A Conservative Prior for Bayesian Hierarchical Models in Biostatistics

by

Md. Shahadut Hossain

M.Sc., University of Dhaka, 1993

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

We accept this thesis as conforming
to the required standard

The University of British Columbia

August 2003

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date 22.08.2003

Abstract

Hierarchical models are suitable and very natural to model many real life phenomena, where data arise in nested fashion. The use of Bayesian approach to hierarchical models has numerous advantages over the classical approach. Estimating a phenomenon with hierarchical model can be viewed as a smoothing problem, and hence while summarizing such a phenomenon via hierarchical model we do not want to undersmooth the phenomenon. That is, in most of the practical applications undersmoothing is more serious type of error than oversmoothing. So, we need an estimation approach which can guard against undersmoothing. If we can control the undersmoothing reasonably, we may get a better calibrated summary of the phenomenon we estimate.

In this study, we have incorporated the aspect of smoothing in estimating the parameters of Bayesian hierarchical models. In doing so we have proposed a conservative prior for the variance component to achieve the adequate degree of smoothness while estimating the phenomenon under study. We have conducted simulation studies to decide about the appropriate values to be used for the hyperparameter while using the conservative prior to ensure the adequate degree of smoothness. We have investigated the performance of the proposed conservative prior in guarding against undersmoothing in simple normal–normal hierarchical models (random effects models for normal response) and in non–parametric regression curve estimation problem via simulation studies. We have also investigated the performance of the proposed prior compared to those of the uniform shrinkage prior and the Jeffreys' prior, with respect to both guarding against undersmoothing and the MSE of the estimated model parameters, through simulation studies.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to convey my gratitude and thanks to my Supervisor Dr. Paul Gustafson for his guidance and help in completing my thesis work. Needless to mention that without his initiative, guidance and co-operation it was not possible for me to complete this piece of work.

Secondly, I am grateful to Dr. Ying Macnab, Assistant Professor, Department of Health Care and Epidemiology, UBC, for being the second reader of my thesis.

Also, I would like to thank all the faculty members in the department of statistics for their high commitment to quality of teaching and mentoring.

Lastly, I have been benefited from the department of statistics, UBC, in many ways. I would like to take the opportunity to thank all students for making my last two years of stay enjoyable and office staff for their co-operation.

MD. SHAHADUT HOSSAIN

The University of British Columbia
August 2003

To my Parents and Wife.

Chapter 1

Hierarchical Models

1.1 Introduction

In many practical field of studies data may arise in nested or hierarchical fashion. Some common scenarios are:

(i) in the field of health, e.g. patients nested within a sample of hospitals may themselves be nested within a geographic region etc.

(ii) in meta–analysis, information from a number of studies on the same phenomenon are combined to produce more accurate inferences and predictions than those available from any single study. In such case of combining information, subjects are nested within studies

(iii) in longitudinal studies, repeated observations on each individual study subject are nested within the individual

In all the above cases if people model data as independent and identically distributed for the sake of simplicity, e.g., pretend that the subjects in the sampling experiment are drawn homogeneously from a single population, then the analysis applied to summarize the data can not be able to depict the real picture that prevails in the data set itself, because in this case the researcher will fail to capture the similarity between groups.

Furthermore, in any hierarchical setup observations are obtained in clusters and the responses from the same cluster can not be assumed independent. So, to incorporate this within cluster

correlation and to reflect the dependence of cluster specific random effects we must consider some hierarchical modeling techniques. In hierarchical models the observable outcomes are modeled conditionally on the group level random effects and then the group level random effects are given a probabilistic specification in terms of the further parameters known as hyperparameters.

1.2 Why Hierarchical Models?

For nested data hierarchical models are good choice for the following reasons:

(i) hierarchical models permit the direct framing of the theories about the effect of structural change at each of the different levels of the hierarchy

(ii) they provide the accurate adjustments to the uncertainty assessment based on the simple random sampling when the data are gathered in hierarchical fashion in the presence of strong intra-cluster correlations

(iii) use of non–hierarchical models is inappropriate for the hierarchical data because with a few parameters they usually can not fit the data accurately. For example, if we consider the observed data y_{ij} $(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n_i)$ from all clusters are homogeneously arise from a common distribution with overall population effect θ, then such a model may fail to account for cluster–to–cluster variability , which may be an important source of variability in the data. Again, with many parameters use of non–hierarchical models tend to overfit the data in the sense of producing models that fit the existing data well but lead to inferior predictions for new data . Hierarchical models can handle both of the problems since they have enough parameters to fit the data well and they can avoid the problem of overfitting by borrowing the strength across clusters using a population distribution to incorporate the similarity among the cluster–specific effects.

Again, in causal analysis it may be necessary to introduce covariates at multiple levels of hierarchy. In such case the assumption of exchangeability of units or subjects at the lowest level breaks down even after conditioning on covariate information because of information sharing among the units at each level. So, we need to introduce covariates relevant to each of the higher level units. But introducing covariate at each level of hierarchy dramatically increases the number of parameters in the model and a sensible estimation is only possible through

further modeling in the form of population distribution, i.e., we need to consider a population distribution for the effects at each level of hierarchy and thus gives rise to hierarchical models.

In a nutshell, hierarchical models can improve the cluster or area specific estimates by combining the data across the units at each level of hierarchy. And in doing so they consider the probabilistic mechanism that gives rise to the data at each level of hierarchy. Hierarchical models offer an explicit framework which can express similarity judgement, combine information across units, and thus produce accurate and well–calibrated predictions of observable outcomes.

1.3 Bayesian Approach to Hierarchical Models

Heuristically, it can be said that hierarchical models force us to be Bayesian because to set up a hierarchy we need to consider a population distribution (prior distribution) for the parameters at each level to incorporate the uncertainty about the parameters at that level. There are both methodological and computational arguments in favor of being Bayesian while dealing with the hierarchical models. Some of them are listed below.

(i) Bayesian methods have the flexibility to incorporate the multiple levels of randomness that prevail in hierarchical models, and the resultant ability to combine information from different sources, while incorporating all reasonable sources of uncertainty in inferential summaries. They naturally lead to smoothed estimates in complicated data structures and consequently have the ability to obtain better real–world answers.

(ii) The psychological reason for adopting Bayesian methods is that in almost all practical cases the users of statisticians' work usually interpret interval estimates provided by statisticians as Bayesian intervals, that is, as probability statements about the likely values of unknown quantities conditional on the evidence in the data. Such direct probability statements require prior probability specifications for unknown quantities, and thus the kinds of answers the clients assume are Bayesian.

Thus from structural point of view the strength of Bayesian hierarchical approach lies in (a) its ability to combine information from multiple sources, (b) its more encompassing accounting of uncertainty about unknowns in a statistical problem and (c) it is the essential tool for achieving partial pooling of estimates and compromising in a scientific way between alternative sources of information. Actually, such partial pooling is more sophisticated and scientific than simple

pooling of individual estimates in the sense that such pooling considers all sources of uncertainty rather than considering only one source of random error within each unit.

Besides theoretical view points there are computational view points also that lead us to be Bayesian. The computational view points are summarized below.

(i) In the case of hierarchical modeling the likelihood approach becomes very difficult to implement, even with the normal response, when the level of hierarchy gets increased. In case of non–normal response the likelihood approach becomes intractable.

(ii) In causal analysis when the response from the same cluster are correlated random or mixed effects models are used. In such cases if the response are not normal the likelihood approach becomes intractable because it requires evaluation of high dimensional integral to calculate likelihood function. Though there are some approximate methods such as, the penalized quasi–likelihood (PQL) and the marginal quasi–likelihood (MQL) (Breslow and Clayton, 1993) based on Laplace's method of integral evaluation they have serious limitation in estimating the variance structure. Both PQL and MQL methods can give rise to a non-positive definite variance structure, which is absurd. The possibility of likelihood approach's producing negative estimate for the variance component can be well understood by considering the case of proto–type normal–normal hierarchical model. Such a model can be written as

Stage 1: $y_i \mid \theta_i \sim N(\theta_i, \omega)$, where ω is known

Stage 2: $\theta_i \sim N(0, \lambda)$

The two stages of the proto–type model can be collapsed to $y_i \mid \lambda \sim N(0, \omega + \lambda)$. The simple likelihood estimate for λ is obtained as $\hat{\lambda} = S^2 - \omega$, where S^2 is the pooled sample variance. Clearly, for $\omega > S^2$ the likelihood estimate of variance component λ will be negative, i.e., $pr(\hat{\lambda} < 0) > 0$. The Bayesian formulation enjoys an advantage here because of the information on the variance components contributed by the prior specification. The Bayesian method is simple to implement due to Markov Chain Monte Carlo (MCMC) methods.

(iii) Even in linear mixed effects models (i.e., when the response are normal) closed–form solutions for the maximum likelihood (ML) and Restricted Maximum Likelihood (REML) estimators of the parameters are often unavailable. In such cases numerical optimization methods like Newton–Raphson and EM algorithm are used to calculate estimates of the model parameters. These algorithms are not guaranteed to locate the global maximum

of the likelihood function from an arbitrary starting point. But a hierarchical Bayesian version of the mixed model does not have such problem of convergence to different points because it leads to simulation from the appropriate posterior distribution. This gives a distributional estimate of the model parameters and thus can provide more complete estimates

1.4 Choice of Prior Specification for Variance Component in Hierarchical Models

In this study our concern is to make a good choice of prior specification for variance components in Bayesian hierarchical models. The choice of prior distribution for variance components is a critical task because often adequate subjective prior information is lacking. For situation where little prior information is available, the usual choice is to use a non–informative type prior. But the problem of using non–informative type priors is that most of them are improper, and so may lead to an improper joint posterior distribution, when applied to a variance component. Impropriety of joint posterior distribution implies that there does not exist a joint density to which an MCMC chain converges, i.e., MCMC algorithm leads to inferences about a non–existent posterior distribution (Hobert and Casella, 1996). So, before using an improper prior we must ensure that the resulting joint posterior distribution is a proper one. Many authors worked on the issue of choosing non–informative prior for the variance components in hierarchical models. The leading ones are Jeffreys (1961), Box & Tiao (1973), Berger & Deely (1988), Berger & Bernando (1992), Daniels (1999) etc.

Daniels (1999) introduced a non–informative prior, known as the uniform shrinkage prior, which itself is proper, and hence leads to proper posterior distribution. It also possesses many desirable frequentist properties. However, none of the authors who worked on this issue previously did consider any explicit criteria to choose the priors for the variance component. In this study we have considered the concept of smoothing in choosing the prior for the variance components. It is true that any estimation procedure is imperfect and we either oversmooth or undersmooth the phenomenon we estimate. But in many application undersmoothing is a worse error than oversmoothing, as undersmoothing involves postulating structure that is not really there. So, in this study we have attempted to choose priors for the variance components in such a way that we can guard against undersmoothing. For example, consider the following normal–normal hierarchical model.

Stage 1: $y \mid \theta, \lambda \sim N(A\theta, v_1)$, where A is a known $n \times p$ matrix, θ is a unknown $p \times 1$ vector, v_1 is an $n \times n$ matrix (may or may not be known).

Stage 2: $\theta \mid \lambda \sim N(0, \lambda v_2)$, where λ is unknown scalar, v_2 is a known $p \times p$ matrix.

Stage 3: $\lambda \sim \pi(\lambda)$, where $\pi(\lambda)$ is a prior distribution for λ, the variance component.

Such models have numerous applications. Some of the scenarios are listed below.

Scenario A: Random effects models. For instance, data are collected on patients at p hospitals. It may be reasonable to think that the patient outcomes are similar but not identical across hospitals. Thus y consists of patient outcomes, θ_i can be taken to be the ith hospital effect.

Scenario B: Spatial models. For instance, data might describe observed disease prevalence in a geographic regions. It may be reasonable to think that the underlying prevalence in adjacent regions is similar but not identical. Thus θ_i is the ith region effect and v_2 is chosen so that component of θ which are geographically closer have higher positive correlation.

Scenario C: Curve-fitting. Say that $Y = f(X)$+noise, and we wish to estimate the function f. The smoothing spline approaches to this problem can be viewed as a hierarchical model with θ being the values of f at some fixed knot values. By appropriate choice of v_2 we end up penalizing functions f which are more wiggly.

All the three scenarios can be thought as smoothing problem. In each case we estimate (rather than fix) λ so that the data decide on "how much smoothness" is appropriate. That is, how similar should the hospitals or regions be in scenarios A and B, and literally, how smooth should the function f be in scenario C. And the choice of prior at Stage 3 of the hierarchical model has some impact on this data–driven smoothing procedure.

Again, in smoothing problem it is assumed that the underlying phenomenon is smooth, and hence we want an estimate of the phenomenon which is also smooth. For example, consider a hospital model about the effectiveness of a cardiac treatment, with patients in the hospital j having the survival probability θ_j, it might be reasonable to expect that estimates of θ_js', which represent a sample of hospitals, should be related to each other. A natural way to incorporate this similarity of θ_js' is to use a prior distribution in which θ_js' are viewed as a sample from a

common population distribution. In such case the observed data, y_{ij}, with units indexed by i within groups indexed by j, can be used to estimate aspects of the population distribution of the θ_js' even though the values of θ_js' are not observed. It is natural to model such a phenomenon hierarchically, with observable outcomes modeled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of further parameters known as hyperparameters. In such kind of hospital model it is important to have estimates of θ_js' which are not more variable than θ_js' themselves. Because, if it is known that the survival rates among patients suffering from cardiac ailment are higher in some hospitals than those in the other hospitals people may rush to the hospitals with higher survival rates, which may cause serious administrative problems. In such a case the idea is that we do not want to declare that some hospitals are better with respect to cardiac ailment care unless we are pretty much sure, i.e., we want estimates of θ_js' to be smooth. So, in choosing a prior for λ in Stage 3 we have emphasized on an estimate of λ which does not undersmooth the phenomenon understudy. Undersmoothing is defined to be $\hat{\lambda} > \lambda$ and oversmoothing is defined to be $\hat{\lambda} < \lambda$, where $\hat{\lambda}$ is the estimate of λ. Since undersmoothing is considered to be more serious type of error, in this study we have focused on choosing the Stage 3 prior in order to guard against undersmoothing. In particular, we consider choosing the prior in order to make the (frequentist) probability that $\hat{\lambda} > \lambda$, i.e., $pr(\hat{\lambda} > \lambda)$, small.

In search for such a prior we can start with the following proto–type normal–normal hierarchical model:

Stage 1: $y_i \mid \theta_i \sim N(\theta_i, \omega)$, where y_i $(i = 1, 2, \ldots, n)$ is the observed summary outcome in the ith cluster; θ_i is the true effect for the ith cluster and ω is the known dispersion parameter.

Stage 2: $\theta_i \mid \lambda \sim N(0, \lambda)$

The above hierarchical model can be collapsed as:

$$\begin{aligned} E(y_i|\lambda) &= E[E(y_i|\theta_i)|\lambda] \\ &= E(\theta_i|\lambda) \\ &= 0 \\ \text{and } var(y_i|\lambda) &= E[var(y_i|\theta_i)|\lambda] + var[E(y_i|\theta_i)|\lambda] \\ &= E(\omega|\lambda) + var(\theta_i|\lambda) \\ &= \omega + \lambda \end{aligned}$$

Therefore, $y_i|\lambda \sim N(0, \omega+\lambda)$. The likelihood function can then be expressed as

$$L(\lambda) \propto \left(\frac{1}{\omega+\lambda}\right)^{\frac{n}{2}} e^{-\frac{\frac{1}{2}\sum_{i=1}^{n} y_i^2}{\omega+\lambda}}$$

A prior density of the form $\pi(\lambda) \propto (\frac{1}{\omega+\lambda})^{(a+1)} e^{-\frac{b}{\omega+\lambda}}$ would be conjugate, where $\omega+\lambda \sim IG(a,b)$ truncated to be $\omega+\lambda \geq \omega$, i.e., a priori λ is distributed as $V_1 - w|V_1 \geq w$, where $V_1 \sim IG(a,b)$. Now, we need to make reasonable choices of a and b to guard against undersmoothing. The posterior distribution of λ, given the data, is obtained as

$$L(\lambda \mid Y) \propto (\frac{1}{\omega+\lambda})^{\frac{n}{2}+a+1} e^{-\frac{\frac{1}{2}\sum_{i=1}^{n} y_i^2+b}{\omega+\lambda}} \sim IG\left(\frac{n}{2}+a, \frac{1}{2}\sum_{i=1}^{n} y_i^2 + b\right)$$

That is, a posteriori λ is distributed as $V_2 - w|V_2 \geq w$, where $V_2 \sim IG\left(a+\frac{n}{2}, b+\frac{1}{2}\sum y_i^2\right)$.

Now, take $\hat{\lambda}$ =posterior mode as the estimator of λ. So, $\hat{\lambda}$ =max$\left(\frac{\sum_{i=1}^{n} y_i^2+2b}{n+2a+2} - \omega, 0\right)$. The probability of undersmoothing can then be calculated as

$$\begin{aligned} pr(\hat{\lambda} > \lambda) &= pr\left(\frac{\sum_{i=1}^{n} y_i^2 + 2b}{n+2a+2} - w > \lambda\right) = pr\left(\frac{(\omega+\lambda)\sum_{i=1}^{n} \frac{y_i^2}{\omega+\lambda} + 2b}{n+2a+2} > \omega+\lambda\right) \\ &= pr\left(\frac{(\omega+\lambda)v + 2b}{n+2a+2} > \omega+\lambda\right) = pr\left(\frac{v-n}{\sqrt{2n}} > \frac{\sqrt{2}(a+1-\frac{b}{\omega+\lambda})}{\sqrt{n}}\right) \\ &= pr\left(z > \frac{\sqrt{2}(a+1-\frac{b}{\omega+\lambda})}{\sqrt{n}}\right) \end{aligned} \tag{1.1}$$

In Equation (1.1) , $v = \frac{\sum_{i=1}^{n} y_i^2}{\omega+\lambda} \sim \chi_n^2$ and $z = \frac{v-n}{\sqrt{2n}} \sim N(0,1)$ as $n \to \infty$. Again, from Equation (1.1) we observe that the chance of undersmoothing increases as λ decreases. Hence, the worst possible case of undersmoothing takes place when $\lambda = 0$, and so we can worry just about the case when $\lambda = 0$, i.e., if we can ensure adequate guard against undersmoothing in case of $\lambda = 0$ then we can automatically ensure the adequate guard against undersmoothing in all the other cases. For $\lambda = 0$, Equation(1.1) can be written as

$$pr(\hat{\lambda} > \lambda) = pr\left(z > \frac{\sqrt{2}(a+1-\frac{b}{\omega})}{\sqrt{n}}\right) \tag{1.2}$$

For fixed n we can choose the values of a and b to make the probabilities given by equation (1.2) small, but there still has the questions about what happens as $n \to \infty$. Theoretically, as

$n \to \infty$, $\frac{\sqrt{2}(a+1-\frac{b}{\omega})}{\sqrt{n}} \to 0$, which implies that the probability of undersmoothing approaches to 0.5. Figure 1.1 displays the probability plots against n corresponding to the Equation (1.2). Form these plots it is observed that probability of undersmoothing approaches to 0.5 as $n \to \infty$. Again, form the Figure it is also observed that bigger values of a help in keeping the probability of undersmoothing small.

Figure 1.1: Probability of undersmoothing as a function of n

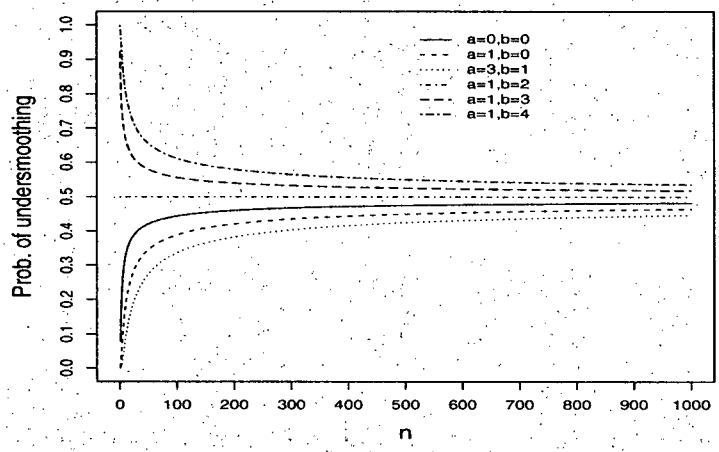
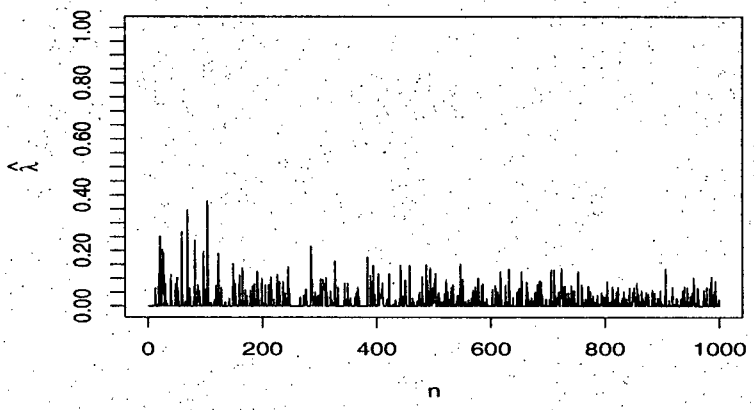


Figure 1.2: Plot of $\hat{\lambda}$ against n



Now, since probability of undersmoothing approaches to 0.5 as the sample size gets larger, there may be frequent upward jumps in the values of $\hat{\lambda}$ and we need to check how big these jumps are. To check the magnitude of the jumps in the values of $\hat{\lambda}$ we have simulated data and

calculated $\hat{\lambda}$ for different values of n and plotted them against n for a true λ value of zero in Figure 1.2. For this plot the values of the hyperparameters a and b have been taken to be 3 and 1, respectively. From Figure (1.2) it is clear that though there are frequent jumps in the values of $\hat{\lambda}$ the jumps get smaller as n gets larger. So, we need not to worry about the jumps in $\hat{\lambda}$ when n is large.

Again, from Equation (1.1) we observe that the probability of undersmoothing is an increasing function of b. So, setting $b = 0$ can ensure maximum possible guard against undersmoothing. For $b = 0$ the probability of undersmoothing is obtained as

$$pr(\hat{\lambda} > \lambda) = pr\left(z > \frac{\sqrt{2}(a+1)}{\sqrt{n}}\right) \tag{1.3}$$

From Equation (1.3) we observe that for fixed n the probability of undersmoothing decreases as a increases. Hence for fixed n we look for a bigger values of a to guard against undersmoothing. Another advantage of setting b to zero is that the probability of undersmoothing does not depend on λ and ω, and we can express the maximum value of probability of undersmoothing as a function of a only.

Sometimes it is desired to express the probability of undersmoothing as the probability that $\hat{\lambda}$ exceeds λ plus a small increment obtained as a function of true λ, i.e.,

$$\begin{aligned} pr(\text{undersmoothing}) &= pr(\hat{\lambda} > \lambda + \epsilon(\lambda + \omega)) \\ &= pr\left(\frac{\sum_{i=1}^{n} y_i^2}{n + 2a + 2} > (\lambda + \omega)(1 + \epsilon)\right) \\ &= pr\left(z > \frac{2(a+1)(1+\epsilon) + n\epsilon}{\sqrt{2n}}\right) \end{aligned} \tag{1.4}$$

The probability of undersmoothing given by Equation (1.4) will be maximum if

$$f(n) = \frac{2(a+1)(1+\epsilon) + n\epsilon}{\sqrt{(2n)}} \tag{1.5}$$

is minimum. The minimum value of $f(n)$ has been found to be

$$f(n)_{min} = 2\sqrt{\epsilon(a+1)(1+\epsilon)} \tag{1.6}$$

Thus

$$pr(\hat{\lambda} > \lambda + \epsilon(\lambda + \omega)) = pr(z > 2\sqrt{\epsilon(a+1)(1+\epsilon)}) \tag{1.7}$$

From Equation (1.7) we observe that the maximum value for the probability of undersmoothing will be smaller for bigger values of a. Table 1.1,constructed based on the maximum probability of undersmoothing given by the Equation (1.7), gives different values of a required to keep the maximum probability of undersmoothing to a certain level α (say), e.g. 5%. From Table 1.1

Table 1.1: Values of the hyperparameter a required to keep the maximum probability of undersmoothing to a certain level for different values of ϵ

α	ϵ		
	0.05	0.10	0.20
0.05	13.89	7.15	3.83
0.10	8.82	4.73	2.71
0.20	4.37	2.61	1.74

we see that quite big values of a is required if we want keep the maximum probability of undersmoothing at 5% level. For this level we need $a = 4$ even if we allow $\hat{\lambda}$ to exceed λ by 20% of the total variability, $\omega + \lambda$. From the Table it is also observed that even if we allow 20% of the $\hat{\lambda}$ exceed the true λ itself by a magnitude of 20% of the data variability we still need a value of almost 2 for a ($a = 1.74$). So, it can be reasonably argued that the uniform shrinkage prior, which considers $a = 1$, can not control undersmoothing adequately.

So, finally by considering all the advantages and mathematical ease it can be concluded that the desired conservative prior for Bayesian hierarchical model can be obtained from truncated inverted gamma prior $\pi(\lambda) \propto (\frac{1}{\omega+\lambda})^{(a+1)} e^{-\frac{b}{\omega+\lambda}}$ by choosing $b = 0$ and a bigger value for a. We term the proposed prior as "conservative prior" because it is conservative against undersmoothing. Now, taking $b = 0$ in the inverted gamma prior leads to a class of priors and is termed as "parametric power" priors (Hobert & Casella, 1996). The advantages of using such priors are that they are obtained from the inverted–gamma prior and for $a > 0$ they are proper and hence lead to a proper joint posterior distribution of unknown parameters, which is essential for adopting MCMC to simulate the values of unknown parameters for valid posterior inferences (Hobert and Casella, 1996). Such priors give the flexibility to choose the value of a to guard against undersmoothing. Furthermore, these priors comprise a class of non–conjugate priors within which some of the most frequently used non–conjugate and relatively non–informative priors belong, e.g., Jeffreys prior and the uniform shrinkage prior. For $a = 0$ this class gives rise to Jeffreys prior and for $a = 1$ it gives rise to the uniform shrinkage prior.

1.5 Application of the Proposed Prior: Simulation Studies

To verify the performance of the proposed class of priors we have conducted simulation studies for the simple normal–normal hierarchical model defined below.

Stage 1: $y_{ij} \mid \theta_i, \omega \sim N(\theta_i, \omega)$; $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$

Stage 2: $\theta_i \mid \mu, \lambda \sim N(\mu, \lambda)$

In stage 1, ω is sometimes considered as nuisance parameter and assumed to be known. But in most of the practical situations ω is unknown and to be estimated form the data. For estimating ω under the Bayesian set up we need to assume a prior for ω in stage 3. So, the full Bayesian version of the above hierarchical model can be written as

Stage 1: $y_{ij} \mid \theta_i, \omega \sim N(\theta_i, \omega)$

Stage 2: $\theta_i \sim N(\mu, \lambda)$

Stage 3: $\pi(\mu) \propto 1$
$\pi(\lambda \mid \omega) \propto (\frac{1}{\lambda+\omega})^{a+1}$
$\pi(\omega) \propto (\frac{1}{\omega})^{c+1} e^{-\frac{d}{\omega}}$

where, $\pi(\omega)$, $\pi(\mu)$ and $\pi(\lambda \mid \omega)$ are the priors for ω, μ and λ, respectively. In stage 3 the prior for ω has been taken to be unit information inverted–gamma prior by setting $c = \frac{1}{2}$ and $d = \frac{1}{2}$ for our simulation studies. The reason of choosing unit information prior is that we are assuming we have little prior information, worth of one data point, about the distribution of ω and allowing ω to be estimated mostly by the data. The flat prior density for μ in stage 3 is reasonable for hierarchical model, because, the combined data from all n experiments (clusters) are generally highly informative about μ and hence we can afford to be vague about its prior distribution. Now, the joint posterior distribution of θs', μ, ω and λ, given the data, can be

obtained as

$$\pi(\theta,\mu,\omega,\lambda \mid Y) \propto \left[\prod_i \prod_j f(y_{ij} \mid \theta_i,\omega)\right] \times \left[\prod_i f(\theta_i \mid \mu,\lambda)\right] \times \pi(\mu) \times \pi(\lambda \mid \omega) \times \pi(\omega) \times \frac{1}{c(\omega)}$$

$$\propto \prod_{i=1}^{n}\prod_{j=1}^{k}\left(\frac{1}{\omega}\right)^{\frac{1}{2}} e^{-\frac{\frac{1}{2}(y_{ij}-\theta_i)^2}{\omega}} \prod_{i=1}^{n}\left(\frac{1}{\lambda}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\frac{(\theta_i-\mu)^2}{\lambda}} \left(\frac{1}{\omega+\lambda}\right)^{a+1}$$

$$\left(\frac{1}{\omega}\right)^{c+1} e^{-\frac{d}{\omega}} \frac{1}{c(\omega)} \tag{1.8}$$

where, $c(\omega)$ is the normalizing constant. This normalizing constant is required to make the joint prior $\pi(\lambda,\omega)$ a proper probability distribution, which in turn make the joint posterior distribution given by the Equation (1.8) proper posterior distribution. Propriety of joint posterior for the model parameters is essential for valid posterior inferences. The normalizing constant for this problem is obtained as

$$c(\omega) = \int_0^\infty \left(\frac{1}{\omega+\lambda}\right)^{a+1} d\lambda = \frac{1}{a\omega^a}$$

Now, for posterior inferences about the model parameters we need to simulate draws for them from the posterior distribution. To draw simulations for each of the unknowns through MCMC method we need to compute the conditional posterior density for each of the unknowns. The conditional posterior for θ is obtained as

$$\pi(\theta \mid \mu,\omega,\lambda,y) \propto \prod_i^n e^{-\frac{1}{2}\left[\sum_j \frac{(y_{ij}-\theta_i)^2}{\omega} + \frac{(\theta_i-\mu)^2}{\lambda}\right]}$$

Since $\theta_1,\ldots,\theta_n$ are assumed exchangeable we can write $\theta_i \mid \mu,\omega,\lambda,y \sim N(\hat{\theta}_i,\tau^2)$, where $\hat{\theta}_i = \frac{\frac{1}{\lambda}\mu+\frac{k}{\omega}\bar{y}_i}{\frac{1}{\lambda}+\frac{k}{\omega}}$, $\tau^2 = \frac{1}{\frac{1}{\lambda}+\frac{k}{\omega}}$.

The conditional posterior density for μ is obtained as

$$\pi(\mu \mid \theta,\omega,\lambda,y) \propto \prod_{i=1}^{n} e^{-\frac{1}{2\lambda}(\theta_i-\mu)^2}$$

$$= e^{-\frac{n}{2\lambda}(\mu-\bar{\theta})^2}$$

Therefore, $\mu \mid \theta,\omega,\lambda,y \sim N(\bar{\theta},\frac{\lambda}{n})$.

The conditional posterior density for ω is obtained as

$$\pi(\omega \mid \theta,\mu,\lambda,y) \propto \left(\frac{1}{\omega}\right)^{\frac{nk}{2}+c+1} e^{-\frac{d+\frac{1}{2}\sum\sum(y_{ij}-\theta_i)^2}{\omega}} \left(\frac{1}{\omega+\lambda}\right)^{a+1} \frac{1}{c(\omega)}$$

Finally, the conditional posterior density for λ is obtained as

$$\pi(\lambda \mid \theta, \omega, \mu, y) \propto \left(\frac{1}{\lambda}\right)^{\frac{n}{2}} e^{-\frac{\frac{1}{2}\sum(\theta_i-\mu)^2}{\lambda}} \left(\frac{1}{\omega+\lambda}\right)^{a+1}$$

Since the conditional posteriors for θ_is' and μ have known parametric form we can use Gibbs sampler to draw simulations for them. But the conditional posteriors for ω and λ do not have known distributional form, and so we need to use random walk Metropolis–Hastings algorithm to draw simulations for them.

It has been already mentioned that the goal of this study is to choose the appropriate value of a in the prior for λ which can sufficiently guard against undersmoothing. Hence, to choose the appropriate value(s) of a and to judge the performance of the proposed prior in estimating λ, the variance component, we have performed simulation studies. For the simulation studies we have proceeded in the following way:

(i) fixed some values for each of the unknown parameters μ, ω and λ as their true values

(ii) for each combination of true values of μ and λ we have generated i.i.d. $\theta_1, \theta_2, \ldots, \theta_n$ from $N(\mu, \lambda)$

(iii) for each of the θ_is' an i.i.d sample $y_{i1}, y_{i2}, \ldots, y_{ik}$ has been generated from $N(\theta_i, \omega)$

(iv) given the generated data set, we have applied MCMC techniques to estimate the unknown parameters for different values of a and tried to identify which value of a is reasonable to sufficiently guard against undersmoothing.

For our simulation studies we have considered three different values of λ, the variance component. The chosen λ values are 0, 1 and 5. For all cases the chosen values of ω is 1 and the value of μ has been taken to be 5. The reason for choosing different values of λ for a fixed value of ω in our simulation study is to see how the proposed class of priors perform in estimating the unknown parameters in different situations when the ratios of the between and within cluster variances are smaller than, equal to and larger than unity and thus make a more general choice for a. The fixed choice of $\omega = 1$ is reasonable in the sense that we can rescale the ω values to 1, which will not affect the other things of our study. For each combination of μ, ω and λ values we have generated ten θ values $\theta_1, \theta_2, \ldots, \theta_{10}$ and from each of the generated θ values we have generated a sample of data values. For our simulations studies we can consider both the small and large sample sizes. But here we have considered only the small sample size case especially, to demonstrate the ability of hierarchical model to give accurate estimates by borrowing

strength across clusters. To identify a reasonable choice for the hyperparameter, a, we have generated 100 data sets and estimated λ by using each of the data sets for different choices of a. Finally, we have observed the number of times, out of 100, that $\hat{\lambda}$ exceeds the true λ value.

1.5.1 Monitoring the Convergence and Mixing the MCMC Run for ω and λ

While using random–walk Metropolis–Hastings algorithm to generate posterior simulations for both ω and λ we need to check the convergence and mixing of MCMC run for them before making any valid inferences about them. For better mixing and convergence we need to choose appropriate jump size for the MCMC algorithm while updating the parameter estimates. In Metropolis–Hastings algorithm there are two problems– (i) the slow movement of the chain toward the target distribution and (ii) slow mixing of the MCMC chain. An MCMC chain may move first (high acceptance rate) but may show slow mixing, i.e, the MCMC chain may move around a specific region of the target distribution for many iterations; on the other hand, it may exhibit good mixing but slow convergence. But in practice, it is always desirable to have a chain which mixes and converges well at the same time to ensure the accurate inference about the target posterior distribution. To have such a chain we need to adjust the jump size of the chain through monitoring the output. There is no hard and fast rule to determine the appropriate jump size. Usually, trial and error method is adopted. A considerable volume of research work has been carried out and suggestions have been made to monitor the mixing and convergence of an MCMC chain. A nice reference in this regard is Gilks, at. el. (1996). Research findings suggest that for better mixing and convergence a desirable acceptance rate is around 50%. So, while implementing an MCMC simulation it is necessary to plot the run to monitor the mixing and the convergence of the chain. Also, we need to adjust the jump size so as to get an acceptance rate of around 50%.

For updating ω and λ using Metropolis–Hastings algorithm we have used exponential scale, i.e., let ω_0 and λ_0 are the initial guesses for ω and λ, respectively, then the candidate states for them are taken as

$$\begin{aligned} \omega_{curr} &= \omega_0 \times exp(N(0, k_1^2)) \\ \lambda_{curr} &= \lambda_0 \times exp(N(0, k_2^2)) \end{aligned}$$

where, k_1 and k_2 are the jump sizes for updating ω and λ, respectively. Each of the Figures 1.3, 1.4 and 1.5 displays four different MCMC chains for the three different true λ values, respectively, considered for the simulation studies. In each case of the three true λ values considered

we have used different initial values, which are widely dispersed from each other, for the four different chains. The convergence plots for ω are displayed in Figure 1.6.

Figure 1.3: Convergence plot for λ when $\lambda = 0$

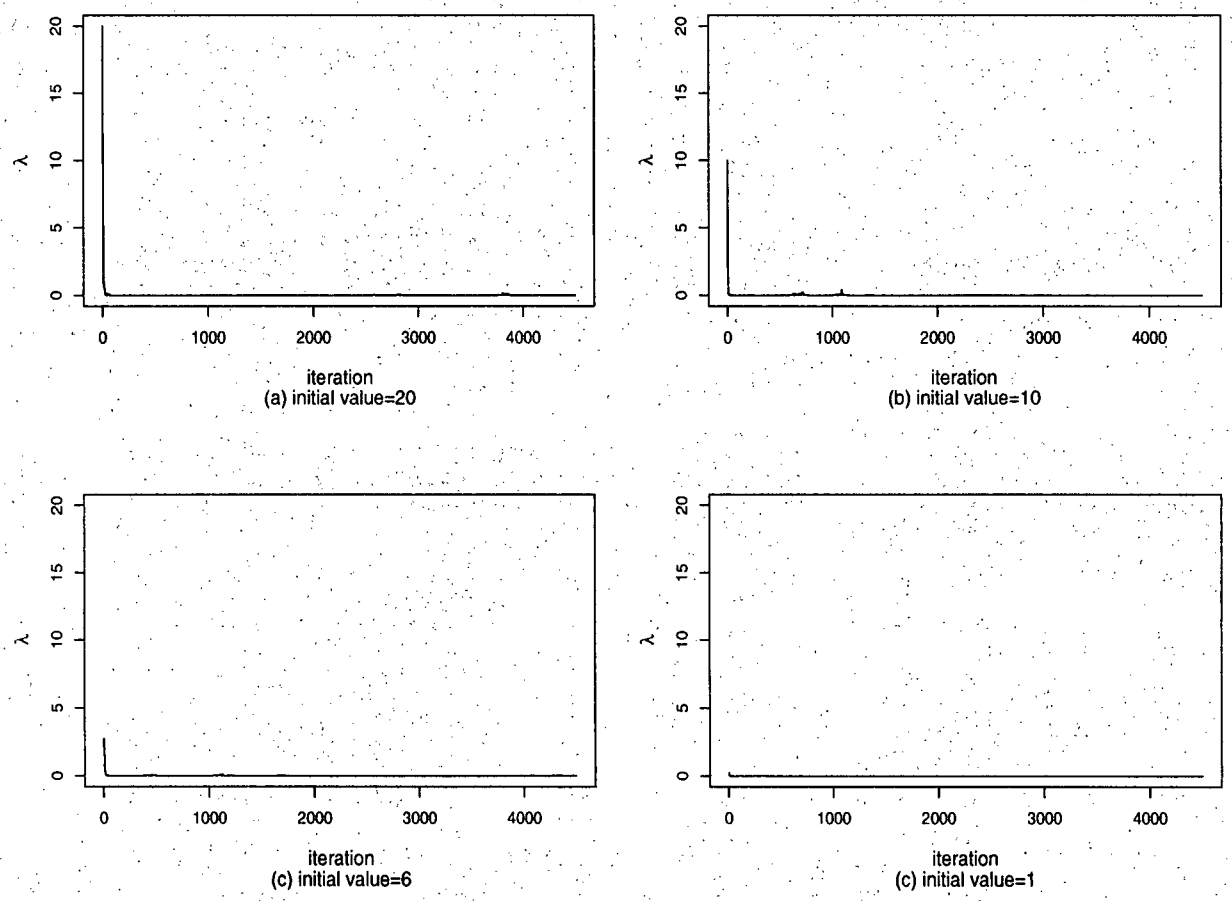


Figure 1.4: Convergence plot for λ when $\lambda = 1$

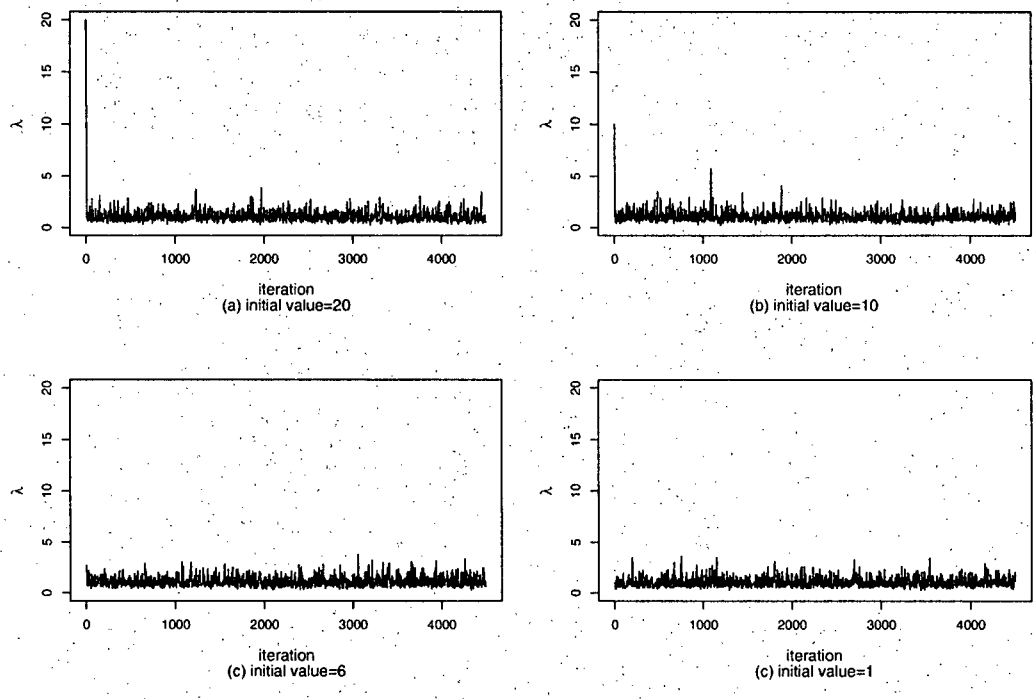


Figure 1.5: Convergence plot for λ when $\lambda = 5$

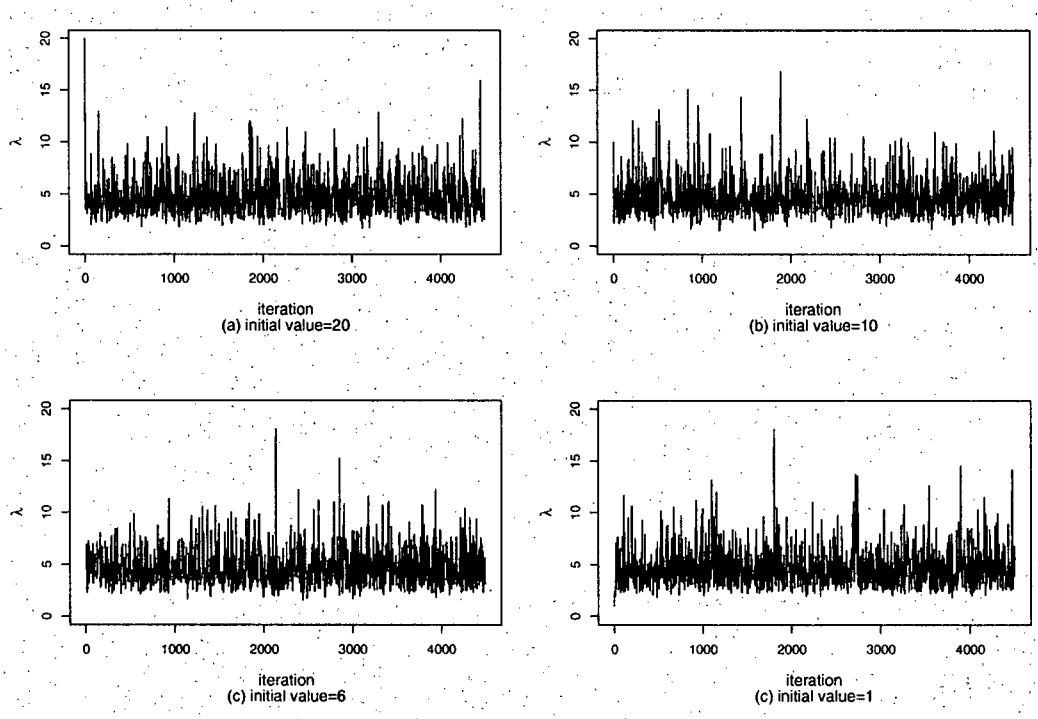
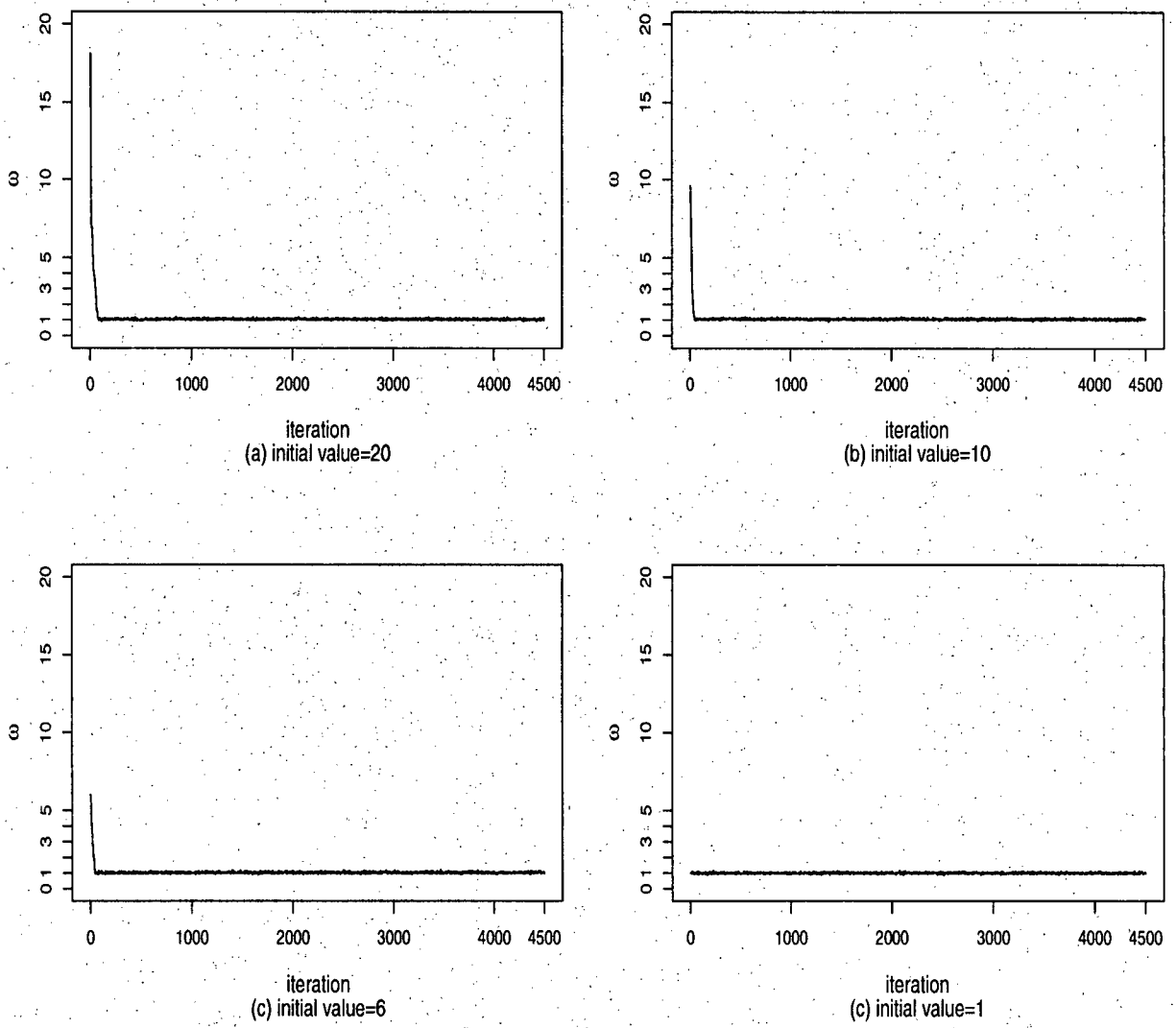


Figure 1.6: Convergence plot for ω



From Figures 1.3, 1.4 and 1.5 we observe that all the chins of each of the true λ values exhibit good mixing since none of the chains in any of the three cases kept moving around slowly in any particular region of the target distribution for many iterations. It is also observed that though different chains for each of the true λ values start at different initial values all of them have stabilized at the respective true λ values, which indicates that the MCMC chains for all of the λ values have converged to the respective target posterior distributions. Furthermore, all the cases the chains have stabilized after very few iterations. So, we can draw posterior inferences about λ by generating even a shorter chain because we do not need to throw away too many iterations as burn–in. Again, if we have a look at the plots of the Figure 1.6 we observe the same feature about the convergence and mixing of the chains for ω as we do in the case of convergence and mixing of the chains of λ.

1.6 Analysis of Output

Since the goal of this study is to choose appropriate value/values of the hyperparameter a for the proposed conservative prior we have estimated λ by using 100 different simulated data sets for different values of a. In estimating λ we have used both the posterior mean and the posterior mode. In each case we have constructed histogram of 100 $\hat{\lambda}$ and see how many of them exceeds the true λ value. For each of the model parameters we have run MCMC simulations for 2500 iterations, among which first 500 iterations have been thrown away as burn–in of the chain and used the remaining 2000 for inference purposes. Here, we have used 2500 iterations because from the convergence study of MCMC chains we have observed that for both ω and λ MCMC chains converge after very few iterations. The histograms of 100 estimates of λ have been displayed in Figures 1.7–1.12 for different true λ values. In each case, we have considered 6 different values of a as $a = 1, 2, 3, 4, 5$ and 10.

Figure 1.7: Histograms of $\hat{\lambda}$ for different choices of a in the case of posterior mean when true $\lambda = 0$

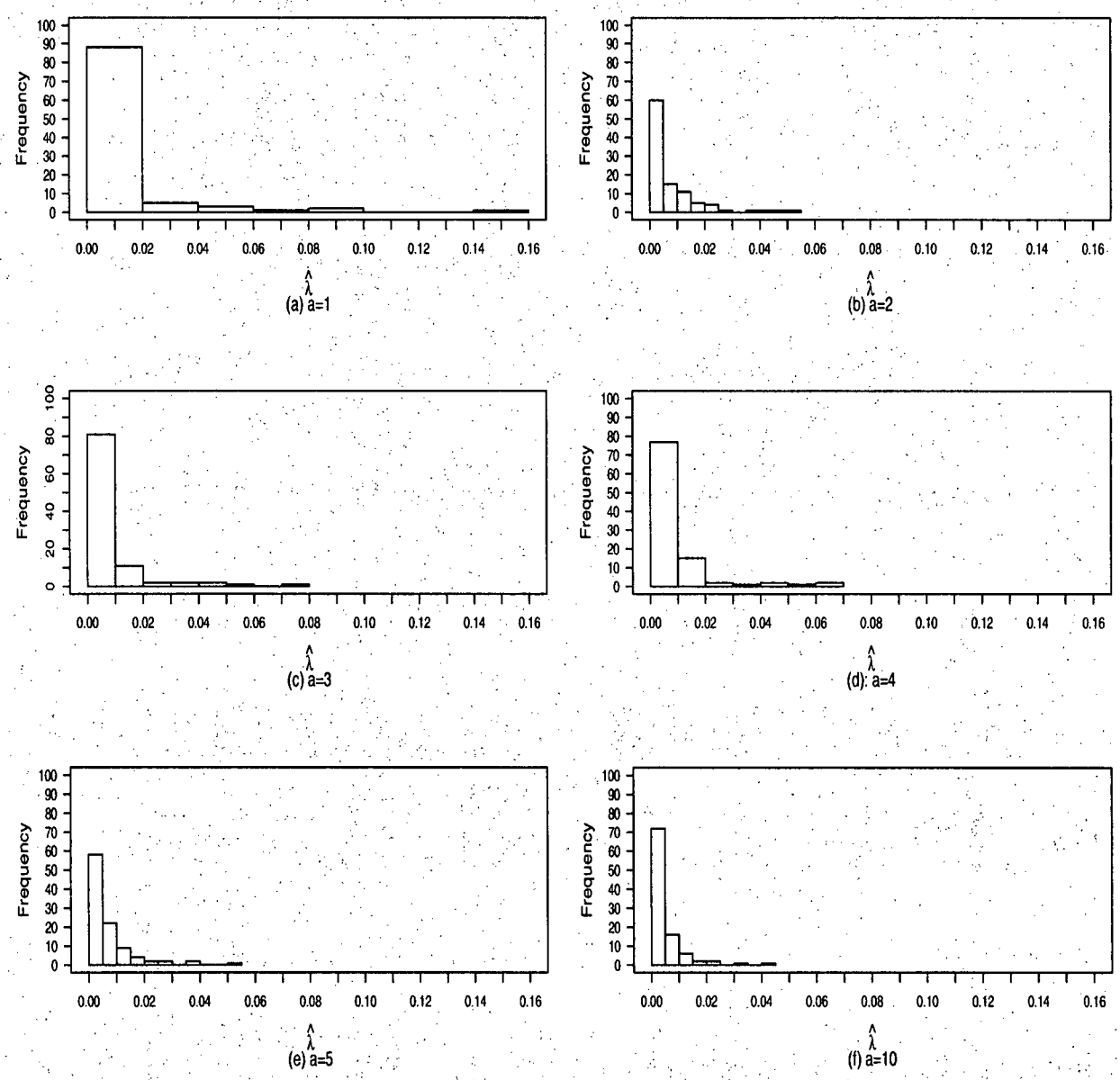


Figure 1.8: Histograms of $\hat{\lambda}$ for different choices of a in the case of posterior mode when true $\lambda = 0$

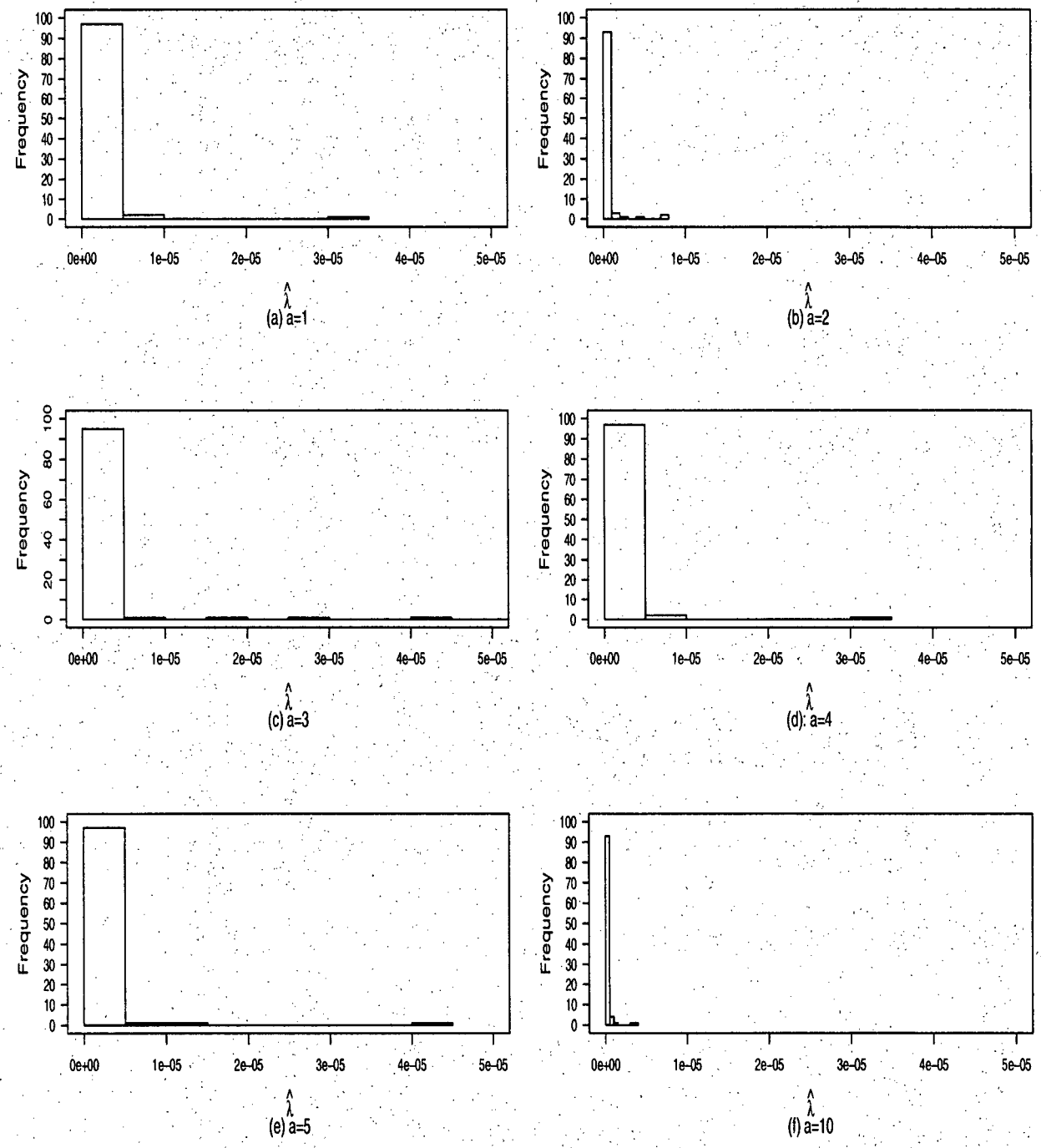


Figure 1.9: Histograms of $\hat{\lambda}$ for different choices of a in the case of posterior mean when true $\lambda = 1$

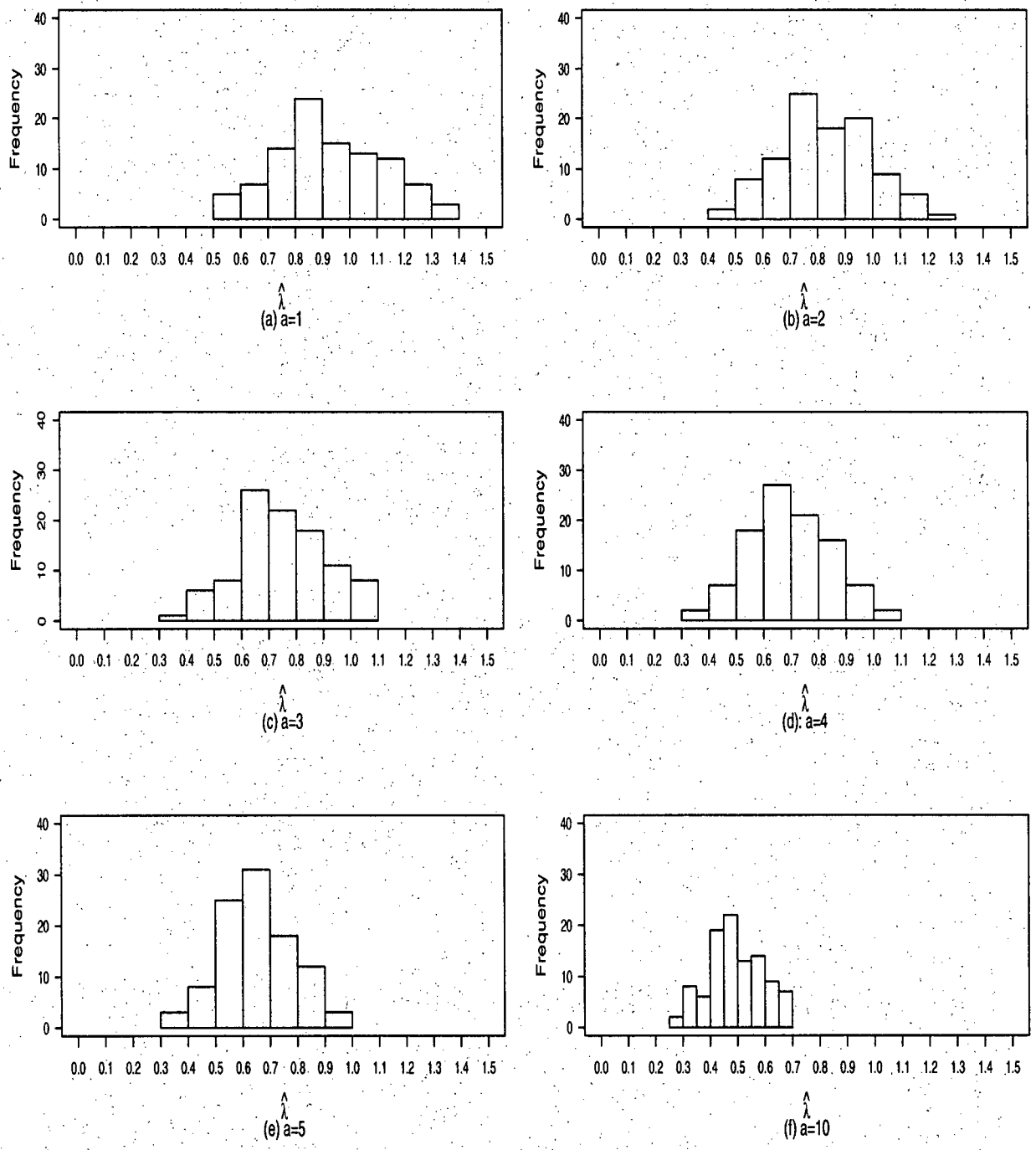


Figure 1.10: Histograms of $\hat{\lambda}$ for different choices of a in the case of posterior mode when true $\lambda = 1$

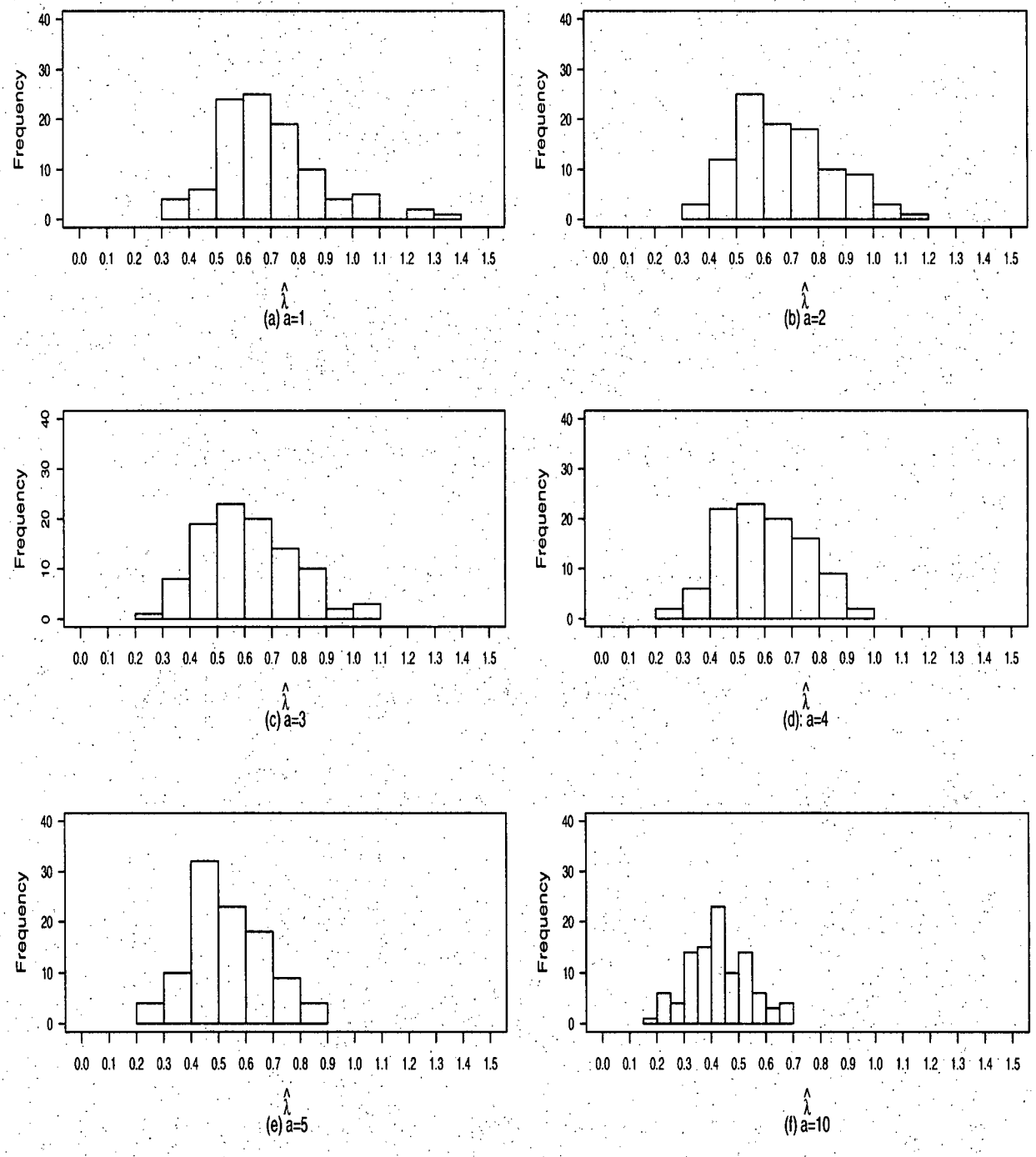


Figure 1.11: Histograms of $\hat{\lambda}$ for different choices of a in the case of posterior mean when true $\lambda = 5$

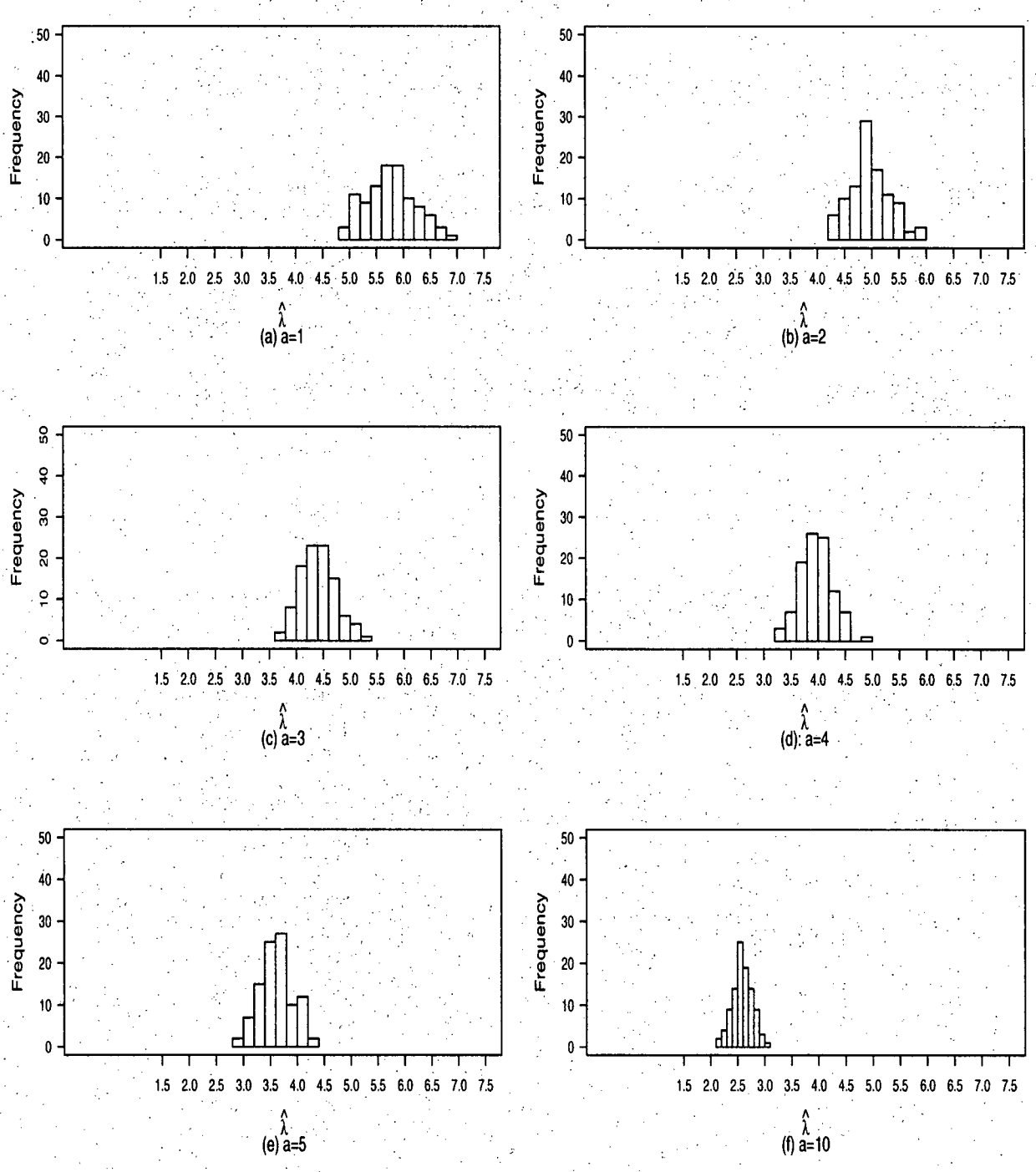
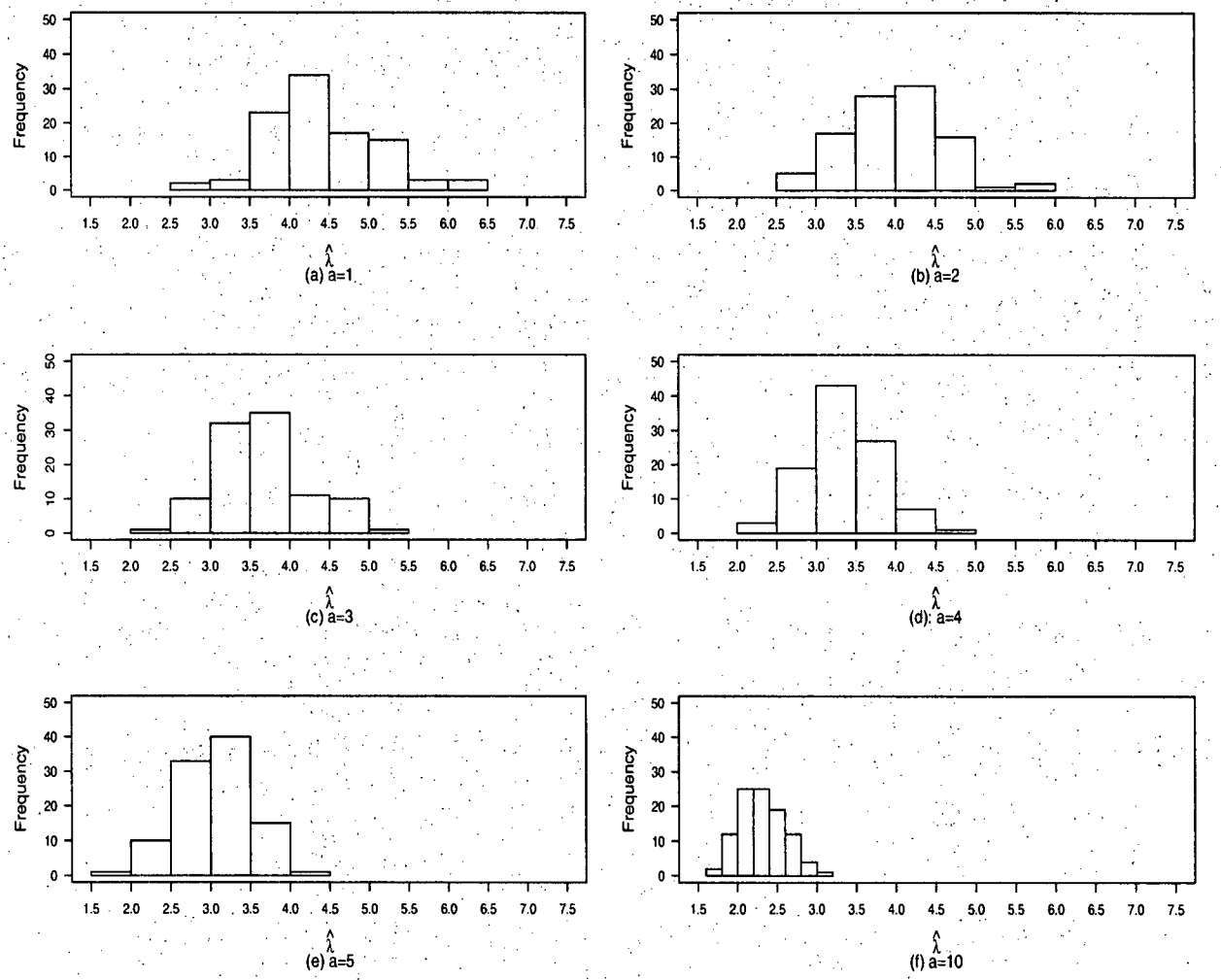


Figure 1.12: Histograms of $\hat{\lambda}$ for different choices of a in the case of posterior mode when true $\lambda = 5$



From the histograms of $\hat{\lambda}$, based on posterior mean, we observe that for $a \leq 3$ most of the estimates of λ exceeds the true value of λ for all the cases considered. But for $a \geq 3$ few of the estimates exceeds the corresponding true parameter value except the cases when $\lambda = 0$. Furthermore, the simulation studies reveal the fact that too big values of a cause serious underestimation. So, the general choice for a should be between 3–5. In the case when true $\lambda = 0$, though some of the estimates exceed the true λ values for values between 3–5 of a, the magnitudes of overestimation are quite small (Figures 1.7 and 1.8). Also, in prototype model, mathematically we have shown that the smaller the value of true λ the bigger the

chance of undersmoothing and for $\lambda = 0$ the probability of undersmoothing is the highest. So, Figures 1.7 and 1.8 confirm that a value between 3–5 of a performs quite well even in the worst case of undersmoothing. So, from the results of simulation studies we can conclude that use of the proposed conservative prior with a value between 3 to 5 for the hyperparameter a can sufficiently guard against undersmoothing when posterior mean is used as point estimate of λ.

Regarding the use of posterior mode as an estimate of λ, the histograms of $\hat{\lambda}$ (based on posterior mode) reveal that any value between 2 to 4 of a can guard quite well against undersmoothing, and in the case of no heterogeneity ($\lambda = 0$) posterior mode performs much better than posterior mean in guarding against undersmoothing. Even with $a = 1$ undersmoothing seems to be under control. In the case of posterior mode all the $\hat{\lambda}$ are very close to 0 (between 0 to 0.00005), the true value of λ. So, as long as the point estimate of λ is concerned posterior mode is preferable to posterior mean. But the problem of using the posterior mode is that it depends on the bin width used to calculate it. There is another problem, probably the more serious one in the context of smoothing, of depending on posterior mode as an estimate of λ while estimating the mean vector θ in Bayesian hierarchical model by using MCMC technique. The use of posterior mode as a point estimate of λ ends up with suggesting the use of smaller values for a in conservative prior compared to the case when posterior mean is used (2 to 4 versus 3 to 5). Use of smaller a values, as suggested by the posterior mode, may lead to undersmoothing of the phenomenon, i.e., smaller a values may produce θ estimates which are more variable than the true θs. Because, while estimating the θs by MCMC technique we generate θs conditional on the specific λ generated at each iteration of the MCMC run, i.e., we do not condition on the point estimate of λ while generating θs, rather we generate θs for each generated λ from its conditional posterior distribution. So, use of smaller a values, as suggested by the posterior mode, can generate a λ bigger than the true λ, which in turn results in generated θs that are more variable than they truly are. This problem seems to be more serious when true λ is 0. In this case the use of posterior mode as a point estimate of λ simulation studies suggest that even $a = 1$ performs greatly in guarding against undersmoothing.

Again, from the histograms of $\hat{\lambda}$ based on the posterior mean we observe that $a = 5$ sometimes can cause to much oversmoothing, especially, when λ is relatively larger than ω. Also, posterior mean suggests a value between 3 to 5 and posterior mode suggests a value between 2 to 4 for a to guard against undersmoothing. Since, both the point estimates of λ suggest a common range of values for a, which is between 3 to 4, hence, in general, irrespective of the choice of point estimate for λ we can conclude that any value between 3 to 4 for a in the conservative prior can reasonably be used to ensure adequate guard against undersmoothing, while at the

same time not oversmoothing the phenomenon too much and thus obtaining better calibrated estimates of the model parameters in normal–normal hierarchical models.

1.7 Comparison of Results Obtained by Using the Conservative Prior with those Obtained by Using the Jeffreys' Prior and the Uniform Shrinkage Prior

For comparing the performance of the proposed conservative prior with those of the uniform shrinkage prior and Jeffreys, priors we have conducted simulation studies. We have estimated λ from 100 different simulated data sets by using all the three competitive priors. We have examined the comparative performance of the priors from two different aspects– (i) first, with respect to their ability to guard against undersmoothing and (ii) second, since by introducing the conservative prior we force the variance component λ to be underestimated (oversmoothed), in a sense, we are introducing some sort of bias in $\hat{\lambda}$. So, the obvious question here is that how much we gain or lose with respect to mean–squared error (MSE) of $\hat{\lambda}$ by accepting some bias while estimating it. So, to reflect this issue we have also compared the MSE of $\hat{\lambda}$ for the competitive priors. For our simulation studies we considered $\lambda = 0, 1, 5$ and $\omega = 1$. In this case we have performed the simulation studies under both the small ($n = 10$) and large ($n = 100$) sample sizes.

In this study we have considered the uniform shrinkage prior and the Jeffreys' prior as the main competitors of the proposed conservative prior. The reason is that these two priors are widely used non–informative priors for Bayesian hierarchical models. Daniels (1999) showed that other priors are not good in estimating the random effects variance component, especially in the situation of no heterogeneity and in the cases when the random effect variance λ is relatively smaller with compared to error variance ω.

1.7.1 Analysis of Output

For comparing the performance of the three competitive priors we have calculated the MSEs of $\hat{\lambda}$ by using each of the three priors. In each case we have calculated the MSE by averaging over 100 simulated data sets. In all the cases we have used $a = 5$ for the conservative prior. MSEs are summarized in Tables 1.2, 1.3 and 1.4 for different true λ values considered.

Table 1.2: MSE of $\hat{\lambda}$ for different priors when $\lambda = 0$

Prior	Sample size	MSE
Conservative	10	0.0001124
Uniform shrinkage		0.0001368
Jeffreys'		0.0001871
Conservative	100	$3.1606e-06$
Uniform shrinkage		$3.5682e-06$
Jeffreys'		$3.5337e-06$

Table 1.3: MSE of $\hat{\lambda}$ for different priors when $\lambda = 1$

Prior	Sample size	MSE
Conservative	10	0.032352
Uniform shrinkage		0.428368
Jeffreys'		0.838135
Conservative	100	0.027283
Uniform shrinkage		0.435985
Jeffreys'		0.810519

From Tables 1.2, 1.3 and 1.4 we see that, irrespective of sample size, MSE of $\hat{\lambda}$ is the minimum for the conservative prior and the maximum for the Jeffreys' prior when the true λ values are smaller or of equal magnitude ($\lambda = 0$ or $\lambda = 1$) with compared to the within group variance ω. But when the true λ is larger compared to the error variance ω the MSE for the conservative prior is the maximum while that for uniform shrinkage prior is the minimum in the case of small sample size. In case of large sample size the MSE of $\hat{\lambda}$ is the maximum for Jeffreys' prior and minimum for the uniform shrinkage prior. The interesting point is that for large sample the MSE increases for both uniform shrinkage prior and Jeffreys' prior but decreases for conservative prior when true λ value is relatively bigger (Table 1.4). This might be due the fact that, as it can be observed from Equation (1.1) and from Figure 1.1, as sample size gets larger the probability of undersmoothing increases toward a limit of 0.5, and so there is more chance of getting bigger values in the posterior simulations of λ, which inflates the estimated λ and thus contributing to higher MSE of $\hat{\lambda}$. But in case of conservative prior the bigger a value suppresses those upward jumps in the posterior simulations for λ, and controls the undersmoothing, which in turn does not let the MSE to go up even when sample size gets larger.

Table 1.4: MSE of $\hat{\lambda}$ for different priors when $\lambda = 5$

Prior	Sample size	MSE
Conservative	10	1.995964
Uniform shrinkage		0.436016
Jeffreys'		0.831997
Conservative	100	1.808555
Uniform shrinkage		0.666834
Jeffreys'		3.77446

Before drawing general conclusion about the comparative performance of the three competitive priors there is one point worth focusing on. So far, for calculating the MSE of $\hat{\lambda}$ we have considered $a = 5$. Previously, from simulation studies, we have observed that any value between 3 to 5 of a is sufficient for guarding against undersmoothing while we use the proposed conservative prior. But if we have look at the histograms of $\hat{\lambda}$ in Figures 1.11 and 1.12 we observe that using $a = 5$ for conservative prior gives a rather underestimation (oversmoothing) for λ when true λ is relatively larger, which might be the cause of bigger MSE of $\hat{\lambda}$ in this case. Since any value between 3 to 5 of a is good enough to guard against undersmoothing and since for the case of relatively larger between group variance use of $a = 5$ gives too much underestimation it may be reasonable to use $a = 3$ and see how the conservative prior perform with respect to MSE of $\hat{\lambda}$ in all the cases we have considered. To check this we have computed the MSE of $\hat{\lambda}$ again for $a = 3$. Tables 1.5, 1.6 and 1.7 summarize these MSEs.

Table 1.5: MSE of $\hat{\lambda}$ for different priors, with $a = 3$ for conservative prior, when $\lambda = 0$

Prior	Sample size	MSE
Conservative	10	0.0001216
Uniform shrinkage		0.0001368
Jeffreys'		0.0001871
Conservative	100	$3.2738e-06$
Uniform shrinkage		$3.5682e-06$
Jeffreys'		$3.5337e-06$

Looking at the output in Tables 1.5, 1.6 and 1.7 we observe that with $a = 3$ the conservative prior produces the best results in terms of both guarding against undersmoothing and MSE in

Table 1.6: MSE $\hat{\lambda}$ for different priors, with $a = 3$ for conservative prior, when $\lambda = 1$

Prior	Sample size	MSE
Conservative	10	0.117817
Uniform shrinkage		0.428368
Jeffreys'		0.838135
Conservative	100	0.126487
Uniform shrinkage		0.435985
Jeffreys'		0.810519

Table 1.7: MSE $\hat{\lambda}$ for different priors, with $a = 3$ for conservative prior, when $\lambda = 5$

Prior	Sample size	MSE
Conservative	10	0.484169
Uniform shrinkage		0.436016
Jeffreys'		0.831997
Conservative	100	0.474004
Uniform shrinkage		0.666834
Jeffreys'		3.77446

estimating λ in all but one of the cases considered. The only case where the uniform shrinkage prior marginally beats (an MSE of 0.4842 for conservative prior versus an MSE of 0.4360 for uniform shrinkage prior) conservative prior is when between group variability is larger than the within group variability and sample size is small. So, in general, we can conclude that though the proposed conservative prior has been introduced with the aim of guarding against undersmoothing simulation studies shows that for a reasonably chosen value of hyperparameter a it can also produce better calibrated estimate for the random effects variance component and hence for the other parameters in hierarchical models.

1.8 Conclusion

In this chapter we have introduced a class of non–informative priors for the random effects variance component in Bayesian hierarchical models. This class of priors can give rise to the widely used priors in Bayesian hierarchical models such as, the Jeffreys' prior and the uniform

shrinkage prior. Finally, for estimating the variance component λ by Bayesian approach we have incorporated the idea of smoothing to choose the appropriate values of the hyperparameter a for the considered class of priors to get a prior which can sufficiently guard against undersmoothing. We named this prior as "conservative prior". Simulation studies reveal that we can achieve desired guard against undersmoothing if we choose a value between 3 to 5 for the hyperparameter a. But guarding against undersmoothing may result in some degree of bias in the parameter estimates and hence an increase in the MSE of of the estimated parameters. Keeping this aspect in mind we have compared the performance of conservative prior in estimating the variance component λ with those obtained by using the Jeffreys' prior and the uniform shrinkage prior with the help of simulation studies. From simulation studies we have found that use of $a = 5$ for conservative prior produces smaller MSE for $\hat{\lambda}$ with compared to those produced by the use of other two priors when between group variability is relatively smaller than the within group variability, but produces higher MSE compared to the other two priors in the reverse situation except the case that it produces smaller MSE than Jeffreys' prior does when sample size is large. From simulation studies it is also observed that relatively higher MSE for $\hat{\lambda}$ for conservative prior with $a = 5$ in the situation of relatively higher between group variation is actually due to the fact that using $a = 5$ in this situation gives to much underestimation of λ (Figures 1.11 and 1.12). But using $a = 3$ nicely controls undersmoothing as well as gives lower MSE by providing more precise estimate of λ, which in turn produces better calibrated estimates of the other model parameters.

Finally, though we have conducted our simulation studies for the situations of between and within group variance ratios of 0, 1 and 5, it can be argued that for most of the practical situations it is very unlikely that the ratio of between group variability to the within group variability would be very high, e.g. as large as 5. For instance, if we think of a hospital model of cardiac treatment it is very unlikely that the survival probability θ_i will differ too much from one hospital to another. In such cases undersmoothing should be main concern and there is not too much to get worried in using any value between 3 to 5 for the hyperparameter a in conservative prior to get better calibrated estimates of the model parameters and thus have a better real life picture. In the case where it is really expected that the between group variability would be much larger than the within group variability we can use $a = 3$ for better calibrated estimates of the model parameters.

Chapter 2

Curve Fitting

2.1 Introduction

In chapter 1, we have studied the performance of the proposed conservative prior in Bayesian normal–normal hierarchical model, which is actually an application of Bayesian hierarchical model to random effects model for normal response. Another potential area of using the Bayesian hierarchical model is non–parametric curve–fitting. In this chapter we have briefly described the idea of non–parametric curve–fitting and the spline–based roughness penalty approach to curve–fitting.

Curve fitting or regression function estimation is a common and most useful tool in statistics. It has two purposes–firstly, it provides a way of analyzing and presenting the dependence of a response variable on the design variables; secondly, it allows prediction of unobservable future responses. Suppose we have n measurements on a response variable y, and a single predictor variable x. In general, the dependence of y on x can be expressed as

$$y = f(x) + \epsilon \tag{2.1}$$

where f is the curve of some sort and ϵ is the random noise. The objective of curve fitting is to estimate the function f in (2.1) from the observed data. There are two approaches to curve fitting– (i) parametric approach and (ii) non-parametric approach.

2.2 Parametric Approach

The parametric approach imposes rigid parametric assumptions about the dependence between the response and the predictor. For instance, in the case of linear regression the regression function f in (2.1) is assumed to be linear. In general, in parametric approach response are assumed to follow a parametric distribution, e.g., a distribution from exponential family. Under the exponential family the dependence of response can be summarized under the framework of generalized linear model (GLM)(McCullah and Nelder, 1989). The simplest version of GLM is the linear regression model where the responses are assumed to follow normal distribution, thus making f a linear function of x, i.e., $y = \alpha + \beta x + \epsilon$. That is, every parametric method requires rigid assumption on the form of f, which may not be true, and hence the question of a non–parametric approach comes through.

2.3 Non-parametric Approach

Data for which none of the parametric method seems reasonable are often envisaged. In such cases, it is wise to let the data to show us the appropriate functional form rather than imposing any definite parametric form. Non–parametric methods are very flexible in allowing the data itself to decide on how the dependence pattern should be. The underlying assumption here is that the dependence of the mean of y on x should not change much if x does not change much. This assumption is very often reasonable. This assumption can be interpreted as that we want an estimate of f, say $\hat{f}$, which is at most as variable as f itself, i.e., we don't want $\hat{f}$ to be more wiggly than f. So, the non–parametric approach involves the choice of a smoothing parameter which controls the balance between goodness of fit and smoothness of the estimated regression function. There are different non-parametric methods for estimating regression function. Interested readers are referred to Simonoff (1996). Since our work is aimed at developing a conservative prior for Bayesian approach to spline–based roughness penalty approach of curve estimation, descriptions of the spline–based roughness penalty approach and associated techniques are given in the next few sections.

2.3.1 Spline–based Roughness Penalty Approach

The basic idea of the spline–based roughness penalty approach is to quantify the notion of a rapid fluctuating curve with the help of a piece–wise polynomial of certain degree (called spline) over each subintervals of the range of the predictor considered, and then pose the estimation problem in such a way that makes the necessary compromise between the two opposing aims of curve estimation. The opposing aims are–(i) goodness of fit, which measures the closeness of the estimated curve to the true underlying curve and (ii) roughness, which measures the wigglyness (local variability) of the estimated curve. In general, it is always desirable to have an estimated curve which provides a good fit as well as not too wiggly. The roughness penalty approach makes a compromise between these two opposing factors. In our discussion we will focus on only cubic splines because polynomials of degree higher than three are too wiggly, while lower degree polynomials are not flexible enough to capture the local variation of the data. A cubic spline can capture the local variation because it has two continuous derivatives and at the same time it is not too wiggly. It is also very amenable for mathematical manipulation. Eubank (1988), Wahba (1990) and Green & Silverman (1994) are excellent reference books on spline–based smoothing techniques. As our work will be based on cubic spline/natural cubic spline, formal definitions of cubic spline and natural cubic spline have been given below.

Cubic Spline (CS): The function f is said to be a cubic spline on the interval $[a, b]$, satisfying $a < t_1 < t_2 < \ldots < t_n < b$, if f is a cubic polynomial on each of the intervals $(a, t_1), (t_1, t_2) \ldots, (t_n, b)$ and the polynomial pieces fit together at the points t_i in such a way that f itself and its first and second derivatives are continuous at each t_i, and hence on the whole interval $[a, b]$. The points t_i's are called the knots.

Natural Cubic Spline (NCS): A cubic spline on an interval $[a, b]$ is said to be natural cubic spline if its second and third derivatives are zero at the boundaries a and b. These conditions are known as natural boundary conditions. These imply that f is linear on the two extreme interval $[a, t_1]$ and $[t_n, b]$. Natural cubic spline has enormous mathematical convenience because it can be exactly specified by finding a finite number of constants (Green and Silverman, 1994).

2.3.2 Mathematical Formulation of the Roughness Penalty Approach

For constructing an estimate of the curve of type 2.1, whose observed values are subject to random error, suppose that $t_1, t_2, \ldots, t_n$ are the points in $[a, b]$ satisfying $a < t_1 < t_2 < \ldots < t_n < b$ and that $y_1, y_2, \ldots, y_n$ are the observed values. Let $S_2[a, b]$ be the space of continuous twice differentiable functions, then for any function f in $S_2[a, b]$ the penalized sum of squares is defined to be

$$S(f) = \sum_{i=1}^{n} \{y_i - f(t_i)\}^2 + \alpha \int_a^b \{f(x)''\}^2 dx \tag{2.2}$$

In the roughness penalty approach, $\hat{f}$ is calculated so as to be the minimizer of $S(f)$ over the class $S_2[a, b]$ of all sufficiently smooth curves on $[a, b]$. The second term in (2.2) is the roughness penalty term. The addition of the term $\alpha \int (f'')^2 dx$ in (2.2) ensures that the cost of $S(f)$ of the particular curve is determined not only by its goodness–of–fit to the data quantified by the first term but also by its roughness $\int (f'')^2 dx$. The smoothing parameter α represents the strength of penalty to be paid. Large value of α represents stronger penalty, which produces a smooth curve. On the other hand, if α is relatively small then the main contribution to $S(f)$ will be the residual sum of squares and the curve estimate $\hat{f}$ will track the data closely, thus producing a wiggly curve. Thus minimization of $S(f)$ can give a compromise between smoothness and goodness-of-fit. In implementing roughness penalty approach the obvious choice is the natural cubic spline because NCS produces the smoothest possible curve among all the polynomials with continuous second derivatives in $S_2[a, b]$ minimizing $S(f)$ (Green and Silverman, 1994). Also knowing that $\hat{f}$ is a natural cubic spline has many advantages. We can specify $\hat{f}$ exactly by finding a finite number of constants because we only need to minimize $S(f)$ over a finite dimensional class of functions, the NCS's with knots at t_i, instead of considering the infinite dimensional set of smooth functions $S_2[a, b]$. Also, with an NCS there is an elegant algorithm of how the minimizing spline curve can be found by solving a set of linear equations.

2.3.3 Elegant Way of Representing NCS

The most elegant way of representing NCS is the value second derivative representation (Green and Silverman, 1994). In this method an NCS can be specified by its value and the second

derivatives at each of the knots t_i. Let the spline have n equally spaced knots satisfying $t_1 < t_2 < \ldots < t_n$. Define $f_i = f(t_i)$ and $\gamma_i = f''(t_i)$ for $i = 1, 2, \ldots, n$. For an NCS we have $\gamma_1 = \gamma_n = 0$. Let f be the vector of $(f_1, f_2, \ldots, f_n)'$ and γ be the vector of $(\gamma_2, \gamma_3, \ldots, \gamma_{n-1})'$. For specification of the NCS with the help of f and γ one needs to define two matrices Q and R using the knot values.

Let $h_i = t_{i+1} - t_i$ for $i = 1, 2, \ldots, n-1$. The Q matrix is an $n \times (n-2)$ matrix with entries $q_{ij}, i = 1, 2, \ldots, n;\ j = 2, \ldots, n-1$, defined as

$$\begin{aligned} q_{j-1,j} &= h_{j-1}^{-1} \\ q_{j,j} &= -h_{j-1}^{-1} - h_j^{-1} \\ q_{j+1,j} &= h_j^{-1} \end{aligned}$$

for $j = 2, 3, \ldots, n-1$, and $q_{ij} = 0$ for $|i - j| \geq 2$. The symmetric matrix R is an $(n-2) \times (n-2)$ matrix with entries $r_{ij},\ i, j = 2, 3, \ldots, n-1$, defined as

$$r_{i,i} = \frac{1}{3}(h_{i-1} + h_i), \text{for } \ i = 2, 3, \ldots, n-1$$

$$r_{i,i+1} = r_{i+1,i} = \frac{1}{6}h_i, \text{for } \ i = 2, 3, \ldots, n-2$$

and $r_{ij} = 0$ for $|i - j| \geq 2$. Using these facts two important results are stated in Green and Silverman (1994), which are given by the following theorem.

Theorem: The vectors f and γ specify a natural cubic spline if the condition $Q^T f = R\gamma$ is satisfied. If the above condition is satisfied then the roughness penalty will satisfy

$$\begin{aligned} \int_a^b f''(t)^2 dt &= \gamma^T R \gamma \\ &= f^T K f \end{aligned}$$

with $K = QR^{-1}Q^T$.

It can be verified that the value of the cubic spline at any point t is given by (Green and Silverman)

$$f(t) = \frac{(t - t_i)f_{i+1} + (t_{i+1} - t)f_i}{h_i} - \frac{1}{6}(t - t_i)(t_{i+1} - t)\left\{\left(1 + \frac{t - t_i}{h_i}\right)\gamma_{i+1} + \left(1 + \frac{t_{i+1} - t}{h_i}\right)\gamma_i\right\}$$
$$\text{for } \ t_i \leq t \leq t_{i+1};\ \ i = 1, 2, \ldots, n-1 \qquad (2.3)$$

In matrix notation, (2.3) can be written as

$$(f(x_1), f(x_2), \ldots, f(x_N)) = Af \tag{2.4}$$

where A is a matrix of order $N \times n$ whose (j)th row is obtained as

$$w_{ji} = \frac{(t-t_i)}{h_i} - \frac{(t-t_i)(t_{i+1}-t)}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)c_{i+1,i} + \left(1+\frac{t_{i+1}-t}{h_i}\right)c_{i,i}\right\}$$

$$w_{j,i+1} = \frac{(t-t_i)}{h_i} - \frac{(t-t_i)(t_{i+1}-t)}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)c_{i+1,i+1} + \left(1+\frac{t_{i+1}-t}{h_i}\right)c_{i,i+1}\right\}$$

$$\text{and} \quad w_{j,i'} = -\frac{(t-t_i)(t_{i+1}-t)}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)c_{i+1,i'} + \left(1+\frac{t_{i+1}-t}{h_i}\right)c_{i,i'}\right\}$$

$$\text{for} \quad i' = 1, 2, \ldots, i-1, i+1, \ldots, n$$

for any t between t_i and t_{i+1}, where c_{ij} is the (ij)th element of $R^{-1}Q^T$ and N is the number of knot points for which f needs to be estimated.

2.4 Selection of Smoothing Parameter for Spline Smoothing

Though spline smoothing is a very popular non–parametric technique for estimating a regression function the performance of it depends on the choice of smoothing parameter. So, the choice of smoothing parameter is the most essential task of spline smoothing. There are different methods for an automated selection of smoothing parameter. The most widely used ones in frequentist approach include the cross–validation (CV) and the generalized cross–validation (GCV) methods.

2.4.1 Cross–validation Method

The basic idea behind CV method is in terms of prediction. It uses the principle of "leave–one–out" prediction. The idea is to leave the data points out one at a time and to select the smoothing parameter α in equation (2.2) under which the removed data points are best predicted by the remaining data.

Let $\hat{f}^{(-i)}(t, \alpha)$ be the curve estimate from all the data except y_i using a smoothing parameter value α. Then the cross–validation choice of α is the value of α that minimizes the CV criterion

defined as

$$CV(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{f}^{(-i)}(t_i,\alpha)\right\}^2. \tag{2.5}$$

An easier computational form of CV criterion defined by equation (2.5) has been given by Craven and Wahba (1979) as

$$CV(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(t_i,\alpha)}{1 - B_{ii}(\alpha)}\right)^2, \tag{2.6}$$

where $\hat{f}$ is the spline smoother calculated from the full data set (t_i, y_i) with the smoothing parameter α and $B(\alpha) = (I + \alpha QR^{-1}Q^T)^{-1}$.

2.4.2 Generalized Cross–validation approach

Generalized cross–validation (Cravan and Wahba, 1979) method is a modified form of CV method. It is obtained by replacing $B_{ii}(\alpha)$ in equation (2.6) by its average value, $n^{-1}trB(\alpha)$ as–

$$GCV(\alpha) = \frac{\frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{f}(t_i,\alpha)\}^2}{\{1 - n^{-1}trB(\alpha)\}^2} \tag{2.7}$$

As in ordinary CV method in GCV method the smoothing parameter α is obtained by minimizing GCV score given by the equation (2.7). Theoretical arguments show that GCV produces asymptotically best possible choice for α in the sense of minimizing the average squared error at the design points. Also, GCV method has some computational advantages over CV method (Silverman, 1984).

2.5 Conclusion

In this chapter we have discussed the roughness penalty approach to curve–fitting. The most important consideration in the roughness penalty approach is choosing the smoothing parameter. Two widely used frequentist methods in this regard are the CV and GCV. Both the methods have some limitations. The most serious one is that though in smoothing problem it is believed that the true underlying regression curve is a smooth one the CV or GCV approach

sometimes gives a smoothing parameter that produces an estimated curve which is very under-smooth (wiggly). Secondly, the basic idea of CV or GCV approach is to choose the value of the smoothing parameter α which minimizes the CV or GCV score. But in some cases CV or GCV score may be a monotone function of α, and in those cases no obvious solution is available. Thirdly, in cases when the function CV or GCV has not a unique minimum a simple grid search is usually used to locate the global minimum. In such cases it is not easy to determine that how small the grid length should be so that the global minimum can be located. From this view point CV or GCV is not automated method in the strict sense. Again, with CV or GCV approach we can just estimate and predict the response as a function of the predictor, but we can not test the hypothesis that whether the response predictor relationship is significant. Considering all the limitations of CV or GCV approach it is desirable to have a more automated approach of selecting the smoothing parameter to get an estimated curve which is not more wiggly than the true curve itself and which can provide means to draw conclusion about the significance of the response–predictor relationship.

Chapter 3

Bayesian Representation of Roughness Penalty Approach of Curve Fitting: Methodology and Simulation Studies

3.1 Introduction

In seeking an approach to regression curve estimation which can guard against undersmoothing, we can think of a Bayesian formulation of roughness penalty approach. Also, in the Bayesian framework of roughness penalty approach we can think of making use of the proposed conservative prior. Because simulation studies show that the proposed prior, with suitably chosen hyperparameter, performs quite well in guarding against undersmoothing in the case of simple normal–normal hierarchical model. In this regard, first of all we need to define the Bayesian formulation of the roughness penalty approach to curve–fitting. The current section gives the details of the Bayesian formulation of the roughness penalty approach.

The roughness penalty approach has straight forward Bayesian representation. Let y be the response vector and x be the vector of the values of the design variable. In smoothing problem an adequate summary of the relationship between a response variable y and a predictor variable x is provided by the smoothing model

$$y = f(x) + \epsilon \tag{3.1}$$

where, ϵ is usually multivariate normal with mean vector 0 and the variance matrix $\sigma^2 I_n$. Define n knots $t_1, \ldots, t_n$ over the values of the design variable x. Let $f_i = f(t_i)$, so that $f = (f_1, \ldots, f_n)'$ be the vector of the values of $f(x)$ at the knots. With $f(x)$ being a natural cubic spline with knots at $t_i = x_i$, the penalized sum of squares can be expressed as

$$S(f) = (y - f)'(y - f) + \lambda f' K f \tag{3.2}$$

where $f'Kf$ is the measure of roughness (Green and Silverman, 1994; Hastie and Tibshirani, 1990). In the frequentist roughness penalty approach the goal is to estimate f such that $S(f)$ is minimum. In Bayesian paradigm there is a nice straightforward representation for this roughness penalty approach. In this chapter we elaborate on the Bayesian representation of roughness penalty approach of smoothing problem. In Bayesian representation the roughness penalty approach to curve–fitting can be expressed as a multivariate normal–normal hierarchical model, where the smoothing parameter can be taken as the function of the random effects variance component. Hence, in order to guard against undersmoothing we have used the conservative prior for the variance component. We have conducted simulation studies to see how the proposed conservative prior can perform to guard against undersmoothing. Also, on the basis of simulation studies, we have compared the performance of conservative prior and that of uniform shrinkage prior (Daniels, 1999) and Jeffreys' prior with respect to guarding against undersmoothing as well as with respect to MSE of the estimated curve. We have considered uniform shrinkage prior as the main competitor of conservative prior with respect to guarding against undersmoothing since it can provide some degree of guard against undersmoothing due to its property of shrinkage toward the situation of no global heterogeneity in the data set. Finally, on the basis of simulation studies, a comparison of Bayesian approach using conservative prior and Bayesian approaches using uniform shrinkage prior and Jeffreys' prior have been made to see how each of the competitive priors perform with respect to the mean squared error of the estimated curve.

3.2 Bayesian Representation of Roughness Penalty Approach

Suppose, we have the data (y_i, x_i); $i = 1, 2, \ldots, n$ and the model $y_i = f(x_i) + \epsilon_i$. Assume that $y_i \sim N(f(x_i), \sigma^2)$. Let $f(x)$ be a natural cubic spline. Define distinct knots at $x = t_0, x = t_1, \ldots, x = t_p, x = t_{p+1}$ over the values of the design variable x with $f(t_0) = \theta_0$, $f(t_1) = \theta_1$, ..., $f(t_p) = \theta_p$, $f(t_{p+1}) = \theta_{p+1}$. Here, $\theta_0, \theta_1, \ldots, \theta_{p+1}$ are unknown parameters with $f(t_j) = \theta_j$; $j = 0, 1, \ldots, p+1$. By defining the vector of parameters $\theta = (\theta_0, \theta_1, \ldots, \theta_{p+1})$ at the selected knots only we represent the true function $f(x)$ with a lower dimensional vector θ. The main advantage of such a parameterization of the true function $f(x)$ is that we are able to estimate the entire function $f(x)$ by estimating a fewer number of parameters. For example, there may be the cases where the number of design points is over several hundred or even thousand. In such cases it is not computationally reasonable to estimate $f(x)$ at all the design points. In such situations even 10 or 20 knots can give adequate approximation to the

true curve. In other words as long as smoothing is concerned, fewer number of knots can give the same level of flexibility as the case of taking knot at each design point.

For a natural cubic spline it is assumed that $f(x)$ is linear outside the boundary knots. Under this assumption we can consider the model (3.1) to be composed of a linear part and a smoothed part. The model can then be written as

$$y = \beta_0 + \beta_1 x + f(x) + \epsilon \tag{3.3}$$

Now, for simplicity, we can assume that linear trend has been removed so that $f(t_0) = 0$ and $f(t_{p+1}) = 0$. In this case, $\theta = (\theta_1, \ldots, \theta_p)'$. Under the assumption that linear part has been removed model (3.3) can be written as $y - \beta_0 - \beta_1 x = f(x) + \epsilon$, or equivalently,

$$\tilde{y} = f(x) + \epsilon \tag{3.4}$$

where $\tilde{y} = y - \beta_0 - \beta_1 x$. Removing the linear part from the model (3.3) helps in defining a proper prior for the parameter vector θ. The smoothed part $f(x)$ and the linear part can be estimated simultaneously by Bayesian back–fitting algorithm (Haste and Tibshirani, 2000). In this study we have concentrated on estimating the smoothed part $f(x)$ only. Under the model (3.4) we have p-dimensional parameter vector $\theta = (\theta_1, \ldots, \theta_p)'$ to represent the true function $f(x)$. In this case, we have $f'Kf = \theta' K\theta = \int [f''(t)]^2 dt$ as the measure of roughness of f. The penalized sum of squares can now be expressed as

$$S(f) = (\tilde{y} - f)'(\tilde{y} - f) + \lambda \theta' K \theta \tag{3.5}$$

Since we represent $f(x)$ with the parameter vector θ we estimate $f(x)$ only at the selected knot points as $\hat{\theta}$. So, to estimate the entire curve we can use the method of interpolating natural cubic spline (Green and Silverman, 1994) to have $\hat{f}(x) = A\hat{\theta}$, where A is an $n \times p$ design matrix. The computation of the matrix A has already been discussed in section 2.3.3.

Now, for Bayesian representation we can write

$$\tilde{y} \sim N(A\theta, \sigma^2 I_n),$$

where $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)'$ is the $n \times 1$ response vector. So, we can write $f = A\theta$, and the penalized sum of squares as

$$S(f) = (\tilde{y} - A\theta)'(\tilde{y} - A\theta) + \lambda \theta' K \theta \tag{3.6}$$

Consider the prior distribution for the parameter vector θ as $\theta | \tau^2 \sim N(0, \tau^2 V)$, where, $V = K^{-1}$ is a known $p \times p$ matrix (K has been defined in section 2.3.3). The hierarchical model for the

smoothing problem (3.4) can be expressed as

$$\begin{aligned} \tilde{y}|\theta,\sigma^2 &\sim N(A\theta,\sigma^2 I_n) \\ \theta|\tau^2 &\sim N(0,\tau^2 V) \end{aligned} \tag{3.7}$$

The posterior distribution for θ, given τ^2, σ^2 and the data, is obtained as

$$\begin{aligned} \pi(\theta|\sigma^2,\tau^2,\tilde{y}) &\propto f(\tilde{y}|\theta,\sigma^2)\pi(\theta|\tau^2) \\ &\propto e^{-\frac{1}{2\sigma^2}(\tilde{y}-A\theta)'(\tilde{y}-A\theta)} e^{-\frac{1}{2\tau^2}\theta' V^{-1}\theta} \\ &= e^{-\frac{1}{2\sigma^2}\{(\tilde{y}-A\theta)'(\tilde{y}-A\theta)+\frac{\sigma^2}{\tau^2}\theta' V^{-1}\theta\}} \end{aligned}$$

where, $\pi(\theta|\tau^2)$ is the prior for θ.

It can be easily verified that the posterior distribution of θ is normal. The Bayesian estimate of θ is given by the posterior mean (or mode) of θ, which can be obtained by minimizing the quantity $Q = (\tilde{y} - A\theta)'(\tilde{y} - A\theta) + \frac{\sigma^2}{\tau^2}\theta' V^{-1}\theta$. If we let $\lambda = \frac{\sigma^2}{\tau^2}$, then $Q = S(f)$, the penalized sum of squares. So, minimizing Q is the same as minimizing $S(f)$ meaning that the Bayesian model (3.7) is the exact representation of the roughness penalty approach of curve fitting. For our model the smoothing parameter is $\lambda = \frac{\sigma^2}{\tau^2}$. Since $\theta \sim N(0, \tau^2 V)$, τ^2 determines the variability among θs. Smaller τ^2 corresponds to a stronger penalty and hence attempts to produce a smoother curve, whereas larger τ^2 attempts to produce an wiggly curve. From model (3.7) it is observed that in Bayesian paradigm we express the roughness penalty approach as hierarchical model and we take the smoothing parameter λ as a function of the random effects variance component τ^2. Since, in Bayesian approach degree of smoothing in curve estimation is controlled by the random effects variance component τ^2 we can use the conservative prior for τ^2 so that the chance of having an estimated τ^2 greater than the true τ^2 is low. Thus we can have an estimated curve which is not more wiggly than the true underlying curve.

For the full Bayesian treatment of the roughness penalty approach we need to specify priors for σ^2 and τ^2. There are many possible choices for the priors of random effect variance component τ^2. Barry (1995) considered Jeffreys' prior jointly for τ^2 and σ^2. Daniels (1999) considered uniform shrinkage prior for the covariance matrix of random effects in normal–normal hierarchical model. Though, in his study, Daniels focused on the simple normal–normal hierarchical model it is possible to construct an intuitive version of uniform shrinkage prior for smoothing problem represented by model (3.7). In any of the above cases, there was no or little attempts to guard against undersmoothing while estimating the random effects, θ, though it is desirable for most of the applications. In this study our goal is to suggest a prior for the random effect variance component τ^2 which can guard against undersmoothing. That is, we would like a prior for τ^2

which encourages smaller τ^2 and therefore smoother $f(x)$. In other words, we do not want $\hat{f}(x)$ to be more wiggly than the real $f(x)$. In this chapter we have conducted simulation studies to estimate the function $f(x)$ by Bayesian roughness penalty approach using the proposed conservative prior for τ^2. Also, besides using the suggested conservative prior we have used the uniform shrinkage prior and the Jeffreys' prior for the random effects variance component τ^2 and have compared the results.

3.2.1 The Conservative Prior for the Random Effects Variance Component τ^2 in Roughness Penalty Approach

The goal of this study is to ensure the smoothness of the estimated curve in non–parametric regression. In simple normal–normal hierarchical model simulation studies suggest that the proposed conservative prior can achieve this goal. The conservative prior suggested for normal–normal hierarchical model can be extended for the smoothing problem. In the case of smoothing we need to deal with the multivariate problem instead of univariate one, and we have to make adjustment for the dimension of the covariance matrix $\tau^2 V$ in defining the conservative prior for τ^2. In smoothing problem we have

$$\begin{aligned} \tilde{y}|\theta,\sigma^2 &\sim N(A\theta,\sigma^2 I_n), \\ \theta|\tau^2 &\sim N_p(0,\tau^2 V). \end{aligned}$$

The conservative prior for τ^2 can then be defined as

$$\pi(\tau^2|\sigma^2) \propto \frac{1}{|\sigma^2 I_p + \tau^2 V|^{\frac{a+1}{p}}}. \tag{3.8}$$

The denominator p in the exponent has been taken to adjust for the dimension of the parameter vector θ to make the prior comparable to the same one in the univariate case (simple normal–normal hierarchical model). So, the full Bayesian model for the smoothing problem (3.4) can be written as

$$\begin{aligned} \tilde{y}|\theta,\sigma^2 &\sim N(A\theta,\sigma^2 I_n) \\ \theta|\tau^2 &\sim N(0,\tau^2 V) \\ \pi(\tau^2|\sigma^2) &\propto \frac{1}{|\sigma^2 I_p + \tau^2 V|^{\frac{a+1}{p}}} \\ \pi(\sigma^2) &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}+1} e^{-\frac{\frac{1}{2}}{\sigma^2}} \end{aligned} \tag{3.9}$$

where $\pi(\sigma^2)$ is the prior for σ^2, which has been taken to be a unit information inverse gamma distribution centered at one.

3.2.2 Posterior Distributions for the Model Parameters Under Conservative Prior

For the posterior inference of the model parameters we need the conditional or marginal posterior distributions of the model parameters θ, σ^2 and τ^2. The joint posterior distribution of θ, σ^2 and τ^2 is obtained as

$$\begin{aligned}\pi(\theta,\sigma^2,\tau^2|\tilde{y}) &\propto f(\tilde{y}|\theta,\sigma^2)\pi(\theta|\tau^2)\pi(\sigma^2)\pi(\tau^2|\sigma^2)\\ &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}}e^{-\frac{1}{2\sigma^2}(\tilde{y}-A\theta)'(\tilde{y}-A\theta)}\frac{1}{(\tau^2)^{\frac{p}{2}}}e^{-\frac{1}{2\tau^2}\theta' V^{-1}\theta}\frac{1}{|\sigma^2 I_p+\tau^2 V|^{\frac{a+1}{p}}}\\ &\quad \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}+1}e^{-\frac{\frac{1}{2}}{\sigma^2}}\frac{1}{c(\sigma^2)} \qquad (3.10)\end{aligned}$$

where, $c(\sigma^2)$ is the normalizing constant for $\pi(\tau^2|\sigma^2)$ which is needed to obtain the joint prior as $\pi(\tau^2,\sigma^2)=\pi(\tau^2|\sigma^2)\pi(\sigma^2)$ and hence to make the joint posterior (3.10) a proper probability distribution. Propriety of the joint posterior distribution of the model parameters is essential for valid inferences about them. The normalizing constant $c(\sigma^2)$ is obtained as

$$\begin{aligned}c(\sigma^2) &= \int_0^\infty \frac{1}{|\sigma^2 I_p+\tau^2 V|^{\frac{a+1}{p}}}d\tau^2\\ &= \int_0^\infty \frac{1}{(\sigma^2)^{a+1}}\frac{1}{|I_p+\frac{\tau^2}{\sigma^2}V|^{\frac{a+1}{p}}}d\tau^2 \qquad (3.11)\end{aligned}$$

By making the substitution $u=\frac{\tau^2}{\sigma^2}$, we have $\tau^2=\sigma^2 u \Rightarrow d\tau^2=\sigma^2 du$. Substituting these facts in the right hand side of the expression (3.11) we have

$$\begin{aligned}c(\sigma^2) &= \int_0^\infty \frac{1}{(\sigma^2)^{a+1}}\frac{1}{|I_p+uV|^{\frac{a+1}{p}}}\sigma^2 du\\ &= \frac{1}{(\sigma^2)^a}\int_0^\infty \frac{1}{|I_p+uV|^{\frac{a+1}{p}}}du\\ &\propto \frac{1}{(\sigma^2)^a}\end{aligned}$$

Now, from equation(3.10), the conditional posterior distribution of θ, given σ^2, τ^2 and the data, can be derived as

$$\begin{aligned}\pi(\theta|\sigma^2,\tau^2,\tilde{y}) &\propto e^{-\frac{1}{2\sigma^2}(\tilde{y}-A\theta)'(\tilde{y}-A\theta)}e^{-\frac{1}{2\tau^2}\theta'V^{-1}\theta} \\ &= e^{-\frac{1}{2\sigma^2}\left\{(\tilde{y}-A\theta)'(\tilde{y}-A\theta)+\frac{\sigma^2}{\tau^2}\theta'V^{-1}\theta\right\}} \\ &= e^{-\frac{1}{2\sigma^2}\left\{(\tilde{y}-A\theta)'(\tilde{y}-A\theta)+\lambda\theta'V^{-1}\theta\right\}} \\ &= e^{-\frac{1}{2\sigma^2}Q}\end{aligned}$$

where, $\lambda = \frac{\sigma^2}{\tau^2}$ and $Q = (\tilde{y}-A\theta)'(\tilde{y}-A\theta)+\lambda\theta'V^{-1}\theta$. The quadratic form Q can be decomposed as

$$\begin{aligned}Q &= \tilde{y}'[I-H(\lambda)]\tilde{y}+[\theta-G(\lambda)\tilde{y}]'(A'A+\lambda V^{-1})[\theta-G(\lambda)\tilde{y}] \\ &= W(\lambda)+U(\theta)\end{aligned}$$

where, $H(\lambda) = A(A'A+\lambda V^{-1})^{-1}A'$ and $G(\lambda) = (A'A+\lambda V^{-1})^{-1}A'$. Therefore, the conditional posterior distribution of θ, given σ^2, τ^2 and the data, is

$$\pi(\theta|\tilde{y}) \propto e^{-\frac{1}{2\sigma^2}[\theta-G(\lambda)\tilde{y}]'(A'A+\lambda V^{-1})[\theta-G(\lambda)\tilde{y}]} \tag{3.12}$$

From equation (3.12) it can easily be verified that the posterior distribution of θ is normal with mean vector $\hat{\theta} = (A'A+\lambda V^{-1})^{-1}A'\tilde{y}$ and covariance matrix $(A'A+\lambda V^{-1})^{-1}\sigma^2$, i.e.,

$$\theta|\sigma^2,\tau^2,\tilde{y} \sim N[G(\lambda)\tilde{y},(A'A+\lambda V^{-1})^{-1}\sigma^2]$$

The conditional posterior distribution of σ^2, given θ, τ^2 and the data, is

$$\begin{aligned}\pi(\sigma^2|\lambda,\tilde{y}) &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}}e^{-\frac{1}{2\sigma^2}(\tilde{y}-A\theta)'(\tilde{y}-A\theta)}\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}+1}e^{-\frac{\frac{1}{2}}{\sigma^2}}\frac{1}{|\sigma^2I_p+\tau^2V|^{\frac{a+1}{p}}}\frac{1}{c(\sigma^2)} \\ &= \left(\frac{1}{\sigma^2}\right)^{\frac{n+1}{2}+1}e^{-\frac{1}{2\sigma^2}\{(\tilde{y}-A\theta)'(\tilde{y}-A\theta)+1\}}\frac{1}{|\sigma^2I_p+\tau^2V|^{\frac{a+1}{p}}}\frac{1}{c(\sigma^2)}\end{aligned} \tag{3.13}$$

Finally, the conditional posterior distribution of τ^2, given θ, σ^2 and the data, is

$$\pi(\tau^2|\lambda,\tilde{y}) \propto \frac{1}{(\tau^2)^{\frac{p}{2}}}e^{-\frac{1}{2\tau^2}\theta'V^{-1}\theta}\frac{1}{|\sigma^2I_p+\tau^2V|^{\frac{a+1}{p}}} \tag{3.14}$$

From equations (3.12), (3.13) and (3.14) it is observed that the conditional posterior distribution of θ has closed form but those of σ^2 and τ^2 do not have closed form. So, we can use Gibbs sampler for posterior simulation of θ, but for σ^2 and τ^2 we need to use the Metropolis–Hastings algorithm to draw posterior simulations.

3.3 Performance of Conservative Prior in Smoothing Problem: Simulation Studies

As in the case of simple normal–normal hierarchical model simulation studies have been performed in smoothing problem also to demonstrate the performance of conservative prior in this area. For simulation studies we have considered the following situations:

(i) simulation studies with large sample size ($n = 111$)

(ii) simulation studies with small sample size ($n = 23$)

For each of the above two cases we have considered two different values of p (the dimension of mean vector θ) as:

(a) $p = 10$

(b) $p = 5$

For $n = 111$ and $p = 10$ the design variable has been taken to be $x = \frac{i}{10}$; $i = 0, 1, \ldots, 110$ and the equidistant knots have been taken at $x = 0$, $x = 1$, $x = 2$, ..., $x = 11$. For $n = 111$ and $p = 5$ the design variable x has been defined to be $x = \frac{1.833333i}{18.4}$; $i = 0, 1, \ldots, 111$ and the equidistant knots have been taken at $x = 1.833333 \times k$; $k = 0, 1, \ldots, 6$.

For each of the above 4 cases data have been generated as:

(i) we have fixed the values of the unknown parameters σ^2 and τ^2 ($\sigma^2 = 1$, τ^2=1)

(ii) given the value of τ^2, we have generated 100 different θ vectors as $\theta \sim N(0, \tau^2 V)$

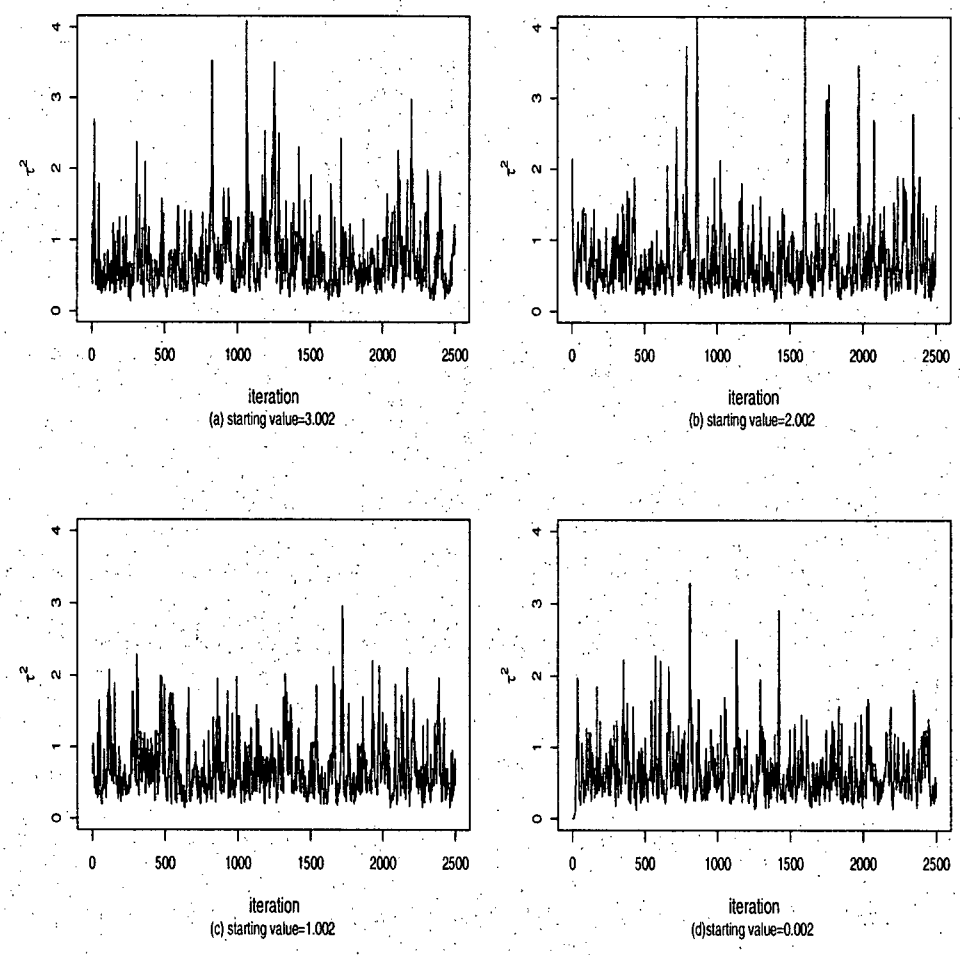
(iii) for each of the generated θ vectors and given σ^2 value we have generated a data set ($\tilde{y}$ vector) as $\tilde{y} \sim N(A\theta, \sigma^2 I_n)$.

From the generated data the model parameters θ, σ^2 and τ^2 have been estimated by using the Bayesian technique that makes the use of the proposed conservative prior with a choice of $a = 5$. For parameter estimates we have run MCMC simulation techniques. We have used Gibbs sampler for posterior simulation of θ and random–walk Metropolis–Hastings algorithm for σ^2 and τ^2.

3.3.1 Monitoring the Convergence of MCMC Simulation

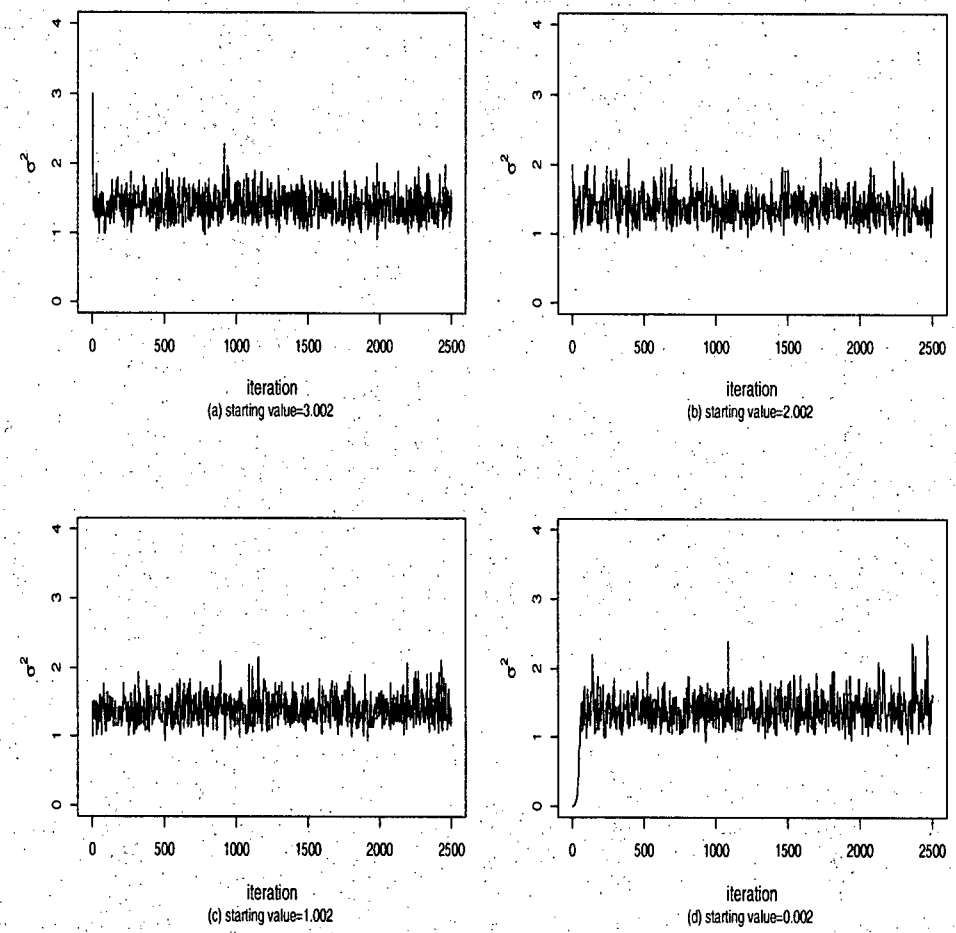
For monitoring the convergence and mixing of the MCMC chains for both σ^2 and τ^2 we have plotted four different chains for each of them. For updating σ^2 and τ^2 we have used exponential scale. Figures 3.1 and 3.2 display four different MCMC chains for each of σ^2 and τ^2, respectively, obtained by using the same data set. For different chains of each parameter widely dispersed initial values have been used. From Figures 3.1 and 3.2 we observe that all the chins of each

Figure 3.1: Plots of MCMC chains for τ^2 with different starting points



of the parameters exhibit good mixing since none of the chains of any of the parameters kept moving slowly around any particular region of the target distribution for many iterations. It is also observed that though different chains for each of the parameters starts at different initial

Figure 3.2: Plots of MCMC chains for σ^2 with different starting points



values all the chains for both σ^2 and τ^2 have stabilized to very close to the respective true values after very few iterations. In all the cases the acceptance rates are found to be between 45% to 50%.

3.4 Output Analysis

Since the main goal of the proposed conservative prior is to control the undersmoothing, i.e., to control the over estimation of the random effects variance component τ^2, in this study

we have estimated the same true τ^2 value by using 100 different data sets, which have been simulated from the distribution with the same true σ^2 and τ^2 values ($\tau^2 = 1, \sigma^2 = 1$). Details of data simulations had been elaborated earlier in section 3.3. For estimation of parameters by Bayesian approach we have considered the proposed conservative prior. We have adopted MCMC simulation to generate draws from the respective conditional posterior distributions. By monitoring the output for several independent chain for each of the model parameters we have found that for both τ^2 and σ^2 the MCMC chains converge after very few iterations. So, for posterior inferences about the model parameters each time we have run the MCMC chain for each of the parameters for 1500 iterations. Finally, we have thrown away first 500 iterations as burn–in of the chain and used the remaining 1000 iterations for inferential purposes. To estimate σ^2 and τ^2 we have used both the posterior mean and the posterior mode. For θ posterior mean and mode are the same because the posterior distribution of θ,given σ^2, τ^2 and the data, is multivariate normal. Figures 3.3 and 3.4 display the histograms of 100 estimates of the same τ^2 value (true $\tau^2 = 1$) obtained by using 100 different data sets for large sample size ($n = 111$) and for 10 interior knots ($p = 10$). Estimates in Figure 3.3 are based on posterior mean and those in Figure 3.4 are based on posterior mode.

From Figures 3.3 and 3.4 we observe that for both cases not too many estimates of τ^2 exceed the true τ^2 value. However, in the case of posterior mean the number of τ^2 estimates that exceed the true τ^2 value is slightly more than in the case of posterior mode (20 in case of mean versus 9 in case of mode). This is reasonable because mode is insensitive to any possible outliers. However, deciding about the value of the hyperparameter a on the basis of posterior mode may not be sufficient for guarding against undersmoothing, especially when τ^2 is relatively smaller than σ^2. The disadvantages of using the posterior mode in the context of undersmoothing have already been discussed in section 1.6 of chapter 1.

Regarding σ^2, it is observed that estimates are very close to the true value and the estimates do not differ too much whether we use the posterior mean or mode. This is because the posterior distribution of σ^2 is not so positively skewed as that of τ^2. This picture will be more clear if we look at the Figures 3.1 and 3.2. From each plot of Figure 3.1 it is observed that there are some big upward jumps in the posterior simulations of τ^2 indicating highly positive skewness of the posterior distribution of τ^2, whereas such big upward jumps are absent in each plot of Figure 3.2 of posterior simulations of σ^2. The histograms of the estimates of σ^2 based on posterior mean and mode are displayed in Figures 3.5 and 3.6, respectively.

Figure 3.3: Histogram of estimated τ^2 based on posterior mean

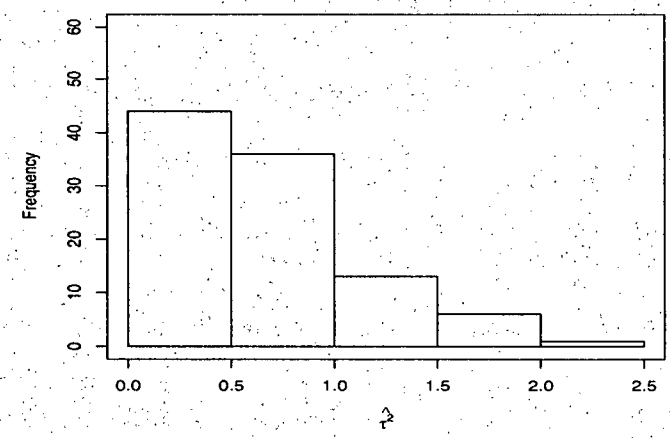


Figure 3.4: Histogram of estimated τ^2 based on posterior mode

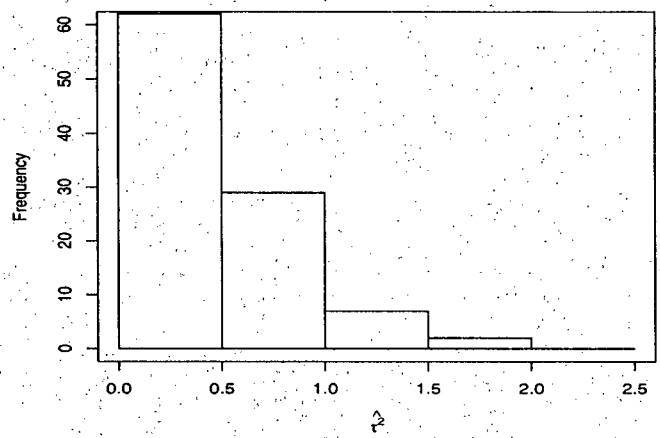


Figure 3.5: Histogram of estimated σ^2 based on posterior mean

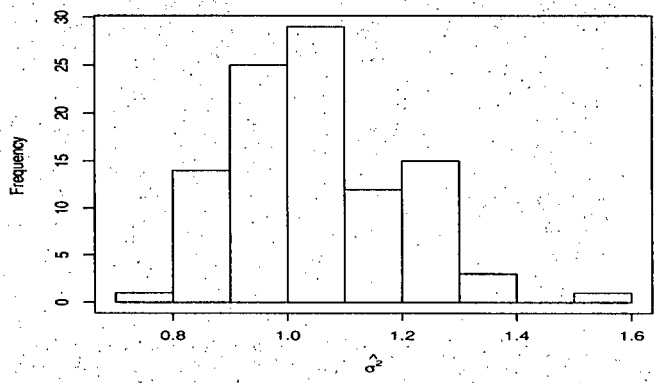
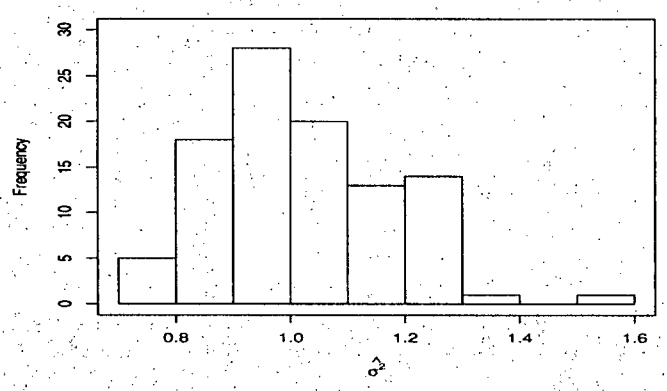


Figure 3.6: Histogram of estimated σ^2 based on posterior mode



From Figures 3.5 and 3.6 it is observed that the spreads of the both histograms are almost same, but the histogram of the estimates of σ^2 based on posterior mean is centered around 1.05, whereas that based on posterior mode is centered around 0.95.

Finally, since the main purpose of smoothing problem is to estimate the curve $f(x)$ itself, which in our study is parameterized by θ vector, we are interested in drawing inferences about θ vector also. By introducing the proposed conservative prior we attempt to estimate the θ vector so that the estimates of the components of θ vector are not more variable than the true values of the components of θ vector. In our simulation study we have generated 100 different θ vectors

and for each θ vector we have generated a data set. Finally, we have used each generated data set to estimate the respective θ. Now, for our estimation problem we have parameterized the true curve $f(x)$ with a lower dimensional vector θ, but in practice it is desirable to estimate the entire curve. The estimate of the entire curve can be obtained by using the method of interpolating natural cubic spline given the curve estimate at the selected knot points. This is obtained as $\hat{f}(x) = A\hat{\theta}$. Figure 3.7 displays the plots of the true data , the true curve as given by $f(x) = A\theta$ and the entire estimated curve for four randomly selected θ vectors out of 100 of them.

From the plots (a), (b) and (c) of Figure 3.7 we observe that the estimated curves are quite smoothed and they track the respective underlying true mean curves very closely. On the other hand, if we look at the panel (d) of Figure 3.7 we observe that true curve is wiggly but estimated curve is smoothed. In this though the estimated curve could not track the true underlying mean curve very closely in some data ranges still it has captured the key features in the data set it represents. Actually, if there are too rapid fluctuation in the data then the true curve may become wiggly but the goal of smoothing problem is to estimate the data in such a way that the estimated curve is a smoothed one and can pick the key pattern in the data set. All the plots of Figure 3.7, especially the last panel, confirms the ability of conservative prior to guard against undersmoothing in estimating the true curve in the smoothing problem.

Again, as it is already known, in hierarchical model and in smoothing problem the key concern on the part of an estimation method is its ability to detect the situation of no heterogeneity. So, to reflect the ability of the conservative prior to detect the situation of no heterogeneity we have plotted the data values, the true underlying mean curve and the estimated curve on the same plane in Figure 3.8 when $\tau^2 = 0.01$. The similar plots for the uniform shrinkage prior and for the Jeffreys' prior for the same data sets used in the case of conservative prior to construct the Figure 3.8 are displayed in Figures 3.9 and 3.10, respectively.

Figure 3.7: Plots of data values, true mean curves and the estimated curves $\hat{f}(x)$ for the selected data sets

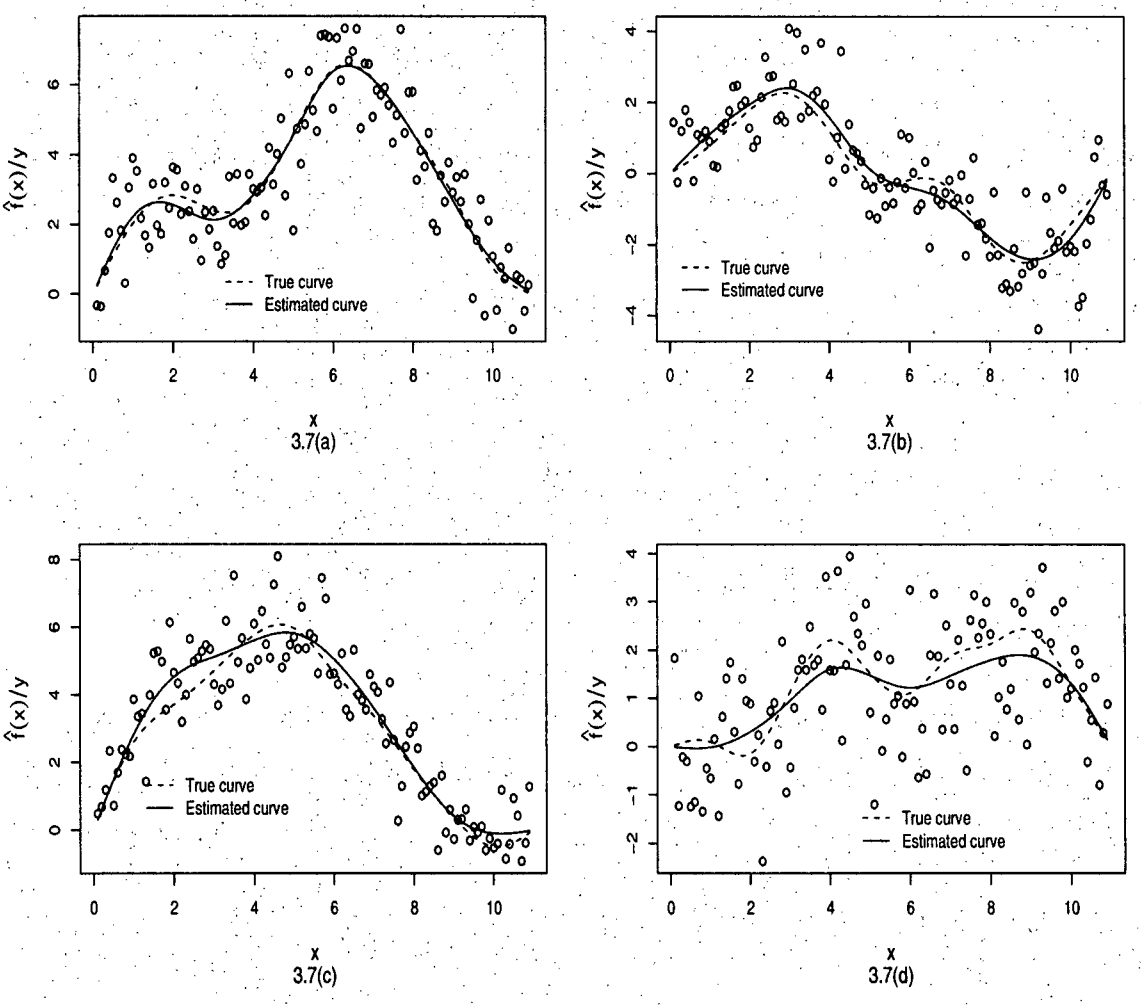


Figure 3.8: Plots of data values, true mean curves and the estimated curves $\hat{f}(x)$ for the selected data sets when $\tau^2 = 0.01$

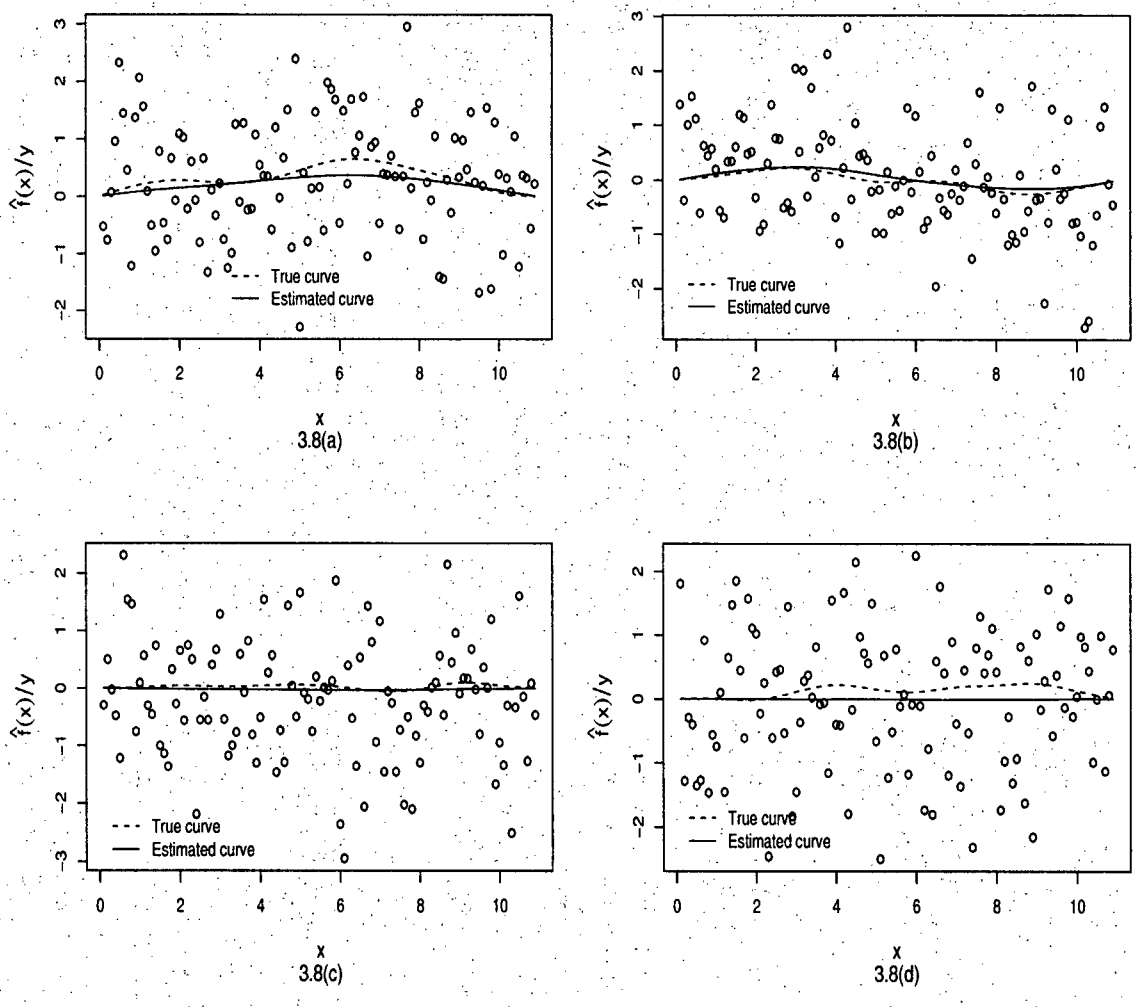


Figure 3.9: Plots of data values, true mean curves and the estimated curve $\hat{f}(x)$ obtained by using uniform shrinkage prior when $\tau^2 = 0.01$

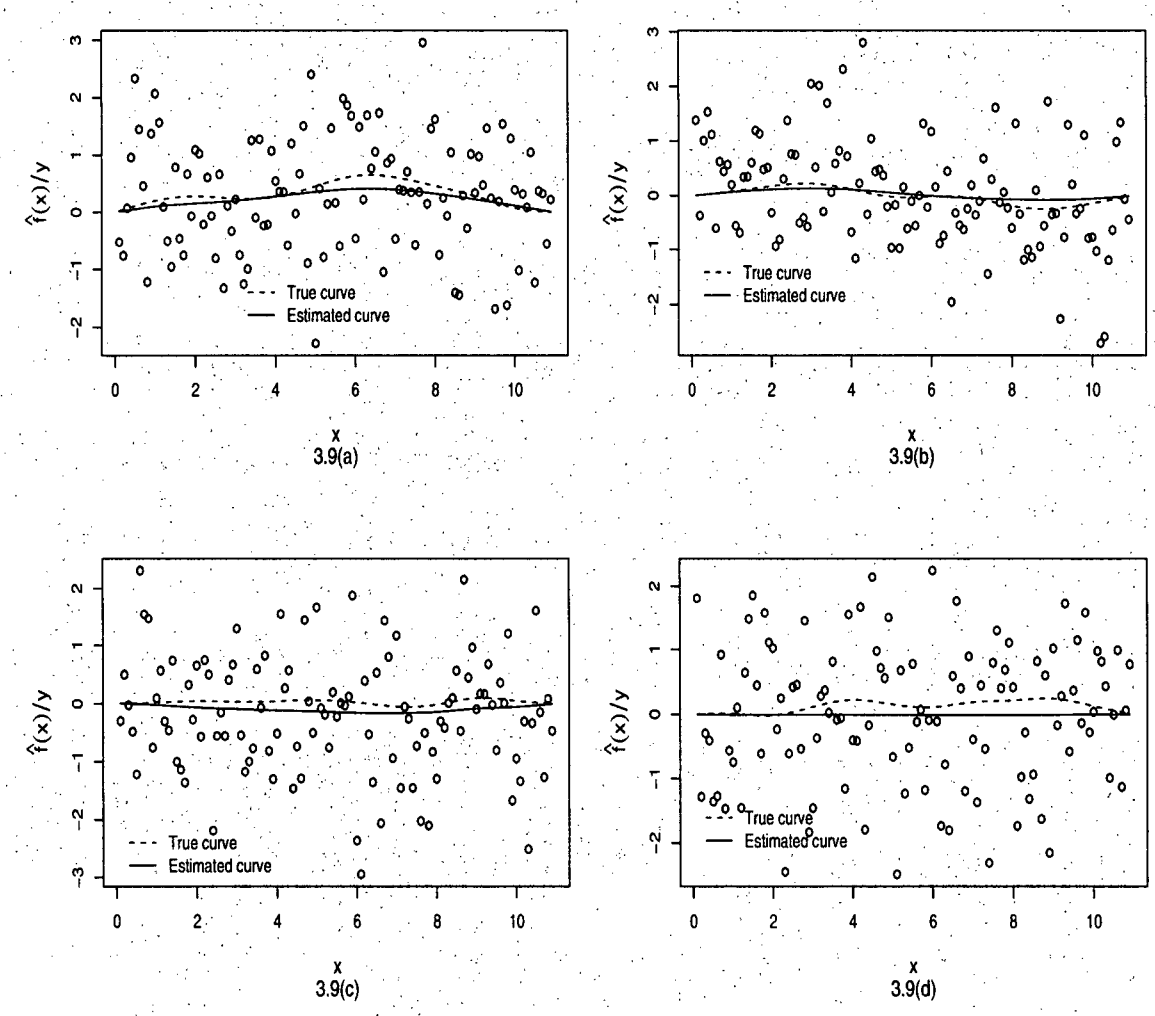
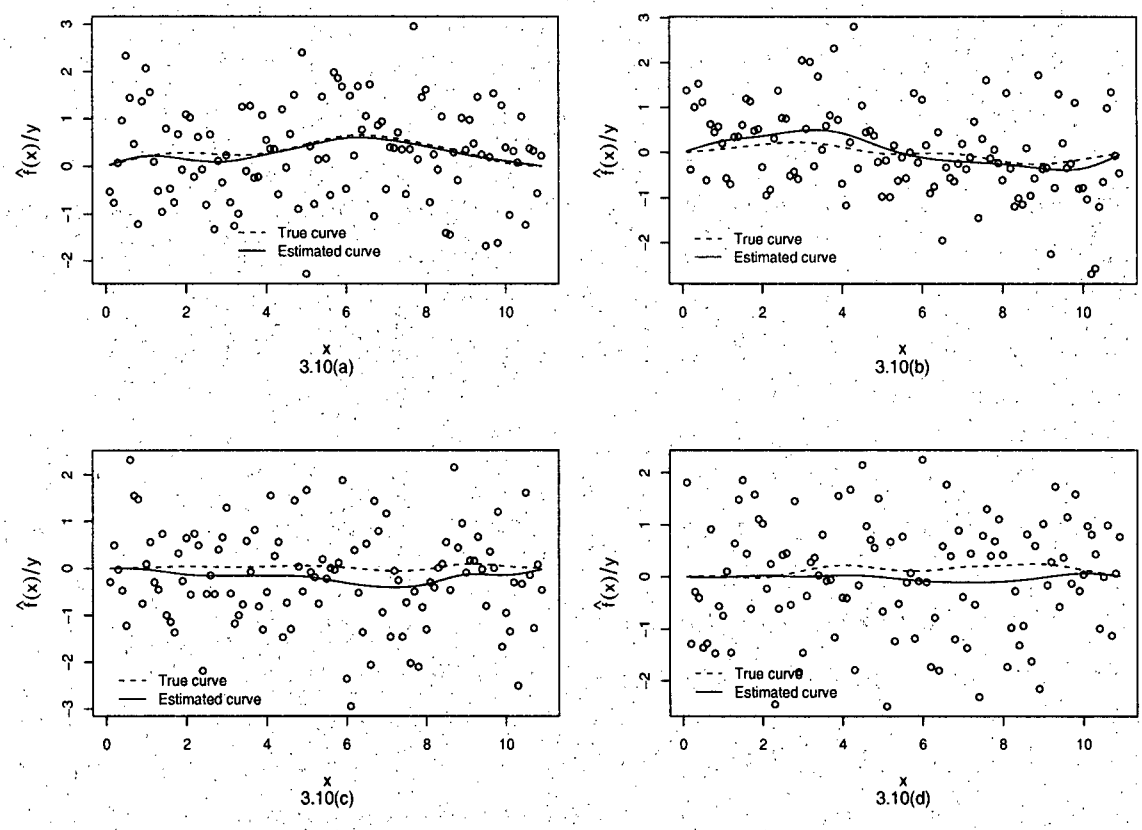
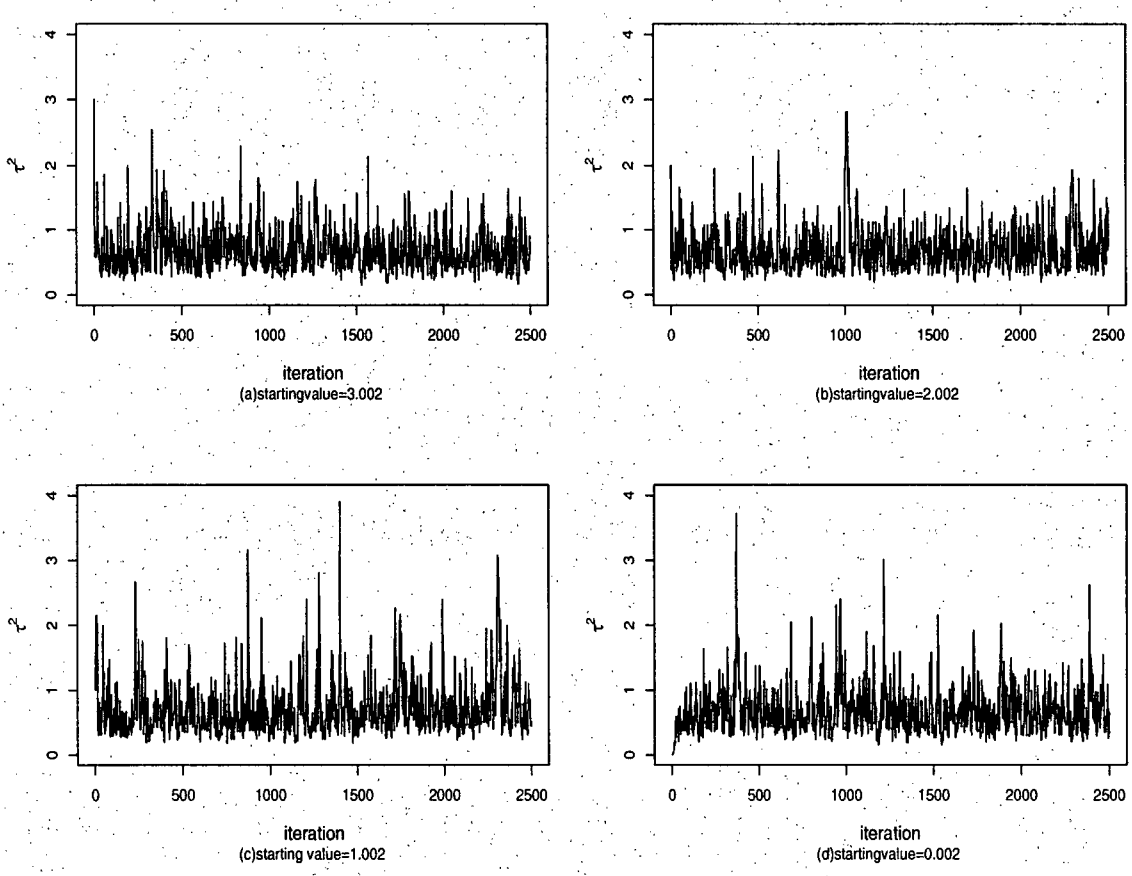


Figure 3.10: Plots of data values, true mean curves and the estimated curves $\hat{f}(x)$ obtained by using Jeffreys' prior when $\tau^2 = 0.01$



The plots of the Figure 3.8 show that all the estimated curves track the respective true underlying curves very closely, and none of them is more wiggly than the corresponding true curve. Also, plots of the original data values show that there is almost no global variation, as it should be for a true τ^2 value close to 0, in the data set and the estimated curves have picked up this feature of the data sets very well. Now, by comparing the plots of Figures 3.9 and 3.10 we see that the conservative prior and the uniform shrinkage prior produce very similar results for 3 out of the 4 selected data sets (panels (a), (b) and (d) of both the Figures). But for the data set corresponding to the panel (c) the estimated curve in the case of uniform shrinkage prior is not so close to the true curve as it is in the case of conservative prior (plot 3.8(c) versus plot 3.9(c)). For this particular data set the estimated curve seems to be more wiggly than the true curve in the case of uniform shrinkage prior, whereas the conservative prior has produced a smoothed estimated curve which also tracks the true curve very closely. The picture is worse for 3 out of 4 data sets when Jeffreys' prior is used. Plots 3.10(b), 3.10(c) and 3.10(d) show that the estimated curves are far more wiggly than the corresponding true curves. So, we can

Figure 3.11: Plots of MCMC chains for τ^2 with different starting points when p=5

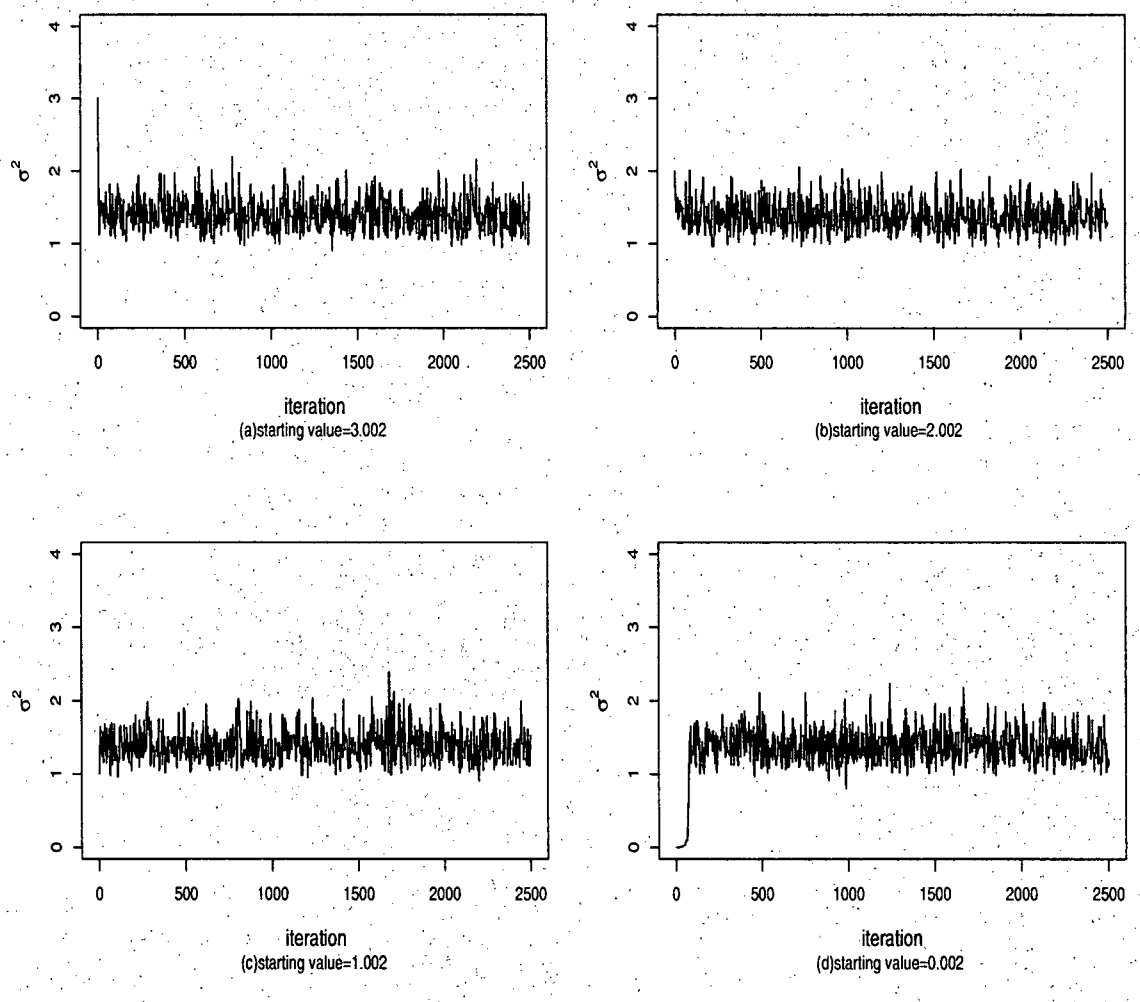


conclude that if undersmoothing can not be reasonably controlled we may get too wiggly and less accurate estimated curves.

To check whether the use of conservative prior produces similar results in different situations we have also analyzed the simulated data for large sample size by specifying fewer number of knots (5 instead of 10). Figure 3.11 displays the convergence plots of MCMC chains for τ^2 with different starting values.

Different plots of Figure 3.11 reveal that though different chains started with different initial values all the chains have stabilized at the same level after very few iterations. Also mixing of MCMC chains looks very good. Similar conclusions can be drawn about the convergence and mixing of the chains for σ^2 (Figure 3.12).

Figure 3.12: Plots of MCMC chains for σ^2 with different starting points when p=5



Regarding the performance of conservative prior to control the undersmoothing in estimating θ vector in case of 5 knots we can conclude that the proposed prior can perform equally well as it does in case of 10 knots. Figures 3.13 and 3.14 confirm this fact.

Figure 3.13: Histogram of estimated τ^2 based on posterior mean while using fewer knots (p=5)

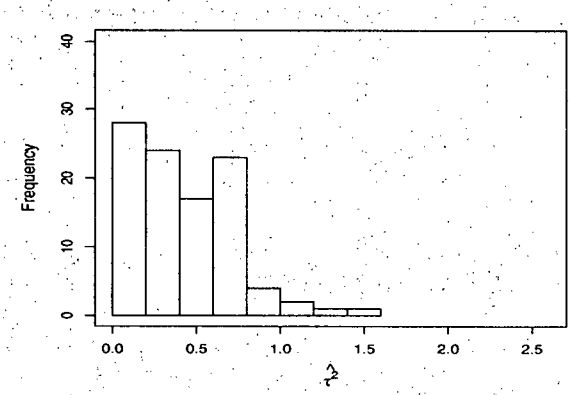
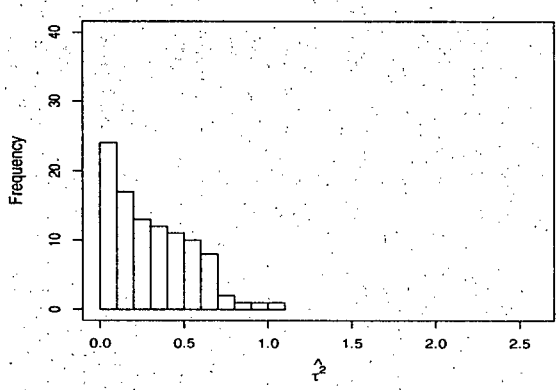


Figure 3.14: Histogram of estimated τ^2 based on posterior mode while using fewer knots (p=5)



If we look at the histogram of $\hat{\tau}^2$ based on the posterior mean (Figure 3.13) and that based on posterior mode (Figure 3.14) we observe that in both cases $\hat{\tau}^2$ is centered around 0.5, which is very much similar to the case when we used 10 knots instead of 5 knots. About the estimates of σ^2 similar conclusions can be drawn as in the case of using 10 knots except that for both posterior mean and posterior mode the histograms of $\hat{\sigma}^2$ are centered around 1.05 instead of

1.05 in case of posterior mean and 0.95 in case of posterior mode when we used 10 knots (Figures 3.15 and 3.16).

Figure 3.15: Histogram of estimated σ^2 based on posterior mean when p=5

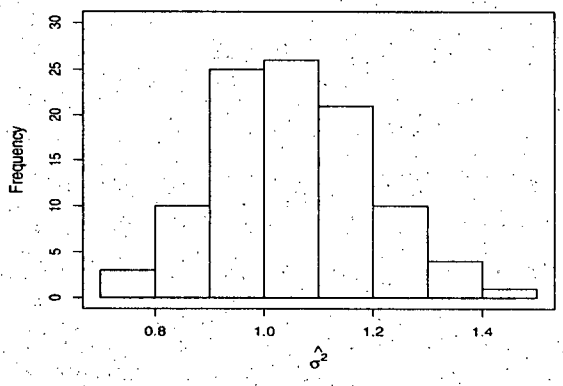
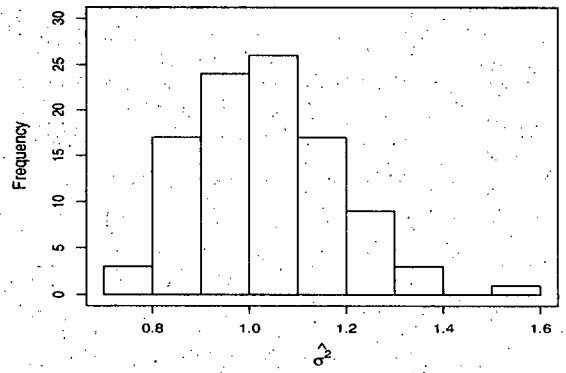


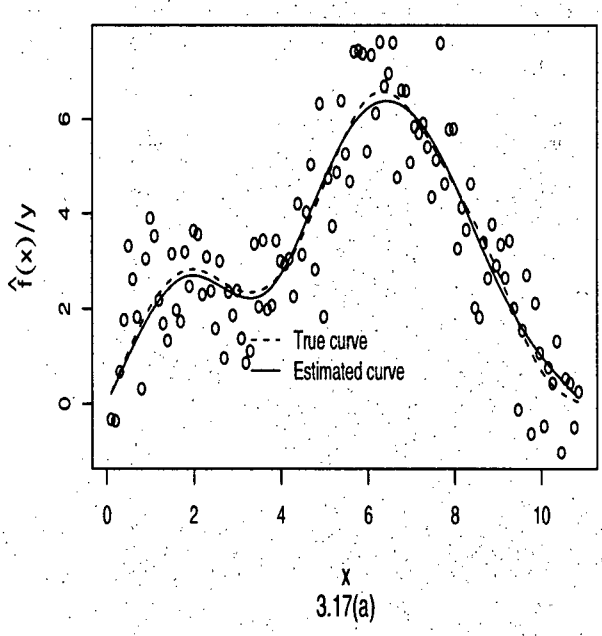
Figure 3.16: Histogram of estimated σ^2 based on posterior mode when p=5



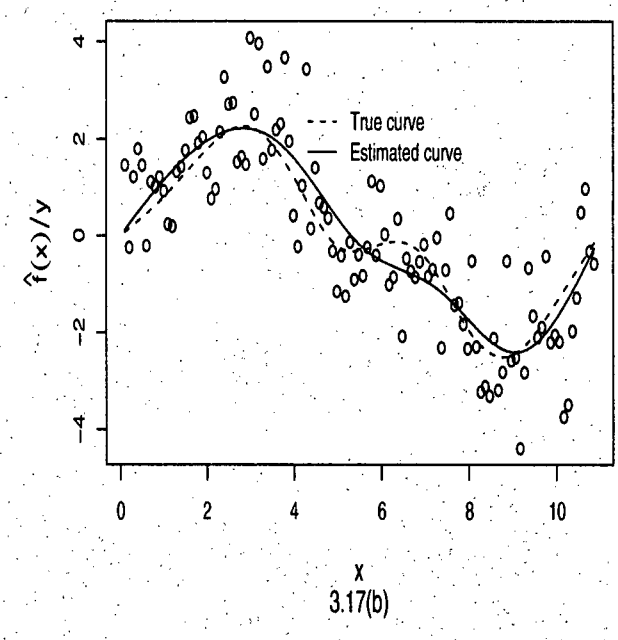
So, in general we can conclude that the proposed conservative prior performs quite well, irrespective of the number of knots used, in controlling the undersmoothing while estimating the true regression curve by Bayesian roughness penalty approach.

Finally, to see how the proposed conservative prior performs in estimating the true data in the case of using fewer knots we have plotted the true data values and the estimated curves on the same plane for the same data sets we have used in the case of 10 knots to construct the Figure 3.7 in Figure 3.17

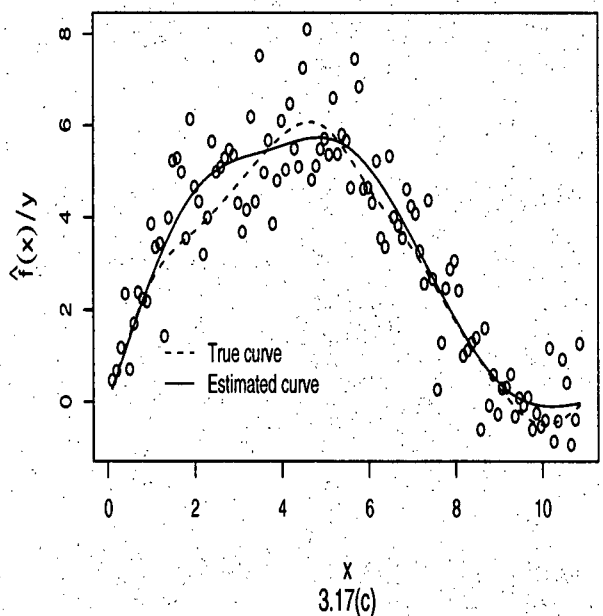
Figure 3.17: Plots of true data, true curves and the estimated curve $\hat{f}(x)$ when p=5



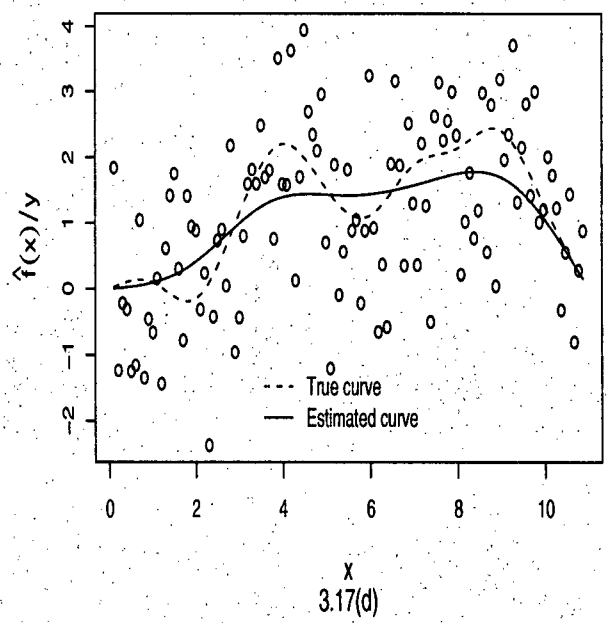
3.17(a)



3.17(b)



3.17(c)



3.17(d)

The plots in Figure 3.17 reveal that the estimated curves are smoother than those in Figure 3.7. Especially, if we look at the plots (a) and (d) of Figure 3.17 we observe that the estimated curves can not track the true curves as closely as they did in Figure 3.7. So, it can concluded that using too few knots may produce too smooth curves which can not pick up the key data feature as well as the estimated curves obtained by using sufficiently large number of knots can do.

3.4.1 Performance Analysis of Conservative Prior in Smoothing in case of Small Sample size ($n = 23$)

As mentioned earlier in Section 3.3, to give strong footing to our conclusion about the performance of the proposed conservative prior in smoothing problem, we have performed simulation studies for small sample size also. In small sample size case we have considered a sample size of $n = 23$. For $n = 23$, the two different cases are:

(i) p=10, i.e., we have considered 10 interior knots

(ii) p=5, i.e., we have considered 5 interior knots

For case (i) the design values have been taken to be $x = \frac{i}{2};\ i = 0, 1, \ldots, 22$ and the equidistant knots are taken at $x = 0, x = 1, \ldots, x = 11$. For case (ii) values of x are taken to be $x = \frac{1.833333i}{3.68};\ i = 0, 1, \ldots, 21$ and $x = 11$, and the knots are taken at $x = 0 \times 1.833333, x = 1 \times 1.833333, x = 2 \times 1.83333 \ldots, x = 10 \times 1.833333 \simeq 11$.

For $n = 23$ and $p = 10$ the convergence plots of 4 independent MCMC chains for τ^2, when true τ^2 value is 1, generated from the same data set with 4 different starting values are displayed in Figure 3.18.

Figure 3.18 shows that all the chains for τ^2 converge to the same level after few iterations and all the chains exhibit good mixing. Note that the jump size in this case is 0.80 for τ^2, and for this jump size the acceptance rates have been found to be between 48% to 50%.

For σ^2 the convergence plots are displayed in Figure 3.19. Similar conclusions can be drawn about the convergence of the chains for σ^2 as was drawn about the convergence of the chains for τ^2. The jump size for updating σ^2 is 0.60, and the acceptance rates have been found to be between 48% to 50%.

Figure 3.18: Plots of MCMC chains for τ^2 with different starting points when n=23 and p=10

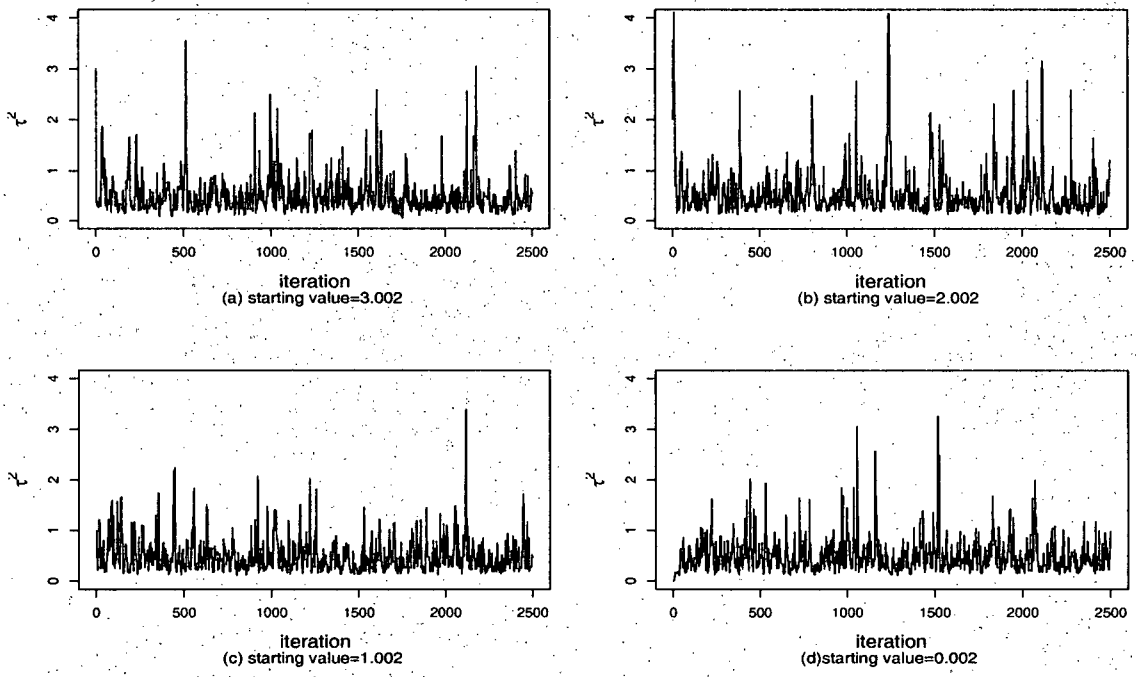
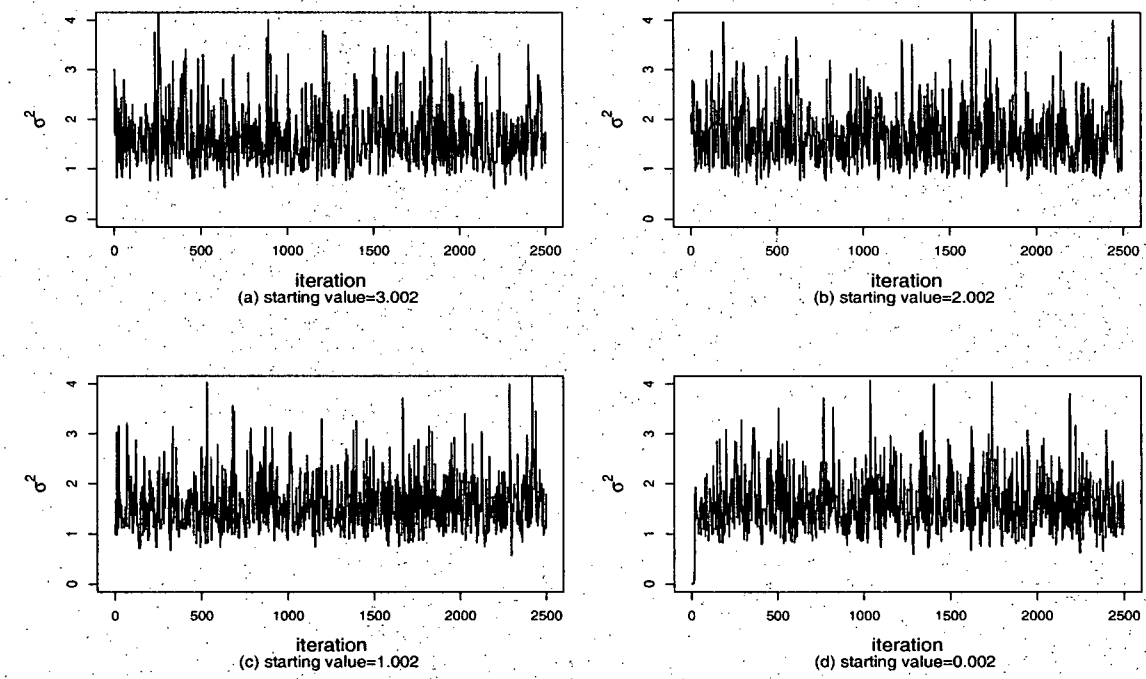


Figure 3.19: Plots of MCMC chains for σ^2 with different starting points when n=23 and p=10



So, finally for drawing posterior inference about τ^2 and σ^2 we have run each chain of both the parameters for 1500 iterations of whom first 500 iterations have been discarded as burn–in and the remaining 1000 have been used for inference purposes. Figures 3.20 and 3.21 display the histograms for the estimates of τ^2 based on posterior mean and mode respectively, when true τ^2 value is 1, obtained by using 100 different simulated data sets.

Figure 3.20: Histogram of estimated τ^2 based on posterior mean for n=23 and p=10

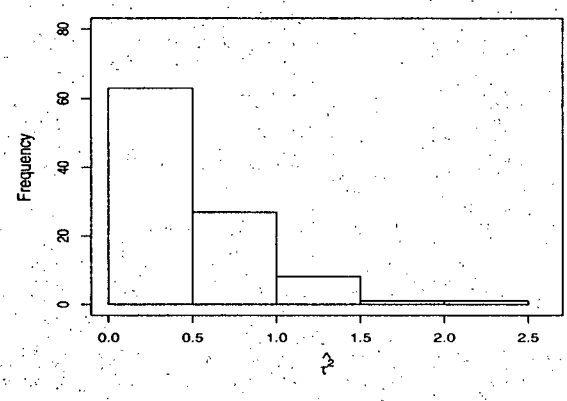
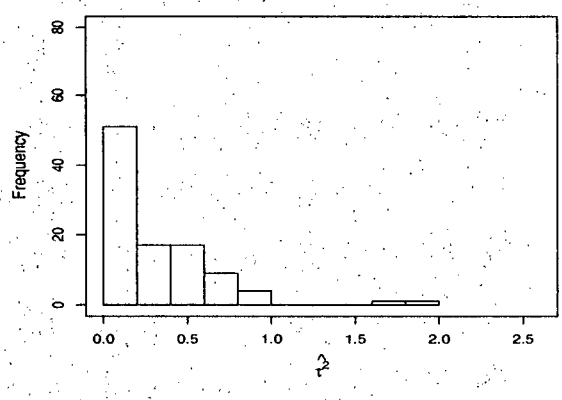


Figure 3.21: Histogram of estimated τ^2 based on posterior mode for n=23 and p=10



From Figures 3.20 and 3.21 we observe the similar picture as we did in case of large sample size.

Again, for $n = 23$ and $p = 5$ the corresponding histograms are displayed in Figures 3.22 and 3.23.

Figure 3.22: Histogram of estimated τ^2 based on posterior mean for n=23 and p=5

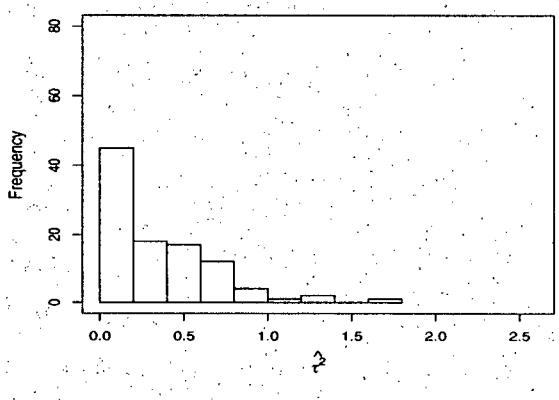
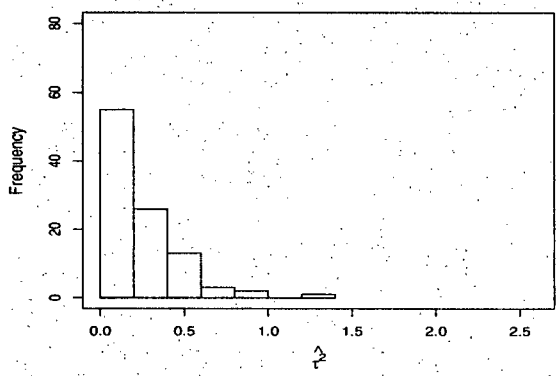


Figure 3.23: Histogram of estimated τ^2 based on posterior mode for n=23 and p=5



From Figures 3.22 and 3.23 we see that in this case also the conservative priors can perform very well in controlling undersmoothing while estimating the true curve in smoothing problem. So, in general we can conclude that the proposed conservative prior performs reasonably well in guarding against undersmoothing in non–parametric regression curve estimation problem.

Finally, to see how the proposed conservative prior performs in estimating the true underlying curve and in capturing the key data feature in the case of small sample size we have plotted the true curve, the data values and the estimated curves on the same plane for the 4 randomly selected data set in Figures 3.24 and 3.25 for $p = 10$ and $p = 5$, respectively.

Figure 3.24: Plots of the true curves, data values and the estimated curves $\hat{f}(x)$ when n=23 and p=10

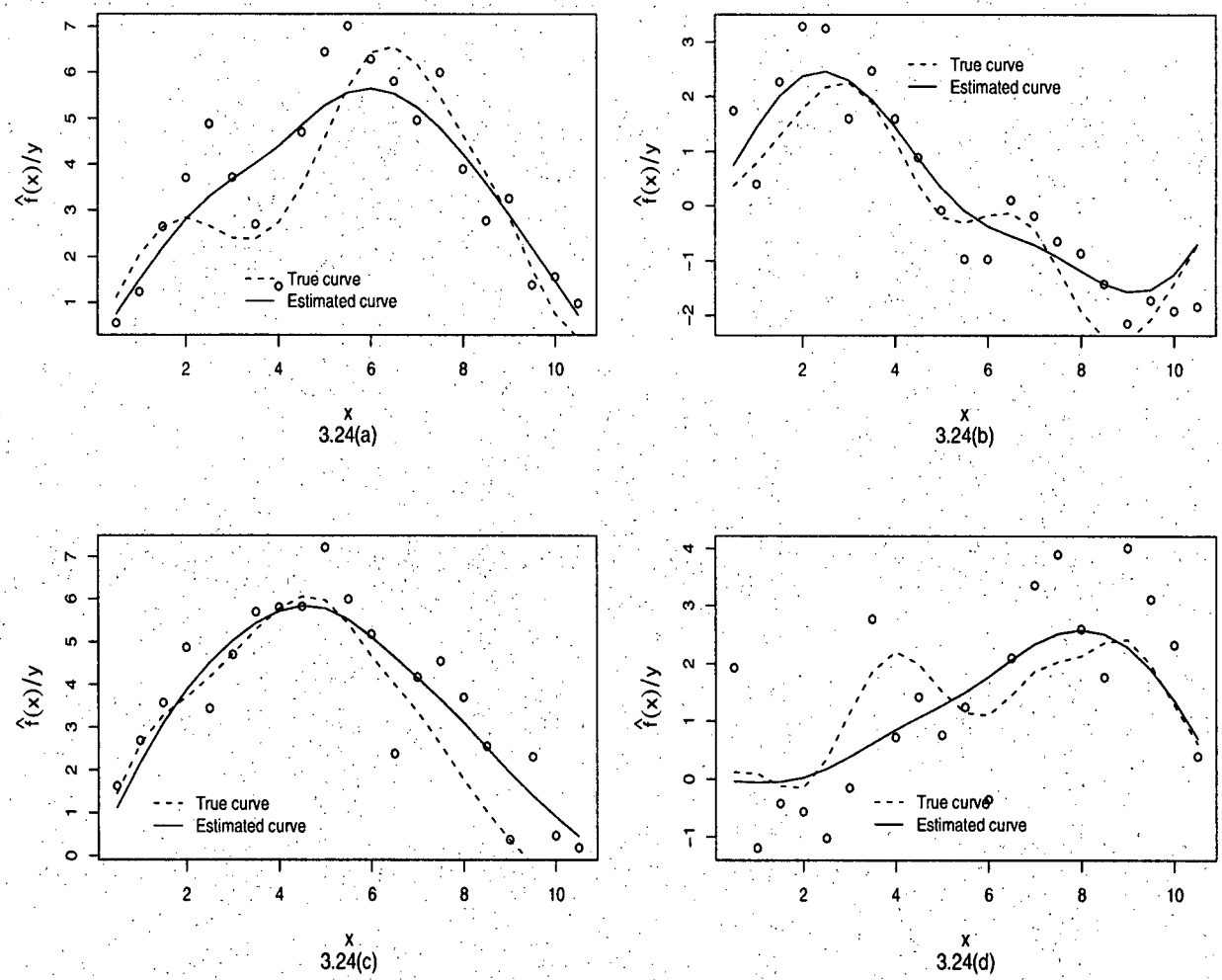
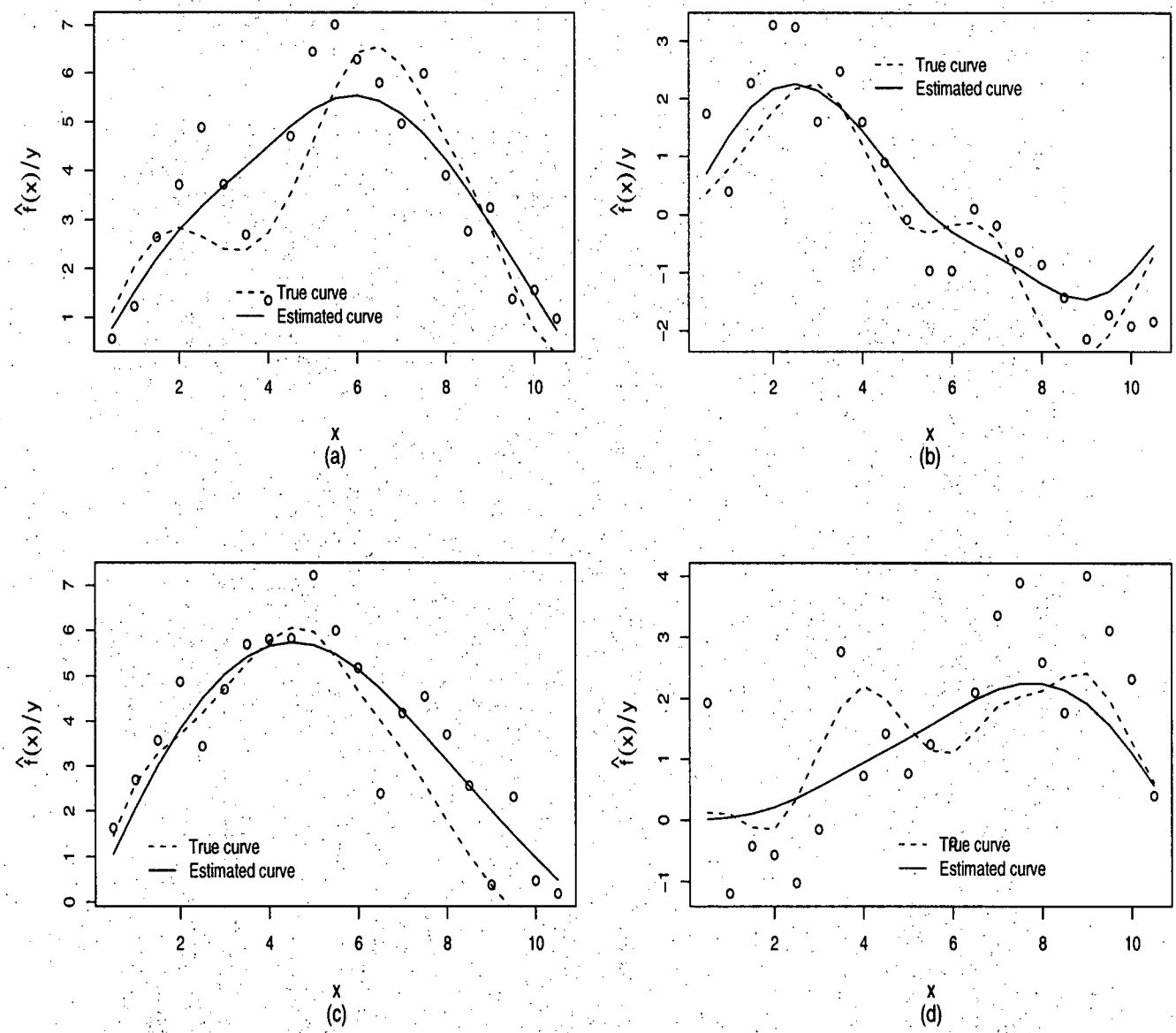


Figure 3.25: Plots of the true curves, data values and the estimated curves $\hat{f}(x)$ when n=23 and p=5



Plots in Figure 3.24 reveal that for the selected data sets the Bayesian roughness penalty approach with the proposed conservative prior seems to perform quite well in estimating the true regression curve. By having a comparative look at the Figures 3.24 and 3.25 we observe the same picture as we did in the case of large sample size. So, in general, we can conclude that irrespective of sample size using too few knots may produce too smooth estimated curves which may not be able to capture the underlying feature of the respective data sets very well as they can do in the case of using sufficient number of knots. Actually, it is always reasonable to use sufficiently large number of knots in smoothing problem.

3.5 Comparison of Results Obtained by Using the Proposed Conservative Prior with those Obtained by Uniform Shrinkage Prior and Jeffreys' Prior

As we mentioned earlier that this study has been aimed at suggesting a conservative prior that can guard against undersmoothing since in many practical situations undersmoothing is considered to be a more serious error than oversmoothing. So, in this section we have studied the comparative performance of conservative prior and that of uniform shrinkage and Jeffreys' priors with respect to the issue of guarding against undersmoothing.

Another important issue in smoothing is that in case of using conservative prior we force the estimates of random effects θ to be close to each other. By forcing the estimates to be closer we can reduce the variability among the estimated θ values but this fact may increase the bias in the estimated curve. So, one obvious question is that how much we gain (or lose) in terms of mean squared error (MSE) by using the conservative prior in comparison with the two competitive priors– the uniform shrinkage prior and the Jeffreys' prior. To check this fact we have compared the estimated MSE of $\hat{\theta}$ for conservative prior with those for the uniform shrinkage and the Jeffreys' priors using simulated data. For smoothing problem, which is a multivariate normal–normal hierarchical model, the uniform shrinkage prior is the special case of conservative prior with $a = 1$, but the derivation of Jeffreys' prior is not so straightforward. However, Barry (1995) derived the Jeffreys' prior for smoothing problem.

3.5.1 Jeffreys' Prior for Smoothing Problem

Barry (1995) derived Jeffreys' prior jointly for σ^2 and τ^2 in smoothing problem. For defining the Jeffreys' prior he considered the smoothing model to be

$$\begin{aligned} y|\theta &\sim N(A\theta, \sigma^2 I_n) \\ \theta|\lambda &\sim N(0, \lambda V) \end{aligned} \tag{3.15}$$

where, $\lambda = \frac{\sigma^2}{\tau^2}$. He found the Jeffreys' prior for σ^2 and τ^2 jointly in smoothing problem to be $\pi(\sigma^2, \lambda) \propto \frac{1}{\sigma^2 \lambda}\sqrt{R(\lambda)}$, where $m = r(V)$, $R(\lambda) = (n+m-p)[tr\{H^2(\lambda)\} + m - p] - [tr\{H(\lambda)\} + m - p]^2$ and $H(\lambda) = A(A'A + \lambda V^{-1})^{-1}A'$. For smoothing model (3.4), $r(V) = p$ and hence $R(\lambda) = n[tr\{H^2(\lambda)\}] - [tr\{H(\lambda)\}]^2$.

Now, it can be easily shown that the conditional posterior distribution of θ, given σ^2, τ^2 and the data, is normal with mean vector $\hat{\theta} = G(\lambda)y$ and covariance matrix $(A'A + \lambda V^{-1})^{-1}\sigma^2$, where $G(\lambda) = (A'A + \lambda V^{-1})^{-1}A'$.

The conditional posterior density of σ^2 given λ is obtained as

$$\pi(\sigma^2|\lambda, y) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}+1}} e^{-\frac{\frac{W}{2}}{\sigma^2}} \sim Inv - gamma\left(\frac{n}{2}, \frac{W}{2}\right)$$

where, $W = y'[I - H(\lambda)]y$.

Finally, the marginal posterior density for λ is

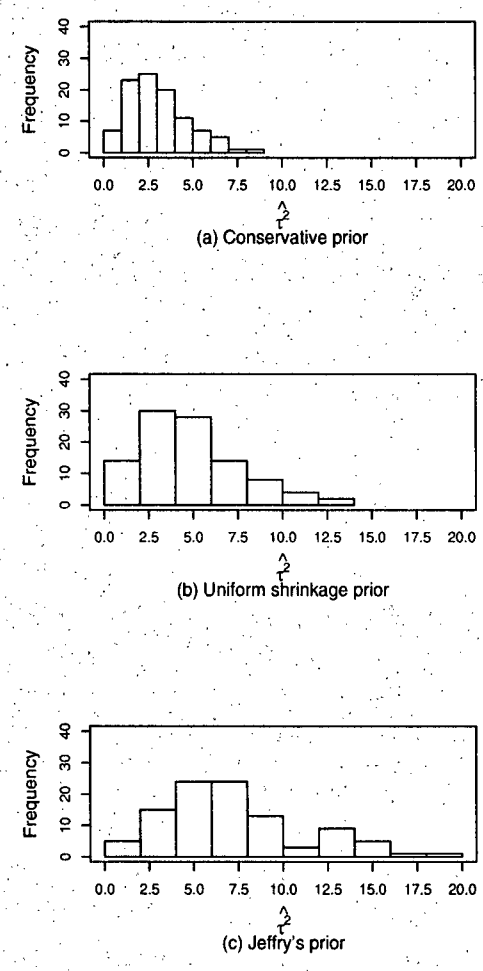
$$\pi(\lambda|y) \propto \frac{\sqrt{(R)}\lambda^{\frac{p-2}{2}}}{|A'A + \lambda V^{-1}|^{\frac{1}{2}} W^{\frac{n}{2}}}.$$

The conditional posterior densities for θ and σ^2 have closed form. So, we can use Gibbs sampler to draw posterior simulations for θ and σ^2. But, we have to use random walk Metropolis–Hastings algorithm to draw posterior simulations for λ. Once we have the estimate of λ we can get the estimate of τ^2 as $\hat{\tau^2} = \frac{\hat{\sigma^2}}{\hat{\lambda}}$.

3.5.2 Comparison of Results: Simulation Studies

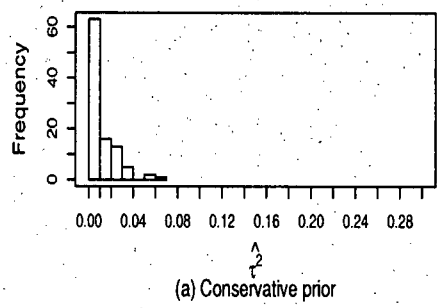
Simulation studies have been performed to compare the results obtained by using the competitive priors in smoothing problem with respect to guarding against undersmoothing. To see whether the proposed conservative prior can perform reasonably well in case of both large and

Figure 3.26: Histograms of $\hat{\tau^2}$ for different priors when $\tau^2 = 5$ and $n = 111$

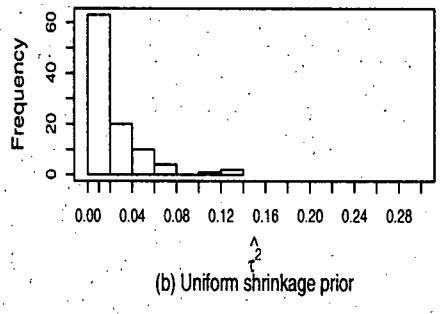


small τ^2 values, we have conducted our simulation studies for $\tau^2 = 5$ and $\tau^2 = 0.01$ while keeping the true σ^2 value fixed at 1. For the purpose of comparison we have used $a = 5$ for the conservative prior. Histograms of $\hat{\tau^2}$ for conservative prior, uniform shrinkage prior and Jeffreys' prior based on 100 different data sets are plotted in Figures 3.26 and 3.27, respectively for $\tau^2 = 5$ and $\tau^2 = 0.01$ when $n = 111$. The same for $n = 23$ are plotted in Figures 3.28 and 3.29.

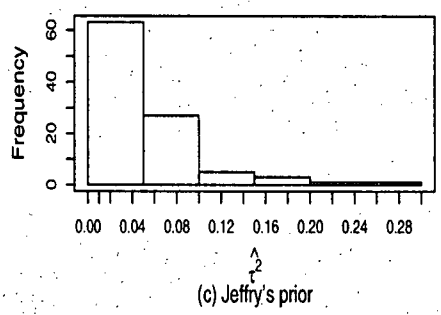
Figure 3.27: Histograms of $\hat{\tau^2}$ for different priors when $\tau^2 = 0.01$ and $n = 111$



(a) Conservative prior

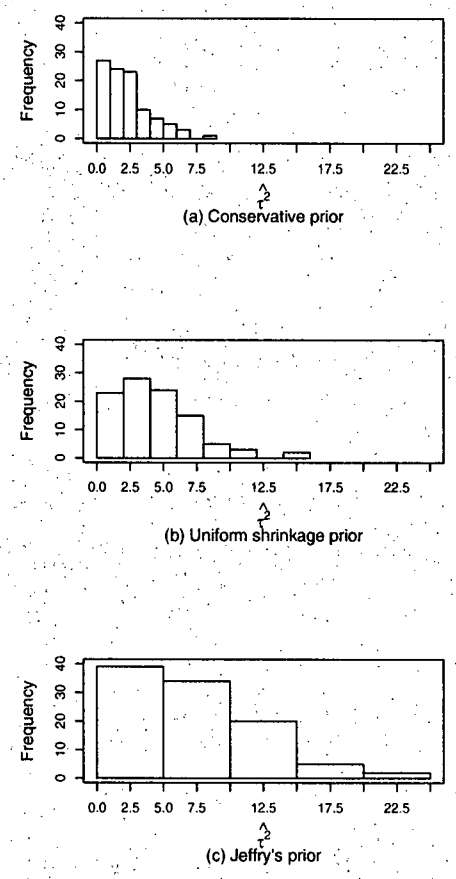


(b) Uniform shrinkage prior



(c) Jeffry's prior

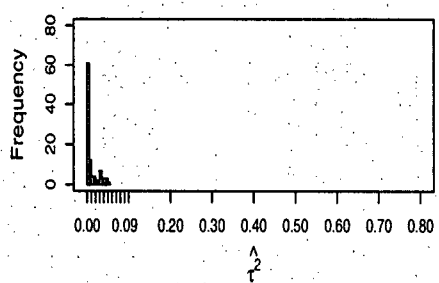
Figure 3.28: Histograms of $\hat{\tau^2}$ for different priors when $\tau^2 = 5$ and $n = 23$



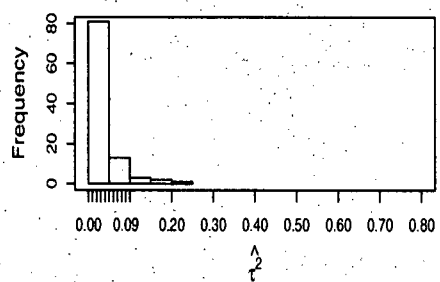
Histograms in Figure 3.26 reveal that the conservative prior can guard against undersmoothing most effectively. Jeffreys' prior produces the worst result in this regard. A similar picture is observed in the case of no heterogeneity ($\tau^2 = 0.01$) (Figure 3.27). It is also observed that in the case of low τ^2 undersmoothing is very severe for the uniform shrinkage and the Jeffreys' priors with more than 50 and 80 out of 100 estimates of τ^2 larger than the true τ^2 value for the uniform shrinkage prior and the Jeffreys' prior, respectively. For small sample size the picture is similar as it is observed in the case of large sample size (Figures 3.28 and 3.29).

Now, to see the performance of the three competitive priors in terms of MSE of $\hat{\theta}$ we have also used the simulated data to compute the differences between MSEs obtained by conservative prior and those obtained by uniform shrinkage prior and Jeffreys' prior for 100 data sets. Finally, we have counted the number of times that MSEs of $\hat{\theta}$ in the case of Jeffreys' and uniform

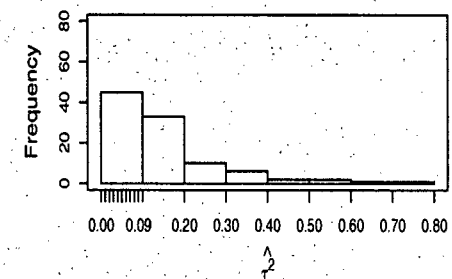
Figure 3.29: Histograms of $\hat{\tau^2}$ for different priors when $\tau^2 = 0.01$ and $n = 23$



(a) Conservative prior



(b) Uniform shrinkage prior



(c) Jeffry's prior

shrinkage priors exceed the MSEs of $\hat{\theta}$ obtained by conservative prior. It can be mentioned that in calculating the MSE of $\hat{\theta}$ we have averaged over the knots, not over the repeated data set, i.e., the MSE of $\hat{\theta}$ is obtained as

$$\text{MSE}(\hat{\theta}) = \frac{1}{p}\sum_{i=1}^{p}(\hat{\theta}_i - \theta_i)^2,$$

where p is the number of knots used to estimate the regression curve.

Tables 3.1, 3.2 and 3.3 represent those counts for $\tau^2 = 5$, $\tau^2 = 1$ and $\tau^2 = 0.01$, respectively.

Table 3.1: Table summarizing the number of times that MSE($\hat{\theta}$) under Jeffreys' prior and uniform shrinkage prior exceed those under conservative prior out of 100 occasions when $\tau^2 = 5$

Sample Size	#[mse$(\hat{\theta})_{usp}$ >mse$(\hat{\theta})_{con}$]	#[mse$(\hat{\theta})_{Jeff}$ >mse$(\hat{\theta})_{con}$]
111	34	43
23	19	33

Table 3.2: Table summarizing the number of times that MSE($\hat{\theta}$) under Jeffreys' prior and uniform shrinkage prior exceed those under conservative prior out of 100 occasions when $\tau^2 = 1$

Sample Size	#[mse$(\hat{\theta})_{usp}$ >mse$(\hat{\theta})_{con}$]	#[mse$(\hat{\theta})_{Jeff}$ >mse$(\hat{\theta})_{con}$]
111	37	52
23	32	48

Table 3.3: Table summarizing the number of times that MSE($\hat{\theta}$)under Jeffreys' prior and uniform shrinkage prior exceed those under conservative prior out of 100 occasions when $\tau^2 = 0.01$

Sample Size	#[mse$(\hat{\theta})_{usp}$ >mse$(\hat{\theta})_{con}$]	#[mse$(\hat{\theta})_{Jeff}$ >mse$(\hat{\theta})_{con}$]
111	47	63
23	97	96

From Table 3.1 we observe that for $\tau^2 = 5$ uniform shrinkage prior and Jeffreys' priors perform better than conservative prior in terms of MSE of $\hat{\theta}$ irrespective of sample size. For both large and small sample sizes the count that MSEs of $\hat{\theta}$ obtained by uniform shrinkage prior and Jeffreys' priors exceed those obtained by conservative prior are less than 50 out of 100 occasions. But the picture is reverse for smaller τ^2 ($\tau^2 = 0.01$). In this case of no heterogeneity

the performance of the conservative prior is far better than that of Jeffreys' prior, irrespective of sample size. For large sample size uniform shrinkage prior and conservative prior perform similarly, but for small sample size the conservative prior performs much better than uniform shrinkage prior. When τ^2 is of equal magnitude of σ^2, the error variance, uniform shrinkage prior still produces smaller MSE for more than 50 occasions out of 100 occasions, but the chance that the Jeffreys' prior produces higher MSE than the conservative prior are 52 and and 48, respectively for large and small sample size out of 100 occasions.

Again, counting the number of times that MSEs of $\hat{\theta}$ for uniform shrinkage prior and Jeffreys' prior exceed those for conservative prior can not give any idea about the magnitudes by which the MSEs of $\hat{\theta}$ for uniform shrinkage prior and Jeffreys' priors exceed those for conservative prior and vice–versa. So, to reflect this picture we have constructed histograms of the differences of MSEs of $\hat{\theta}$ for each pair of the competitive priors. Figures 3.30, 3.31, 3.32, 3.33, 3.34 and 3.35 display these histograms.

Figure 3.30: Histograms of differences of MSE($\hat{\theta}$) for pairs of different priors when $\tau^2 = 5$ and $n = 111$

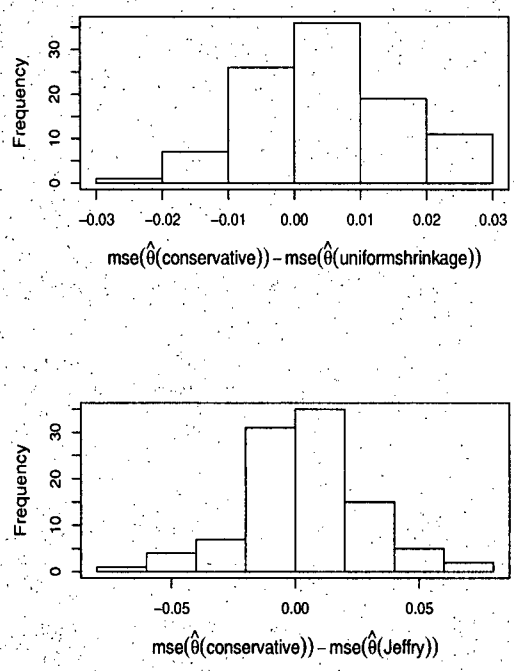


Figure 3.31: Histograms of differences of MSE($\hat{\theta}$) for pairs of different priors when $\tau^2 = 5$ and $n = 23$

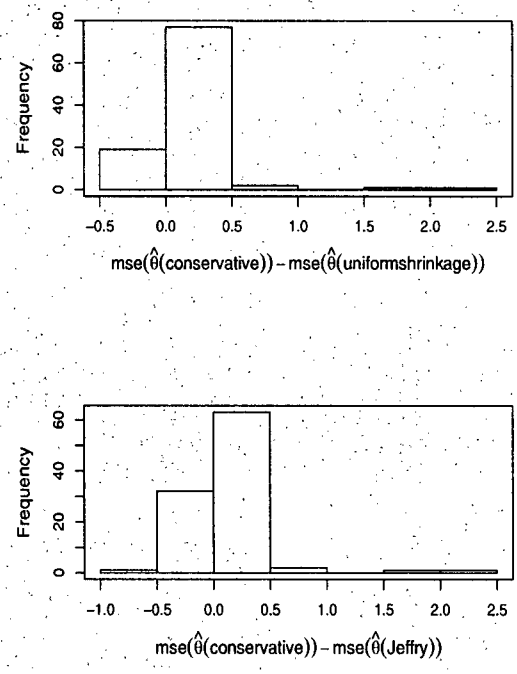


Figure 3.32: Histograms of differences of MSE($\hat{\theta}$) for pairs of different priors when $\tau^2 = 1$ and $n = 111$

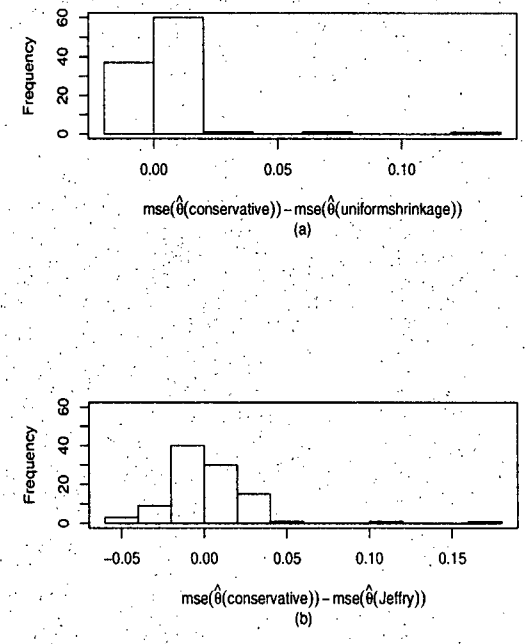


Figure 3.33: Histograms of differences of MSE($\hat{\theta}$) for pairs of different priors when $\tau^2 = 1$ and $n = 23$

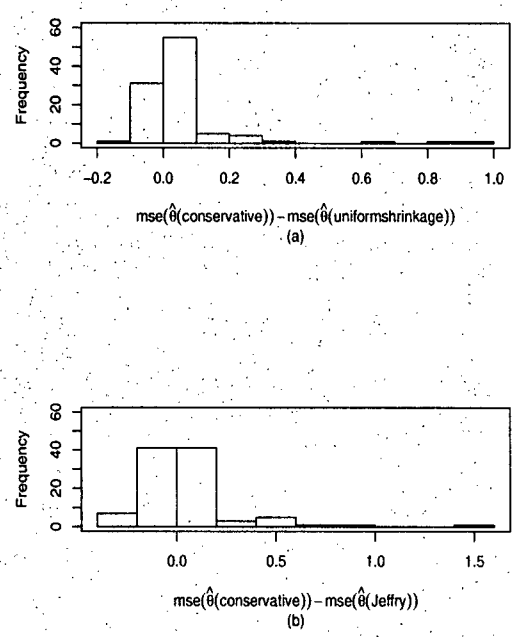


Figure 3.34: Histograms of differences of MSE($\hat{\theta}$) for pairs of different priors when $\tau^2 = 0.01$ and $n = 111$

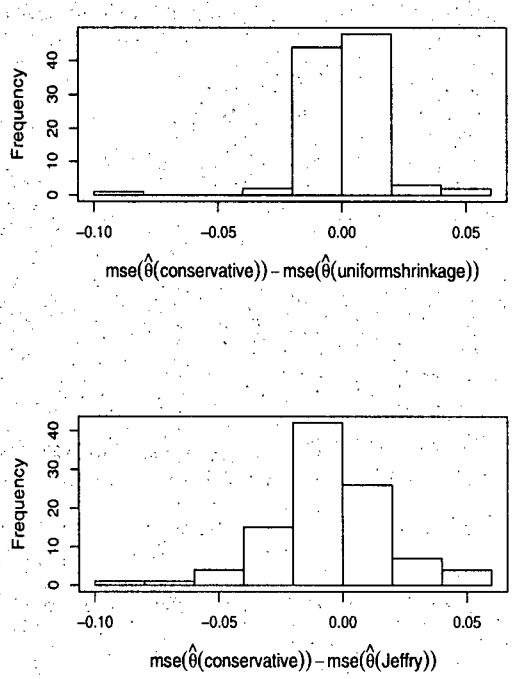
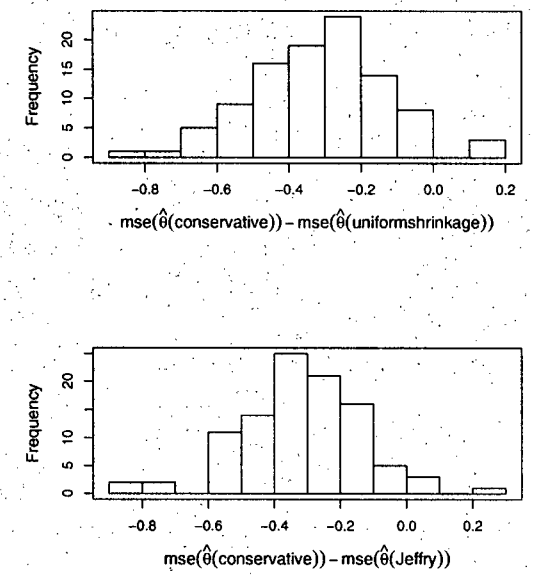


Figure 3.35: Histograms of differences of MSE($\hat{\theta}$) for pairs of different priors when $\tau^2 = 0.01$ and $n = 23$



Figures 3.30 and 3.31 display the histograms of MSE differences for bigger τ^2 value ($\tau^2 = 5$) for large and small sample sizes, respectively. For large sample size the magnitudes of difference are the same both ways around zero, which indicates that though the conservative prior produces large MSE for $\hat{\theta}$ in more occasions the magnitudes by which the MSEs of $\hat{\theta}$ for conservative prior exceed those for uniform shrinkage prior and Jeffreys' prior are the same as those in the case of other way around (Figure 3.30). For small sample size the magnitudes by which the MSEs of $\hat{\theta}$ for conservative prior exceed those for the other two priors are bigger in very few cases than those in the reverse case (Figure 3.31). Histograms for MSE differences in the case of $\tau^2 = 1$ reveal the similar picture as they do in the case of larger τ^2 value. Finally, by looking at the histograms of Figures 3.34 and 3.35 we observe that in the case of no heterogeneity the conservative prior beats the other two priors with big margin in all respect.

3.5.3 Comparison of Results for $a = 3$

In the previous section, for conservative prior we considered the value of the hyper parameter a to be equal to 5. But for normal–normal hierarchical model, simulation studies showed that any value from 3 to 5 can guard against undersmoothing well. Actually, the bigger the value of a the smoother the estimated function will be. So, using $a = 5$ may oversmooth too much

and that is why the performance of conservative prior, in terms of MSE of $\hat{\theta}$, might be worse than those of the other two priors for large τ^2 value. Lowering the value of a to 3 may improve the performance of conservative prior in terms of MSE of $\hat{\theta}$, while keeping undersmoothing in control at the same time. To check this fact we have estimated the model parameters by using conservative prior with $a = 3$ and compared the results as we did in the case of $a = 5$ in the previous section.

Figures 3.36 and 3.37 display the histograms of $\hat{\tau^2}$ for the three competitive priors for large ample size when $\tau^2 = 5$ and $\tau^2 = 0.01$, respectively. The same for small sample size are displayed in Figures 3.38 and 3.39.

Figure 3.36: Histograms of $\hat{\tau^2}$ for different priors when $\tau^2 = 5$ and $n = 111$

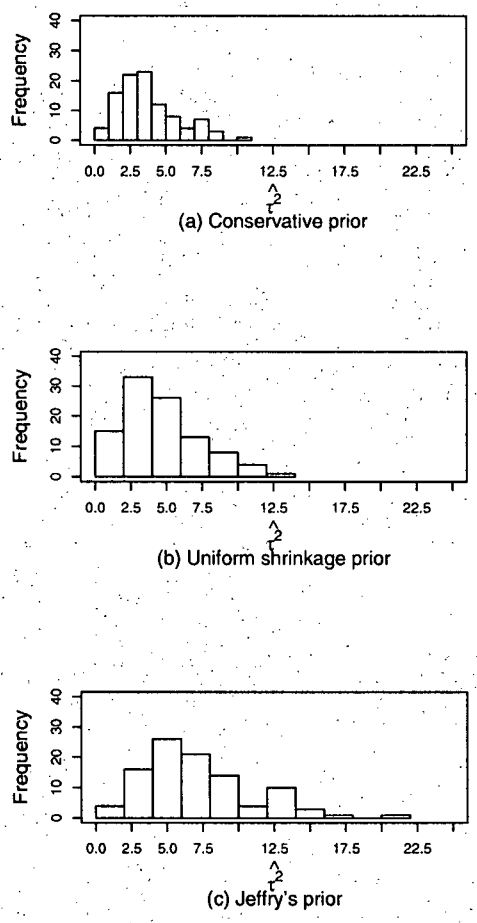
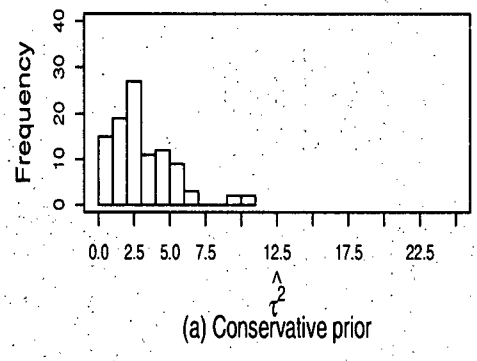
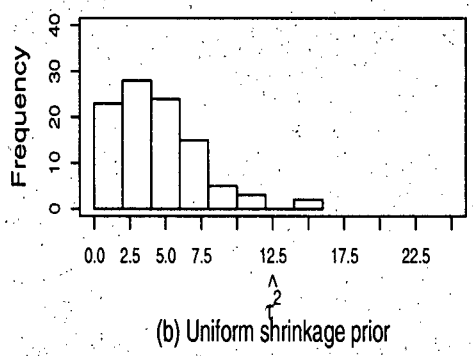


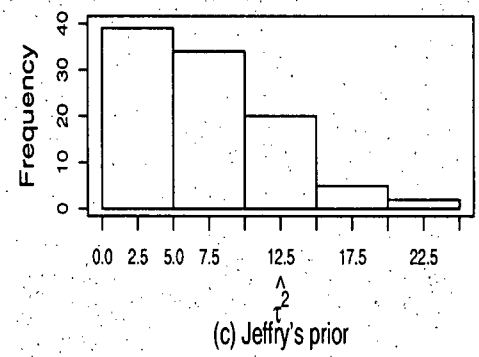
Figure 3.37: Histograms of $\hat{\tau^2}$ for different priors when $\tau^2 = 5$ and $n = 23$



(a) Conservative prior

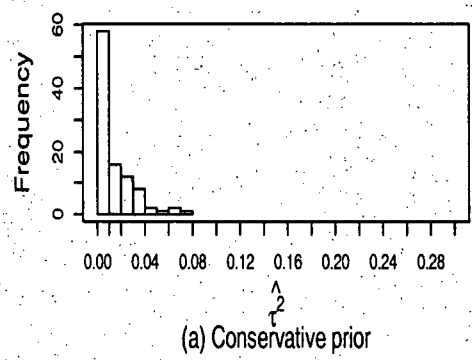


(b) Uniform shrinkage prior

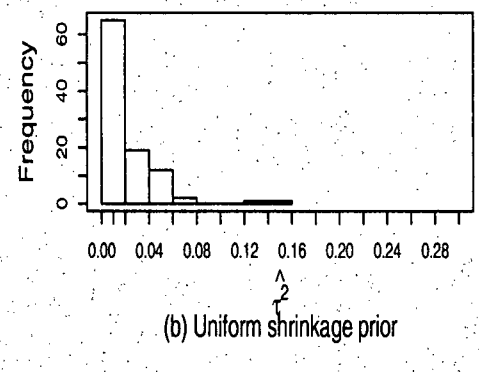


(c) Jeffry's prior

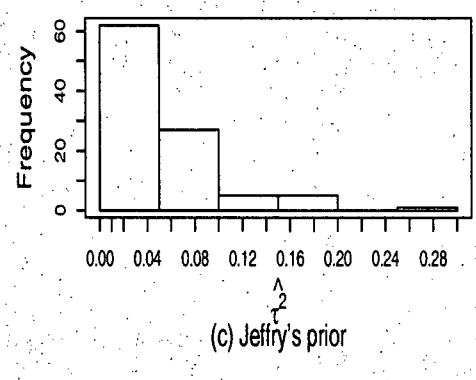
Figure 3.38: Histograms of $\hat{\tau^2}$ for different priors when $\tau^2 = 0.01$ and $n = 111$



(a) Conservative prior

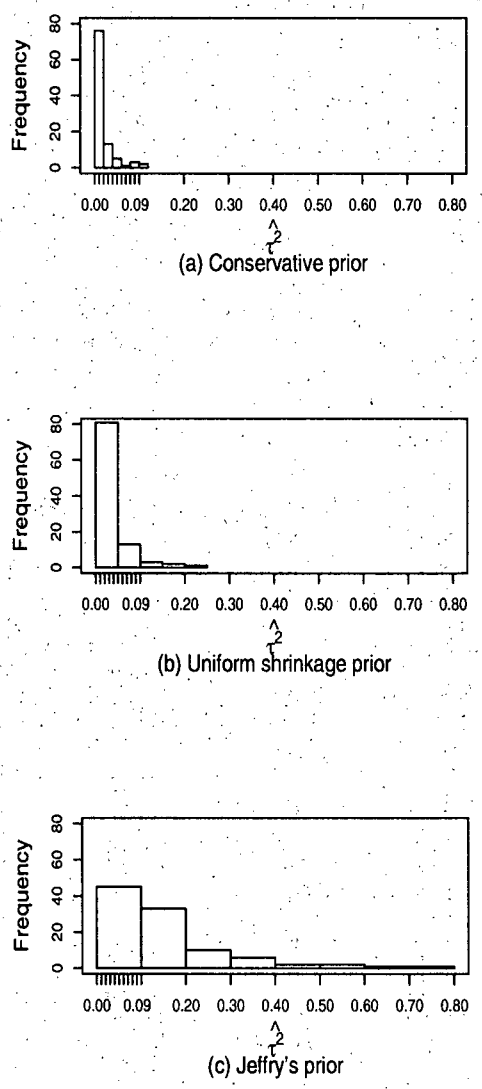


(b) Uniform shrinkage prior



(c) Jeffry's prior

Figure 3.39: Histograms of $\hat{\tau^2}$ for different priors when $\tau^2 = 0.01$ and $n = 23$



All of the Figures 3.36, 3.37, 3.38 and 3.39 confirm that conservative prior performed best in controlling undersmoothing in all the situations considered. To reflect the performance of conservative prior with $a = 3$ compared to the other two priors considered with respect to MSE of $\hat{\theta}$ we have summarized the number of times that MSE of $\hat{\theta}$ in the case of Jeffreys' and uniform shrinkage priors exceed the MSEs of $\hat{\theta}$ obtained by conservative prior in Tables 3.4, 3.5 and 3.6.

From the counts of the Tables 3.4, 3.5 and 3.6 we observe that when τ^2 is relatively larger than

Table 3.4: Table summarizing the number of times that MSE($\hat{\theta}$) under Jeffreys' prior and uniform shrinkage prior exceed those under conservative prior with $a = 3$ out of 100 occasions when $\tau^2 = 5$

Sample Size	#[mse$(\hat{\theta})_{usp}$ >mse$(\hat{\theta})_{con}$]	#[mse$(\hat{\theta})_{Jeff}$ >mse$(\hat{\theta})_{con}$]
111	37	46
23	25	46

Table 3.5: Table summarizing the number of times that MSE($\hat{\theta}$) under Jeffreys' prior and uniform shrinkage prior exceed those under conservative prior with $a = 3$ out of 100 occasions when $\tau^2 = 1$

Sample Size	#[mse$(\hat{\theta})_{usp}$ >mse$(\hat{\theta})_{con}$]	#[mse$(\hat{\theta})_{Jeff}$ >mse$(\hat{\theta})_{con}$]
111	37	52
23	35	52

the error variance σ^2 the uniform shrinkage prior performs better, in terms of MSE of $\hat{\theta}$, in more occasions than the proposed conservative prior. In case the when τ^2 is relatively lower than the error variance the conservative prior with $a = 3$ performs better in more occasions. For the situation when the global variability is of equal magnitude of the local variability, though the performance of the conservative prior improves in the case of small sample size the counts that the MSEs of $\hat{\theta}$ for uniform shrinkage prior exceed those for the conservative prior are still less than 50 out of 100 occasions. Regarding the comparison of the performance of conservative prior and Jeffreys' prior, the output suggest that in case of lager τ^2 conservative prior and Jeffreys' prior performs almost similarly but in the reverse case conservative prior always performs far better than Jeffreys' prior in terms of MSE of the estimated curve.

Table 3.6: Table summarizing the number of times that MSE($\hat{\theta}$)under Jeffreys' prior and uniform shrinkage prior exceed those under conservative prior with $a = 3$ out of 100 occasions when $\tau^2 = 0.01$

Sample Size	#[mse$(\hat{\theta})_{usp}$ >mse$(\hat{\theta})_{con}$]	#[mse$(\hat{\theta})_{Jeff}$ >mse$(\hat{\theta})_{con}$]
111	55	74
23	58	83

3.6 Conclusion

In this chapter we studied the performance of the proposed conservative prior with respect to its ability to controlling undersmoothing through simulation studies. We also studied the performance of the proposed prior with compared to the Jeffreys' prior and the uniform shrinkage prior. The reason for considering the uniform shrinkage prior and the Jeffreys' prior as the competitor of the proposed prior was that these two priors are common choices in Bayesian hierarchical models. Also, they perform better than other priors used for Bayesian hierarchical models (Daniels, 1999). Furthermore, the uniform shrinkage prior can do some degree of smoothing due to its property of shrinkage toward zero. Like the uniform shrinkage prior, in a Bayesian hierarchical model the proposed conservative prior make possible to get a proper posterior distribution of the model parameters, which is essential for valid posterior inferences about the parameters. From simulation studies we observed that the proposed conservative prior performs better in controlling undersmoothing than the other two priors considered in this study. Also, in terms of MSE of the estimated curve the proposed prior exhibits better performance for the data sets when the global variation is relatively smaller than the local variation. In the other case also the proposed prior with $a = 3$ showed better performance in terms of MSE of the estimated curves for many data sets. So, from the result of simulation studies it can argued that using reasonably chosen value of the hyperparameter a, e.g. $a = 3$, for the conservative prior can nicely control the excessive degree of undersmoothing while, at the same time, not oversmoothing too much, and hence can give more accurate estimate of τ^2 and consequently of the other model parameters. Thus the Bayesian approach with conservative prior can give a better picture of the true underlying relationship between response and the covariate in smoothing problem.

Again, if it is believed that the true underlying relationship between the response and the

covariate is smooth then it is desirable to have an estimated curve which is also smooth. In such cases controlling undersmoothing may be the main concern and hence the use of conservative prior, with a value between 3 to 5 of the hyperparameter a, can ensure the smoothness of the estimated curves. Finally, from the results of the simulation studies we can summarize that for a response variable fluctuating rapidly as a function of the design variable it is better to use the conservative prior while using the Bayesian roughness penalty approach in regression curve estimation. Actually, this is the case of larger relative error variability with compared to the global variability, and in such case a smooth estimated curve is really desirable. Also, since the Bayesian roughness penalty approach with the conservative prior for the random effects variance component provides smooth estimate of the true mean curve it is robust against any wild local observations.

Chapter 4

Discussion and Conclusion

The work of this dissertation focuses on finding an appropriate prior distribution for the variance component in the Bayesian hierarchical models. In many real life situations data arise in hierarchical fashion. Some common situations where the hierarchical models arise are the random effect models, spatial models and non–parametric curve–fitting. All the above mentioned models can further be viewed as smoothing problem, and it is often desirable to have an estimated mean structure which is not more variable than the true underlying structure or the estimated regression curve which is not more wiggly than the true underlying curve. That is, we do not want to undersmooth the phenomenon we estimate. The motivation for this work comes from the thought that we should have an estimation technique which can control the undersmoothing while estimating a phenomenon by hierarchical models. In searching such a technique for hierarchical models we have considered the Bayesian approach because of its flexibility to incorporate the multiple level of randomness– the main feature of hierarchical models. Bayesian approach to hierarchical models have many other advantages over the classical approach which have already been discussed in Chapter 1.

Finally, while adopting the Bayesian approach for hierarchical models we have thought of choosing a prior for the random effects variance component which can guard against undersmoothing. We have conducted simulation studies to choose the appropriate values for the hyperparameter for the proposed prior so that it can provide adequate guard against undersmoothing and at the same time do not oversmooth the phenomenon too much and thus can provide better calibrated estimates of the models parameters.

In Chapter 1, we have formulated the Bayesian version of the normal–normal hierarchical model.

Simulation studies reveal that the proposed conservative prior, with values between 3 to 5 of the hyperparameter a, perform quite well in guarding against undersmoothing. With compared to the two of its competitors– the uniform shrinkage prior and the Jeffreys' prior– the proposed prior perform much better in controlling undersmoothing in all the situations considered for the simulation studies. Also, in terms of the MSE of the estimated variance component the proposed prior, with $a = 3$, perform much better than its competitors in all the situations considered. Even with $a = 5$ the proposed prior performs better than the Jeffreys' prior in all but one cases considered and performs better than the uniform shrinkage prior in the case when the between group variability is relatively smaller than the within group variability.

In Chapter 3, we have formulated the Bayesian hierarchical model for the roughness penalty approach to curve–fitting. Then we have suggested a modified version of the conservative prior for that hierarchical model. Finally, we have conducted simulation studies to see how the conservative prior performs in smoothing problem to guard against undersmoothing. Also, we have investigated the performance of the conservative prior with those of the uniform shrinkage prior and the Jeffreys' prior via simulation studies. The results are encouraging in terms of guarding against undersmoothing. In terms of MSE of the estimated curve the proposed conservative prior performs better than its competitors in the case when global variability is less than the local variability. Also, in the other two cases the conservative prior performs better in more than 50% of the data sets considered with compared to the Jeffreys' prior. With compared to the uniform shrinkage prior the conservative prior produces smaller MSE in many occasions (though less than 50%). Again, it is always believed that the true underlying curve in non–parametric regression is usually smoothed. So, for most of the applications in non–parametric curve–fitting the smoothness of the estimated curves is the main concern, and hence the use of conservative prior with the suggested values of the hyperparameter a can ensure the desired degree of smoothness of the estimated curves. If someone is concerned about the MSE of the estimated curves also he/she may use the lowest value from the suggested range (3 to 5) for the hyperparameter.

In a nutshell, on the basis of the results of simulation studies we can conclude that in the case of simple normal–normal hierarchical model the proposed conservative prior with $a = 3$ can be used without any reservation because with $a = 3$ the conservative prior beats the other two priors considered in this study– both in terms of guarding against undersmoothing as well as in terms of MSE of the estimated variance component. Also, in many situations where the smoothness of the estimates is the only concern, e.g., in the case of hospital model of recovery rates of a cardiovascular treatment in different hospitals, even bigger values of a (not exceeding

5) can be used. Again, in smoothing problem researchers can choose between the uniform shrinkage prior and the conservative prior if they think that the global variability is relatively larger than the local variability. In the reverse case the use of conservative prior is always better. Also, in the case when the researchers is concerned only about the smoothness of the estimated curve the use of the conservative may get preference over the uniform shrinkage prior.

As in all other areas there is a huge scope for further research in this area. Firstly, in the present study we have proposed the conservative prior and investigated its performance in the context of continuous response only. Further work is required to make the use of the conservative prior to the response arise from any distribution belongs to the exponential family. In this study, in normal–normal hierarchical model we did not consider the incorporation of predictors to predict the mean response. The hierarchical model that incorporates the response–predictor relationship for the Gaussian response is the linear mixed effects model (LMM). In LMM, the classical approach considers random effects variance component to be fixed and be estimated from the data. In doing so, the classical approach ignores the uncertainty in the random effects variance component. So, the reasonable alternative is the Bayesian hierarchical models. And in the case of Bayesian hierarchical models, we can consider the use of the conservative prior for the random effects variance component.

Again, further work needs to be done to extend the use of the proposed conservative prior in the broad areas of generalized linear mixed models (GLMM). In the case of GLMM, the existing classical approaches lead to inefficient inference about the random effects. In GLMM, the estimate of random effects variance components becomes very difficult when the number of random effects becomes large. The approximate method proposed by Breslow and Clayton may produce estimate of the random effects covariance matrix which is negative definite (Breslow and Clayton, 1993). But the Bayesian formulation enjoys an advantage here because of the information on the variance component provided by the prior distribution. Also, Bayesian procedures avoid the need for numerical integration by taking repeated samples form the posterior distribution. But the main concern in the Bayesian approach is to choose an appropriate prior for the random effects variance components, and the proposed conservative prior, if can be extended for GLMM, may provide a good alternative in this area. Natarajan and Kass (Ranjini Natarajan and Robert E. Kass, 2000) proposed the approximate uniform shrinkage prior and the approximate Jeffreys' prior for the random effects covariance matrix in GLMM. But the main concern with the use of those prior is undersmoothing. So, we can think of extending the use of proposed conservative prior for the random effects covariance matrix in GLMM to check undersmoothing and to have better calibrated estimates of the model parameters.

Bibliography

[1] Barger, J. O. and Deely, J. J. A bayesian approach to ranking and selection of related means with alternative to analysis–of–variance methodology. *Journal of the American Statistical Association*, 83:364–373, 1988.

[2] Barger, J. O. and Bernando, J. M. On the development of reference priors. In Barger, J. O., Bernando, J. M , Dawid, A. P., and Smith,F. M., editors, *Bayesian Statistics-4*, pages 35–60. Oxford University Press, 1992.

[3] Barry, Daniel. A bayesian analysis for a class of penalized likelihood estimates. *Communications in Statistics*, 34(4):1057–1071, 1995.

[4] Box, G. E. P. and Tiao, G. C. *Bayesian Inference in Statistical Analysis.* Wiley, New York, 1973.

[5] Breslow, N. E. and Clayton, D. G. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

[6] Michael J. Daniels. A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, 27, No.3:567–578, 1999.

[7] Eubank, R. L. *Spline Smoothing and Nonparametric Regression.* Marcel Drekker, New York, 1988.

[8] Gilks, at. el. *Markov Chain Monte Carlo in Practice.* Chapman and Hall/CRC, 1996.

[9] Green, P. J. and Silverman, B. W. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall, London, 1994.

[10] Hastie, T. and Tibshirani, R. *Generalized Additive Models.* Chapman and Hall, New York, 1990.

[11] Hastie, T. and Tibshirani, R. Bayesian backfitting. *Statistical Science*, 15, No.3:196–223, 2000.

[12] Hobert, J. P. and Casella, G. The effect of improper priors on gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91:1461–1476, 1996.

[13] Jeffreys, H. *Theory of Probability.* Oxford University Press, 1961.

[14] McCullagh, P.and Nelder, J. A. *Generalized Linear Models.* Chapman and Hall, London, 1989.

[15] Ranjini Natarajan and Robert E. Kass. Reference bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95:227–237, 2000.

[16] Silverman, B. W. . A fast and efficient cross–validation method for smoothing parameter choice in spline regression. *Journal of the American Statistical Association*, 79:584–589, 1984.

[17] Simonoff, Jeffrey S. *Smoothing Methods in Statistics.* Springer–Verlag, New York, 1996.

[18] Wahhba, G. *Spline Models for Observational Data.* SIAM, 1990.