. Moreover, we expect to reach an easy-to-compute estimate, an MCMC approach, applicable for all sizes of data.
Our developed model, with respect to (1.2), is

$$\begin{aligned} y_i &\sim N(\theta_i, g(z_i)) \\ \theta_i|\phi_i,\tau,\psi_i &\sim N(\mu, \psi_i\phi_i^2\tau^2 h(z_i)) \\ \mu &\sim (\mu_0, A_0) \\ \psi_i &\sim Exp(1/2) \\ \phi_i &\sim Beta(a,(n-1)a) \\ \tau &\sim G(na,1/2) \\ a &\sim G(a_0,b_0), \end{aligned} \tag{2.8}$$

where σ_i^2 in (1.2) is equal to $\psi_i\phi_i^2\tau^2$, and μ_0 and A_0 are some known constants. Integrating over ψ leads to the equivalent model

$$\begin{aligned} y_i &\sim N(\theta_i, g(z_i)) \\ \theta_i|\phi_i,\tau,\mu &\sim DE(\mu, \tau\phi_i h^{1/2}(z_i)) \\ \mu &\sim (\mu_0, A_0) \\ \phi_i &\sim Beta(a,(n-1)a) \\ \tau &\sim G(na,1/2) \\ a &\sim G(a_0,b_0). \end{aligned} \tag{2.9}$$

From Bayesian perspective, some appealing and applicable properties of hierarchical model (2.9) are as follows:

a) From practical point of view, one of interesting properties of the model (2.9) is that it allows incorporation of the explanatory variable z_i into analysis in order to highly control the local scale variation.
b) Although some beautiful classes of shrinkage priors, within the Gaussian global-local scale mixture family, have been proposed by many statisticians, Dirichlet-Laplace priors, given in (2.9), have minimax optimality under weak sparsity assumptions from a frequentist perspective, which is crucial in studying posterior contraction in high-dimensional settings, for more details see Bhattacharya et al. (2014).
c) Model (2.9) induces a fully heteroscedasticity for hierarchical models, so it is suitable for any weighted/unweighted regression analysis, covariance analysis or even for regression error detections whenever the analyst is challenged due to the lack of exact zero errors.
d) It is important to note that one may be interested in extending model (2.9) from a simple linear model framework to a multiple one with p regressors such that p may be less, equal, or greater than the number of subjects. In such a case, assuming explanatory variables $z_1,\cdots,z_p$, the function h can be rewritten as $h_p(z_1,\cdots,z_p)$ where subscript p is to

show the number of the explanatory variables. The primary objective in our Bayesian variable selection methodology is to choose a subset of regressors that play important role in constructing coefficients of shrinkage estimators (2.2). Without loss of generality, assume that the first $r(r \leq p)$ variables are effective and hence, the rest can be ignored. In this case, using $h_r(z_1, \cdots, z_r)$ instead of $h_p(z_1, \cdots, z_p)$ should not affect, effectively, the predictions of response y. Therefore, this approach introduces an alternative Bayesian model selection with respect to the other competitors.

e) An interesting property of model (2.9) is that one can evaluate the validity of hypothesis $H_0 : \sigma_1^2 = \cdots = \sigma_n^2$ by computing the posterior probability

$$Prob = \Pr(\max(\sigma_i) - \min(\sigma_i) < \eta \mid y_1, \cdots, y_n), \tag{2.10}$$

for some tiny value η, which is priori chosen by the analyst. The small value of the probability (2.10) points to rejection of hypothesis H_0.

Although these properties are important in their owns, our practical aim is to obtain the posterior distribution (or at least some features) of the parameter of interest θ_i. Therefore, in order to sample from posterior distributions, the posterior full conditional of θ_i is needed. In the next section, we illustrate the subject with more details.

3. Theoretical Results

In this section, we establish main results containing two theorems. The first is to illustrate the asymptotic properties of SURE estimators (2.3). In the previous section, we mentioned that our primary purpose in applying model (2.9) is how one can sample from posterior distributions. Therefore, the second theorem provides the full conditional posteriors.

Theorem 3.1. *Assume the conditions (C1)-(C3) given in the appendix and let $\sigma_i^2 = \tau^2 \psi_i \phi_i^2$. Then for shrinkage estimators (2.2), we have*

$$\sup_{\substack{0 \leq \tau^2 \leq \infty,\ 0 \leq \psi_i \leq \infty, \\ 0 \leq \phi_i^2 \leq \infty,\ |\mu| \leq M \\ i=1,\cdots,n,}} \mid MGS(\sigma_1^2, \cdots, \sigma_n^2, \mu) - l_q(\theta, \hat{\theta}_t^{\sigma_1^2, \cdots, \sigma_n^2, \mu}) \mid \to 0 \text{ in } L^2 \text{ as } n \to \infty.$$

The following result is obvious. It shows that in probability (as $n \to \infty$)

$$\mid l_q(\theta, \hat{\theta}_t^{\sigma_1^{2(SURE)}, \cdots, \sigma_n^{2(SURE)}, \mu^{SURE}}) - l_q(\theta, \hat{\theta}_t^{\sigma_1^{2(OL)}, \cdots, \sigma_n^{2(OL)}, \mu^{OL}}) \mid \to 0.$$

This means that under conditions (C1)-(C3), the SURE estimators are asymptotically as good as the oracle estimators. Here, we assume
$\sigma_i^{2(SURE)} = \tau^{2(SURE)} \psi_i^{SURE} \phi_i^{2(SURE)}$ and $\sigma_i^{2(OL)} = \tau^{2(OL)} \psi_i^{OL} \phi_i^{2(OL)}$.
Since the loss function $l_q(\theta, \hat{\theta}_t^{\sigma_1^2, \cdots, \sigma_n^2, \mu})$ is a convex function of μ and shrinkage factors $U = \frac{\tau^2 \psi_i \phi_i^2 h(z_i)}{\tau^2 \psi_i \phi_i^2 h(z_i) + g(z_i)}$, the optimal Bayesian estimators, alternative to the SURE estimators (2.2), are given as

$$\hat{\theta}_i^B = E(U|y_i) y_i + E\{(1-U)\mu | y_i\}. \tag{3.1}$$

For numerical computation, we need the following theorem.

Theorem 3.2. *The full conditional distributions for efficient posterior computation of model (2.8) are as follows:*

i). $\theta_i|(\psi_i,\phi_i,\tau,\mu,z_i,y_i) \sim N(\mu_i,V_i^2)$, *where*

$$\mu_i = (1+\frac{g(z_i)}{\tau^2\psi_i\phi_i^2h(z_i)})^{-1}y_i + (1+\frac{\tau^2\psi_i\phi_i^2h(z_i)}{g(z_i)})^{-1}\mu,$$
$$V_i^2 = (\frac{1}{g(z_i)}+\frac{g(z_i)}{\tau^2\psi_i\phi_i^2h(z_i)})^{-1}.$$

ii). $\mu|(\psi_i,\phi_i,\theta_i,\tau,z_i) \sim N(\mu_1,A_1)$

$$\mu_1 = (\frac{1}{A_0}+\sum_i\frac{1}{\tau^2\psi_i\phi_i^2h(z_i)})^{-1}(\frac{\mu_0}{A_0}+\sum_i\frac{\theta_i}{\tau^2\psi_i\phi_i^2h(z_i)}),$$
$$A_1 = (\frac{1}{A_0}+\sum_i\frac{1}{\tau^2\psi_i\phi_i^2h(z_i)})^{-1}.$$

Practically, we ignore this item if we assume $\mu=0$ *in model (2.8).*

iii). $\psi_i|(\phi_i,\theta_i,\tau,\mu,z_i) \sim iG(\frac{\phi_i\tau h^{1/2}(z_i)}{|\theta_i-\mu|},1)$,*where iG is an inverse-Gaussian distribution.*

iv). $\tau|(\phi_i,\theta_i,\mu,a,z_i) \sim giG(n(a-1),1,2\sum_{i=1}^n\frac{|\theta_i-\mu|}{h^{1/2}(z_i)\phi_i})$, *where giG, defined below, denotes a generalized inverse-Gaussian distribution.*

v). $\phi_i|(\theta_i,\mu,a,z_i) \sim T_i/\sum_{i=1}^n T_i$, *and* $T_1,\cdots,T_n$ *are distributed independently according to* $T_i \sim giG(a-1,1,\frac{2|\theta_i-\mu|}{h^{1/2}(z_i)})$. *By definition,* $X \sim giG(p,a,b)$ *if* $f(x)=\frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})}x^{p-1}e^{-(ax+b/x)/2}$, *where* K_p *is a modified Bessel function of the second kind with* $a>0$, *and* $b>0$.

vi). $a|(\phi_i,\tau) \sim \pi(a) \propto \frac{\Gamma(na)}{\Gamma(a)^n}a^{a_0-1}e^{-\{b_0-n\log(\tau)-\Sigma\log(\phi_i)\}a}$, *where* $\Gamma(x)=(x-1)!$. *In this case, the Metropolis-Hastings algorithm is used to produce samples for* a.

The last five items show that the joint distribution of $(\psi_i,\phi_i,\tau,\mu,a)|z_i$ depends on the response y_i through θ_i. So the joint posterior distribution of $(\psi_i,\phi_i,\tau,\mu,a)|z_i$ is conditionally independent of y_i given θ_i. Therefore, the introduced sampler cycles in Theorem 2 give us a draw from the posterior distribution of θ_i, which our concentration is on its entries.
It is important to note that the estimators (3.1) are different from the marginal mean of θ_i produced from Theorem 2(i). We call this estimator, θ_i^P, and evaluate its performance in the next section.

4. Application

In this section, we apply our methodology for Bayesian analysis of hierarchical heteroscedastic linear models using Dirichlet-Laplace priors. To illustrate its performance, we have considered two examples including 1: a simple weighted hierarchical regression model for the study of heteroscedasticity and 2: a multiple linear regression with two explanatory variables for variable selection purposes.

4.1. *Simple weighted regression model*

Consider Bid at auction data in Table (1), see Ghoreishi and Mostafavinia (2015). These data belong to a big state company which wanted to survey its recent 12 auctions. It contains an explanatory variable z : Bid at auction(in million dollars) and response variable y : Cost of auction(in million dollars).

The least squares regression shows $R^2 = 0.880$ and the mean square error $MSE = 34.45$, whereas the results of the weighted least square, where the weights are proportional to $\frac{1}{z^2}$, lead to $R^2 = 0.915$ and the weighted mean square error $WMSE = 0.792$. This means the weighted least square provides a suitable fit. Therefore, we consider the following simple hierarchical heteroscedastic regression

Table 1. The Bid at Auction data and corresponding shrinkage estimates

NO.	Bid at auction(z)	Cost of auction (Y)	$\hat{\theta}^P$
1	2.13	15.5	15.94
2	1.21	11.1	11.39
3	11.00	62.6	59.23
4	6.00	35.4	32.98
5	5.60	24.9	22.45
6	6.91	28.1	24.72
7	2.97	15.0	15.76
8	3.35	23.2	21.78
9	10.39	42.0	36.86
10	1.10	10.0	10.24
11	4.36	20.0	19.07
12	8.00	47.5	44.60

model:

$$y_i \sim N(\theta_i, \delta^2 \frac{1}{z_i^2})$$
$$\theta_i = \beta_{0i} + \beta_{1i} z_i,$$

and following Edwards et al. (1963), we assume

$$\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} \sim N_2(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \sigma^2 \mathbf{V}(z_i))$$

where (β_{0i}, β_{1i}) are independent for $i = 1, 2, \cdots, n$ and $\mathbf{V}$ is a known 2×2 covariance matrix. Therefore it is easy to see that

$$\begin{aligned} y_i &\sim N(\theta_i, \delta^2 \frac{1}{z_i^2}) \\ \theta_i &\sim N(\mu, \sigma^2 \begin{pmatrix} 1 & z_i \end{pmatrix} \mathbf{V}(z_i) \begin{pmatrix} 1 \\ z_i \end{pmatrix}). \end{aligned} \tag{4.1}$$

Let us generalized the model (4.1) by considering all local factor effects that may disturb the variance and are ignored in our assumed simple regression model. This is achieved by changing σ^2 to σ_i^2 and assuming $\sigma_i^2 = \psi_i \phi_i^2 \tau^2$. Moreover, for simplicity, in the following, we assume matrix $\mathbf{V}$ is the identity matrix of dimension 2. Adopting the notations of model (2.8) for model (4.1), we have $h(z_i) = 1 + z_i^2$ and the plug-in estimate $\hat{\delta}^2 = 0.792$. So, the corresponding hierarchical

heteroscedastic model is given by

$$\begin{aligned} y_i &\sim N(\theta_i, \frac{\hat{\delta}^2}{z_i^2}) \\ \theta_i|\phi_i,\tau &\sim DE(\mu, \tau\phi_i(1+z_i^2)^{1/2}) \\ \mu &\sim (\bar{y}, 100) \\ \phi_i &\sim Beta(a,(n-1)a) \\ \tau &\sim G(na, 1/2) \\ a &\sim G(5,5). \end{aligned}$$

We call this weighted regression model as M_1. To evaluate the performance of our methodology we also consider two other competing models M_2(unweighed regression model) and M_3 (ignoring the effect of explanatory variable z and using sample variance of y's as the estimate of δ^2). We applied the Gibbs sampler of Theorem 2, for $N = 10000$ draws for this data. The implementation results are given in Table 2. From this table, it is easy to see that model M_1 has a very small mean predictive

Table 2. The prediction errors and Bayesian estimate of a for various choices

Model	$\hat{\delta}^2$	a-estimate	MPE
M_1	0.792	1.300	6.10
M_2	34.45	1.094	13.52
M_3	259.968	0.978	216.44

errors,

$$MPE = \frac{1}{n}\sum(\hat{\theta}_i^p - y_i)^2,$$

which implies a great fit. In contrast, model M_2 and model M_3 (and even the weighted least square (WLS) analysis with MPE=29.87) give an unrealistic MPE, implying unsatisfactory fit. The corresponding shrinkage estimates of model M_1 are given in Table 1 which are acceptable from practical point of view. Since, the MPE and DIC criteria give the same results on these data, we only report the MPEs in Table 2.

4.2. *Multiple regression model*

Consider Systolic Blood Pressure data which is available at *http://college. cengage.com/mathematics/brase/understandable_statistics/7e/students/ data sets/ mlr/frames/frame.html*. Table 3 shows the data. It contains two explanatory variables z_1:Age in years and z_2: Weight in pounds. The response variable y is the systolic blood pressure of 11 patients. A multiple regression analysis shows $MSE = 4.84$.
Similar to the previous simple weighted regression, we consider the following multiple unweighted

Table 3. Systolic Blood Pressure data and corresponding shrinkage estimates

No.	z_1	z_2	y	$\hat{\theta}^P$
1	52	173	132	132.3
2	59	184	143	143.5
3	67	194	153	153.0
4	73	211	162	161.5
5	64	196	154	153.8
6	74	220	168	167.7
7	54	188	137	137.3
8	61	188	149	149.7
9	65	207	158	157.4
10	46	167	128	128.2
11	72	217	166	165.6

regression hierarchical model for these data.

$$\begin{aligned} y_i &\sim N(\theta_i, \delta^2), \\ \theta_i &\sim N(0, \sigma_i^2(1+z_{1i}^2+z_{2i}^2)). \end{aligned}$$

To match this model with model (2.8), consider $g(z_{1i}, z_{2i}) = \delta^2$ and $h(z_{1i}, z_{2i}) = 1+z_{1i}^2+z_{2i}^2$. Therefore, the corresponding hierarchical model is given as

$$\begin{aligned} y_i &\sim N(\theta_i, \hat{\delta}^2) \\ \theta_i|\phi_i, \tau &\sim DE(\mu, \tau\phi_i(1+z_{1i}^2+z_{2i}^2)^{1/2}) \\ \mu &\sim (\bar{y}, 100) \\ \phi_i &\sim Beta(a, (n-1)a) \\ \tau &\sim G(na, 1/2) \\ a &\sim G(5,5). \end{aligned} \tag{4.2}$$

To evaluate the performance of our variable selection approach, we consider four different models. Model M_1 ignores the effects of both explanatory variables and only uses the sample variance of y's as an estimate for δ^2 . Model M_2 incorporates the effect of regressor z_1 while model M_3 considers only the effect of z_2. Finally, model M_4 incorporates the effects of both explanatory variables z_1 and z_2 in the Bayesian analysis. We use the classical mean square error as an estimate of δ^2 in three last models. Table 3 shows the results for $N = 10000$ draws. Table 4 shows that ignoring the explanatory variables, model M_1, produces very unreasonable value for MPE. Model M_4 which employs both variables z_1 and z_2 makes no improvement over models M_2 and M_3. This is true because there is a big correlation 0.946 between z_1 and z_2. So, from practical point of view, we prefer model M_2 and only select the explanatory variable z_1 in our Bayesian model selection approach. The weighted least square (WLS) analysis gives MPE=29.87. Moreover, the corresponding shrinkage estimates of model M_2 are given in Table 3 which are acceptable from practical point of view.

An interesting feature of examples 1 and 2 is the computation of the posterior probability given in (2.10). Table 5 shows these probabilities. As one can see, heteroscedasticity for these data and especially for example 2 is more likely and lets us analyze the data based on model (1.2).

Table 4. The prediction errors and Bayesian estimate of a for various settings

Model	$\hat{\delta}^2$	a-estimate	MPE
M_1	185.69	0.970	148.9
M_2	4.84	0.593	0.169
M_3	4.84	0.421	0.263
M_4	4.84	0.418	0.261

Table 5. The posterior probability

η	0.1	0.01	0.005	0.001
Example 1	1.000	0.872	0.331	0.000
Example 2	0.998	0.194	0.011	0.000

5. Conclusion

In this paper, a clear Bayesian methodology for fitting of a completely heteroscedastic hierarchical linear model, using Dirichlet-Laplace priors, has been presented. This method provides a good fit for all sizes of data. Providing an efficient posterior computation algorithm is another property of this approach. Moreover, good performance of the methodology has clearly been shown for two real datasets.

Acknowledgements

I wish to thank Professor M.R. Meshkani, the editor and two anonymous referees whose comments greatly improved the article.

Appendix

For establishing the asymptotic theorem, the following three conditions are required,

C1) $\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} g(A_t) < \infty$.
C2) $\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \theta_t^2 < \infty$.
C3) For given $M > 0$, $\mid \mu \mid \leq M < \infty$.

Proof of Theorem 3.1
Consider $\sigma_i^2 = \psi_i\phi_i^2\tau^2$. It is easy to see

$$MGS(\sigma_1^2,\cdots,\sigma_n^2,\mu) - l_q(\theta,\theta^{\sigma_1^2,\cdots,\sigma_n^2,\mu}) = \frac{1}{\sum q_i}\sum\{\frac{y_i^2 - g(z_i) - \theta_i^2}{g(z_i)} - \frac{2\psi_i\phi_i^2\tau^2 h(z_i)}{g(z_i)+\psi_i\phi_i^2\tau^2 h(z_i)} \times (\frac{y_i^2 - g(z_i) - \theta_i y_i}{g(z_i)})\} - \frac{2\mu}{\sum q_i}\sum q_i \frac{g(z_i)(y_i-\theta_i)}{\psi_i\phi_i^2\tau^2 h(z_i)+g(z_i)}.$$

For given ψ_i and ϕ_i, we have

$$MGS(\sigma_1^2,\cdots,\sigma_n^2,\mu)-l_q(\theta,\theta^{\sigma_1^2,\cdots,\sigma_n^2,\mu}) = \frac{1}{\sum q_i}\sum\{\frac{y_i^2-g(z_i)-\theta_i^2}{g(z_i)} - \frac{2\tau^2}{\tau^2+u(z_i,\psi_i,\phi_i)}\times(\frac{y_i^2-g(z_i)-\theta_i y_i}{g(z_i)})\} - \frac{2\mu}{\sum q_i}\sum q_i\frac{u(z_i,\psi_i,\phi_i)(y_i-\theta_i)}{\tau^2+u(z_i,\psi_i,\phi_i)},$$

where $u(z_i,\psi_i,\phi_i)=\frac{g(z_i)}{\psi_i\phi_i^2h(z_i)}$. From Theorem (3.2) in Ghoreishi and Meshkani (2014), under Conditions (C1)-(C3), it easy to see

$$\sup_{0\le\tau^2\le\infty,\ |\mu|\le M} |\ MGS(\sigma_1^2,\cdots,\sigma_n^2,\mu)-l_q(\theta,\theta^{\sigma_1^2,\cdots,\sigma_n^2,\mu})\ |\to 0\ in\ L^2\ as\ n\to\infty. \qquad (A.1)$$

Since (A.1) is valid for all ψ_i and ϕ_i, the proof of Theorem 1 is straightforward. □

Proof of Theorem 2

The proof is straightforward.

References

[1] J. Berger, and W.E. Strawderman, Choice of Hierarchical Priors:Admissibility in Estimation of Normal Means, *Annals of Statistics* **24** (1996) 931–951.

[2] A. Bhattacharya, D. Pati, N.S. Pillai, D.B. Dunson, Dirichlet-Laplace priors for optimal shrinkage, (2014) arXiv:1401.5398v1.

[3] K.W. Broman and T.P. Speed, A model selection approach for the identification of quantitative trait loci in experimental crosses, *J. Roy. Stat. Soc. B* **64** (2002) 641–656.

[4] L.D. Brown, and E. Greenshtein, Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High-Dimensional Vector of Means, *Annals of Statistics* **37** (2009) 1685–1704.

[5] P. Dellaportas, J.J. Forster, and I. Ntzoufras, Bayesian Variable Selection Using the Gibbs Sampler, Generalized Linear Models: A Bayesian Perspective (D. K. Dey, S. Ghosh, and B. Mallick, eds.). New York: Marcel Dekker, (2000) 271–286.

[6] W. Edwards, H. Lindman, and L.J. Savage, Bayesian statistical inference for psychological research, *Psychological Review* **70** (1963) 193–242.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression, *Annals of statistics* **32** (2004) 407–499.

[8] S.K. Ghoreishi and M.R. Meshkani, On SURE estimates in hierarchical models assuming heteroscedasticity for both levels of a two-level normal hierarchical model, *J. of Multivariate Analysis* **132** (2014) 129–137.

[9] S.K. Ghoreishi and A. Mostafavinia, Shrinkage estimates for multi-level heteroscedastic hierarchical normal linear models, *J. of Statist. Theory and Appl.* **14** (2015) 204–213.

[10] C. Robert and G. Casella, Monte Carlo Statistical Methods, (Springer, New York, 2004).

[11] C.M. Stein, Confidence Sets for the Mean of a Multivariate Normal Distribution (with discussion), *J. Roy. Statist. Soc. Ser. B,* **24** (1962) 265–296.