Characterizing Modes of the Likelihood, Restricted Likelihood and Posterior for Hierarchical Models

Jiannong Liu and James S. Hodges

September 30, 1999

Division of Biostatistics
School of Public Health
University of Minnesota
Contents

1 Introduction 1

2 Basic Concepts 4
   2.1 Maximum Likelihood Estimate 6
   2.2 Restricted Maximum Likelihood Estimate 7
   2.3 Bayes Estimate 9

3 The Modality of the Likelihood and Restricted Likelihood for the One-way Random Effect Model 12
   3.1 The Maximum Likelihood Estimate 12
   3.2 The REML Estimate 13

4 The Modality of the Joint Posterior for the One-Way Random Effect Model 14
   4.1 The Modality of the Joint Posterior 14
   4.2 One-Way Random Effect Model with \( \tau^2 \) Having an Improper Prior 23
   4.3 Intuitive Explanation of Theorem 4.1; Sensitivity Analysis of Modes 24
   4.4 Examples 27

5 Future Work 32

6 References 33

7 Appendix 37
   7.1 Appendix A: The Proof of Lemma 1 37
   7.2 Appendix B: The Proof of Theorem 4.1 38
List of Tables

1  Symbol table ............................................................. iv
2  Values of $\Delta$, $\phi$ and $f(\phi)$ for different values of $\lambda$ ......................... 28
3  Values of $\Delta$, $\phi$ and $f(\phi)$ for different values of $b$ ................................. 29
4  Values of $\Delta$, $\phi$ and $f(\phi)$ for different values of $\alpha$ ............................. 30
5  Values of $\Delta$, $\phi$ and $f(\phi)$ for different values of $a$ ................................. 30
6  Values of $\Delta$, $\phi$ and $f(\phi)$ for different values of $S_B^2$ ............................ 31

List of Figures

1  The plot of $f(\phi)$ for $\Delta < 0$ ........................................ 18
2  The plot of $f(\phi)$ for $\Delta = 0$ ....................................... 19
3  The plot of $f(\phi)$ for $\Delta > 0$ ....................................... 20
4  The plot of $f(\phi)$ and $\ell(\phi)$ for different $\lambda$ ......................... 52
5  The plot of $f(\phi)$ and $\ell(\phi)$ for different $b$ ............................ 53
6  The plot of $f(\phi)$ and $\ell(\phi)$ for different $\alpha$ ............................ 54
7  The plot of $f(\phi)$ and $\ell(\phi)$ for different $a$ ............................ 55
8  The plot of $f(\phi)$ and $\ell(\phi)$ for different $S_B^2$ ............................ 56
9  The plot to show how $\dot{\phi}^{(1)}$, $\dot{\phi}^{(s)}$ move as $b$ increases ..... 57
Table 1: Symbol table

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Explanation</th>
<th>page (formula)</th>
</tr>
</thead>
<tbody>
<tr>
<td>model (2)</td>
<td>the two level hierarchical model</td>
<td>5 (2)</td>
</tr>
<tr>
<td>model (6)</td>
<td>the one-way random effect linear model with balanced data</td>
<td>11 (6)</td>
</tr>
<tr>
<td>model (8)</td>
<td>the one-way random effect bayes linear model with balanced data</td>
<td>13 (8)</td>
</tr>
<tr>
<td>( \nu )</td>
<td>all the unknown parameters in the covariance matrices</td>
<td>5</td>
</tr>
<tr>
<td>( L(\beta, \nu; Y) )</td>
<td>the likelihood w.r.t. ( \beta ) and ( \nu ) for model (3)</td>
<td>6 (4)</td>
</tr>
<tr>
<td>( l(\beta, \nu; Y) )</td>
<td>the log likelihood, ( \log(L(\beta, \nu; Y)) )</td>
<td>6</td>
</tr>
<tr>
<td>( I_\nu(Y) )</td>
<td>the profile of the log likelihood w.r.t. ( \nu )</td>
<td>6</td>
</tr>
<tr>
<td>( I_1(Y) )</td>
<td>the REML log likelihood w.r.t. ( \nu )</td>
<td>8</td>
</tr>
<tr>
<td>( I_k )</td>
<td>( k \times k ) identity matrix</td>
<td></td>
</tr>
<tr>
<td>( J_k )</td>
<td>( k \times 1 ) matrix with elements 1</td>
<td></td>
</tr>
<tr>
<td>( U_k )</td>
<td>( k \times k ) matrix with elements 1</td>
<td></td>
</tr>
<tr>
<td>( S_B^2 )</td>
<td>( = n \sum_{i=1}^{N} (\bar{y}_i - \bar{y})^2 ), sum of squares between groups</td>
<td>12</td>
</tr>
<tr>
<td>( S_W^2 )</td>
<td>( = \sum_{i=1}^{N} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 ), sum of squares within groups</td>
<td>12</td>
</tr>
<tr>
<td>( \alpha, \lambda, a, b )</td>
<td>the parameters in the priors of ( \sigma^2 ) and ( \tau^2 )</td>
<td>13</td>
</tr>
<tr>
<td>( \ell )</td>
<td>the log posterior ( \ell(\theta_1, \ldots, \theta_N, \mu, \sigma^2, \tau^2) ) for model (8)</td>
<td>14 (10)</td>
</tr>
<tr>
<td>( \ell_1 )</td>
<td>the profile log posterior ( \ell_1(\sigma^2, \tau^2) )</td>
<td>14 15</td>
</tr>
<tr>
<td>( \phi )</td>
<td>( = \frac{\sigma^2}{\sigma^2 + \tau^2} ), the “shrinkage factor”</td>
<td>15</td>
</tr>
<tr>
<td>( e )</td>
<td>( nN + 2\alpha + 2 )</td>
<td>16</td>
</tr>
<tr>
<td>( f )</td>
<td>( N + 2\alpha + 2 )</td>
<td>16</td>
</tr>
<tr>
<td>( \ell_2 )</td>
<td>( \ell_2(\sigma^2, \tau^2, \phi) ), formed by replacing ( \frac{\sigma^2}{\sigma^2 + \tau^2} ) in ( \ell_1 ) by ( \phi )</td>
<td>15 (12)</td>
</tr>
<tr>
<td>( \hat{\tau}^2 )</td>
<td>the critical point of ( \ell_2 ) given ( \phi )</td>
<td>16(15)</td>
</tr>
<tr>
<td>( \hat{\sigma}^2 )</td>
<td>the solutions of ( f(\phi) = 0 ) with ( \hat{\sigma}^2 ) and ( \hat{\tau}^2 ) are the critical points of ( \ell_2 ) with restriction ( \phi = \frac{\sigma^2}{\sigma^2 + \tau^2} )</td>
<td>16 (18)</td>
</tr>
<tr>
<td>( f(\phi) )</td>
<td>the solutions of ( f(\phi) = 0 )</td>
<td>16</td>
</tr>
<tr>
<td>( f'(\phi) )</td>
<td>the derivative function of ( f(\phi) ).</td>
<td></td>
</tr>
<tr>
<td>( \Delta )</td>
<td>( 4(2e + f)^2 - 12(e + f) \left( \frac{2f \lambda + 2n \mu \phi}{2} + \frac{fN(n-1)}{N} \right) + e )</td>
<td>17 (19)</td>
</tr>
<tr>
<td>( \phi_1, \phi_2 )</td>
<td>the solutions of ( f'(\phi) = 0 )</td>
<td>17 (19)</td>
</tr>
<tr>
<td>( \hat{\phi}^{(1)} &lt; \hat{\phi}^{(2)} &lt; \hat{\phi}^{(3)} )</td>
<td>the solutions of ( f(\phi) = 0 )</td>
<td>20</td>
</tr>
</tbody>
</table>
1 Introduction

Recent computing advances have made it much easier to use hierarchical models to analyze data. For example, PROC MIXED in SAS, BUGS (or WinBUGS), HLM and other software can analyze highly complex hierarchical models. Many software packages involve maximizing a likelihood, restricted likelihood, posterior or similar function, while BUGS and other Monte Carlo routines draw repeated, not necessarily independent, samples from the posterior distribution to compute various approximate integrals. However, even for simple hierarchical models the shapes of these functions (likelihood, posterior and others) are unknown, and thus, all these computer packages may be subject to errors of unknown type and severity.

SAS PROC MIXED uses the Newton-Raphson algorithm to optimize the likelihood function or restricted likelihood function. Although the Newton-Raphson algorithm is preferred to other algorithms (Lindstrom and Bates 1988), it is possible for PROC MIXED to converge to a point that is not the global optimum of the likelihood (SAS Institute Inc. 1996, page 610). Suppose we use the mode of the restricted likelihood function as the estimator of the parameters in the covariance matrix and we do not get a global maximum point but a local maximal point, if we want to use the covariance matrix to make inference of other effects, the error may be more severe. So far, no one algorithm is guaranteed to find the global optimum if the function has more than one mode. Thus, we need to concern ourselves about whether the relevant function is unimodal. If the function is unimodal, we need not worry about the optimality of the solution, provided the algorithm has no problem with convergence. If the function is multimodal, we need to be careful when we explain the results or we need to do more to check the results, such as using different initial values for the algorithm, comparing the results from some other methods,
In Bayesian analysis, unimodality is also very important. For example, if the posterior is symmetric and unimodal, the Bayes estimator of a parameter does not change with the loss function if the loss function is from a specified class (Rukhin, 1978). Albert (1992) showed that if the posterior of a single parameter is unimodal, we can use some simple methods to summarize the posterior density. If the posterior is not unimodal, it is difficult to do so.

Even for simple Bayes inferences, unimodality is also important. For example, if the loss function is the 0-1 loss, i.e.

\[
l(\theta, a) = \begin{cases} 
0 & \text{if } a = \theta \\
1 & \text{if } a \neq \theta
\end{cases},
\]

where \( \theta \) is the parameter we are interested in and \( a \) is an estimator of \( \theta \), then the Bayes estimator of \( \theta \) is the mode of the posterior density of \( \theta \) given the data. If the loss function is the squared error, i.e.

\[
l(\theta, a) = (\theta - a)^2,
\]

then the Bayes estimator of \( \theta \) is the posterior mean. But sometimes we use the posterior mode as a substitute for posterior mean because the mean is sensitive to the tails of the posterior density while the posterior mode is not (Harville 1977). Also the computation for finding the mode of a function is in general much easier than for getting the mean. But if the posterior is multimodal, which mode makes sense? Which mode will an algorithm find?

If the posterior is unimodal, the highest posterior density (HPD) credible set is a connected region; if the posterior is multimodal, the HPD credible set can be a collection of disconnected regions. For the unimodal case, we can use some simple and fast algorithms to find the HPD set. But for the multimodal case, we may need a more complicated algorithm.
From the above, we see that the modality of likelihood, restricted likelihood and posterior is very important in statistical inference. But most papers related to the likelihood mode or posterior mode are aimed at finding searching methods to locate the modes for some specific models (Fahrmeir 1992, Eaves and Chang 1992, Chib 1996). A few papers have dealt with modality. With simulation, Mardia and Watkins (1989) showed that for Gaussian random fields, the profile likelihood can be multimodal with regard to covariance scale and range parameters. Dietrich (1991), continued Mardia and Watkins’ work, proved that under some conditions the likelihood in the spatial model can be reduced to unimodal or asymptotically unmodal. O’Hagan (1985) discussed the posterior modality of a hierarchical model and tried to figure out why sometimes the joint posterior has one mode but the marginal posterior has two modes. In the paper, he gave the concept of a “shoulder” of the joint posterior. Hoeschele (1988) showed the possible bimodality of a likelihood function, restricted likelihood function and a posterior density corresponding to a particular hierarchical model.

My research interest is to study the modality of these functions systematically for some classes of hierarchical (linear) models. These functions include the likelihood functions, restricted likelihood functions and posterior densities. I will characterize the modality of these functions and study how the unimodality of the posterior densities is sensitive to data and the parameters in the priors.
2 Basic Concepts

Hierarchical data structures are a common phenomenon in medical, social and other research. For example, in a clinical trial, we may have several centers and each center randomizes some patients from the local area to each treatment arm. Both center (area) and patient information are collected. The patient status \( x_{ij} \), like age, gender, treatment arm, etc.) may affect the treatment results \( y_{ij} \) directly, and different center characteristics \( z_i \) may make the average outcome different in each center.

For another example, in cross-national studies demographers might examine how differences in national economic development interact with adult education attainment to influence fertility rates. Such research combines economic indicators collected at the national level with household information on education and fertility. Both households and countries are units in the research with households nested within countries, and the basic data structure is again hierarchical (Bryk and Raudenbush 1992).

It is common to analyze this kind of data using hierarchical linear models. In the clinical trial example, we can write the model as

\[
y_{ij} | \eta_i, \gamma \sim N(\eta_i + x_{ij} \gamma, \sigma^2), \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, N, \\
\eta_i | \nu \sim N(z_i \nu, \tau^2), \quad i = 1, 2, \ldots, N, \tag{1}
\]

where \( N \) is the number of centers and \( n_i \) is the number of patients in center \( i \). The variables (parameters) \( x_{ij}, \gamma, z_i \) and \( \nu \) may be vectors. In model (1), \( \eta_i \) is the effect of center \( i \), \( \gamma \) is the effect of patient status on the treatment effect and \( \nu \) shows how the center characteristics affect the patient-average outcomes.

The above model is an example of a two-level hierarchical model. In general, a two-level
Hierarchical linear model with normal errors can be written as follows (Lindley and Smith 1972):

\[ Y|\theta \sim N(C\theta, R) \]
\[ \theta \sim N(A\beta, D). \] (2)

In the model, \( Y \) is an observation of an \( n \times 1 \) random variable; \( A \) and \( C \) are known matrices of “regressors” with dimensions \( q \times p, n \times q \), with rank \( p \) and \( q \), respectively; \( \theta \) and \( \beta \) are unobservable parameter vectors with dimensions \( q \times 1 \) and \( p \times 1 \), respectively, and \( R \) and \( D \) are covariance matrices.

Let \( \nu \) be the vector of all the unknown parameters in \( R \) and \( D \), where \( \nu \in \Omega \) and \( \Omega \) is some subset of a Euclidean space such that \( R \) and \( D \) are positive definite for \( \nu \in \Omega \). Usually, our interest is in the parameters \( \beta \) and \( \nu \). Sometimes, we may also be interested in \( \theta \). The most-often used approaches for estimating these parameters are the Maximum Likelihood method, the Restricted Maximum Likelihood method, the Bayes method and sometimes the Empirical Bayes method.

We can see that model (1) is an example of model (2) with \( C = I, R = \sigma^2 I, \nu = \begin{pmatrix} \sigma^2 \\ \tau^2 \end{pmatrix} \),

\[
A = \begin{pmatrix} z_1 & x_{11} \\ \vdots & \vdots \\ z_1 & x_{1n_1} \\ z_2 & x_{21} \\ \vdots & \vdots \\ z_2 & x_{2n_2} \\ \vdots & \vdots \\ z_N & x_{N1} \\ \vdots & \vdots \\ z_N & x_{Nn_N} \end{pmatrix}, \quad D = \tau^2 \begin{pmatrix} U_{n_1} & & \\ & U_{n_2} & \\ & & \ddots \end{pmatrix}, \quad \beta = \begin{pmatrix} w \\ \gamma \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_{11} \\ \vdots \\ \theta_{1n_1} \\ \vdots \\ \theta_{N1} \\ \vdots \\ \theta_{Nn_N} \end{pmatrix},
\]

where \( I \) is the \( \sum_{i=1}^{N} n_i \times \sum_{i=1}^{N} n_i \) identity matrix and \( U_{n_i} \) is an \( n_i \times n_i \) matrix with all elements 1. Actually, we reparameterized model (1).
2.1 Maximum Likelihood Estimate

When we are just interested in $\beta$ and $\nu$, we can consider the marginal distribution of $Y$ with parameters $\beta$ and $\nu$. From model (2), the marginal distribution of $Y$ is (Lindley and Smith 1972):

$$Y \sim N(CA\beta, R + CDC')$$.

Denote $V(\nu) = R + CDC'$ and $X = CA$; the marginal distribution of $Y$ becomes

$$Y \sim N(X\beta, V(\nu)). \quad (3)$$

Then the likelihood with respect to $\beta$ and $\nu$ is

$$L(\beta, \nu; Y) = \frac{1}{(2\pi)^{n/2}|\text{det}(V)|^{1/2}} \exp \left( -\frac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta) \right). \quad (4)$$

The log likelihood is

$$l(\beta, \nu; Y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\text{det}(V)) - \frac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta).$$

By definition, the maximum likelihood estimates (MLEs) of $\beta$ and $\nu$ are the values $\hat{\beta}$ and $\hat{\nu}$ satisfying

$$l(\hat{\beta}, \hat{\nu}; Y) = \sup_{\{\beta, \nu|\nu \in \Omega\}} l(\beta, \nu; Y).$$

We know that $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$ maximizes $l(\beta, \nu; Y)$ with respect to $\beta$ for fixed $\nu$. Since the elements of $V$ are functions of $\nu$, $\hat{\beta}$ is a function of $\nu$. Let $l_\nu(\nu; Y) = l(\hat{\beta}(\nu), \nu; Y)$ be the profile log likelihood with respect to $\nu$; then $\hat{\nu}$ is a maximum likelihood estimator of $\nu$ if and only if $\hat{\nu}$ maximizes $l_\nu(\nu; Y)$. Miller (1977) showed the consistency, asymptotic normality and the efficiency of the MLE for this model.
2.2 Restricted Maximum Likelihood Estimate

The disadvantage of the MLE is that it differs from the analysis-of-variance estimator in variance components model with balanced data, even though the latter have been shown to be the best quadratic unbiased estimators in balanced data and the best unbiased estimators if the data are balanced and normally distributed (Graybill and Hultquist 1961). Indeed, the ML estimates of the variance components are generally biased downwards, sometime severely so (Patterson and Thompson 1974; Corbeil and Searle 1976). The problem can be eliminated by using the Restricted Maximum Likelihood (REML) method proposed by Thompson (1962) for balanced data and generalized to unbalanced data by Patterson and Thompson (1971). The REML method is also called Residual Maximum Likelihood method because it actually uses the residuals to estimate the parameters $\nu$.

The idea of the Restricted Maximum Likelihood Estimate is to factor the likelihood into two parts, one of which does not depend on $\beta$, but only on $\nu$. We call this part the REML likelihood. The other part is a function of both $\beta$ and $\nu$ and we can use this part to get an estimator of $\beta$ that maximizes the likelihood function for fixed $\nu$.

For model (3), the REML likelihood is actually the likelihood with respect to $\nu$ based on a linearly transformed set of data $Z_1 = BY$ whose distribution does not depend on $\beta$, where the matrix $B$ is $(n - p) \times n$ with rank $n - p$, where $p$ is the rank of $X$ and the dimension of $\beta$.

To describe the factorization of the likelihood, let $B$ denote an $(n - p) \times n$ matrix satisfying

$$BB' = I_{n-p} \quad \text{and} \quad B'B = I_n - X(X'X)^{-1}X';$$

then

$$P = \left( \begin{array}{c} B \\ (X'V^{-1}X)^{-1}X'V^{-1} \end{array} \right)$$
is an \( n \times n \) matrix. It is easy to prove that this \( B \) exists and is orthogonal to \((X'V^{-1}X)^{-1}X'V^{-1}\), and that \( P \) is rank \( n \) (Hocking 1996). Let \( Z = PY \), a linear transformation of \( Y \), and let

\[
Z_1 = BY \sim N(0, BB'),
\]

\[
Z_2 = (X'V^{-1}X)^{-1}X'V^{-1}Y \sim N(\beta, (X'V^{-1}X)^{-1}),
\]

where \( Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \) and \( Z_1 \) and \( Z_2 \) are independent (Hocking 1996). The likelihood with respect to \( \beta \) and \( \nu \) based on \( Z \) is proportional to the likelihood based on \( Y \) because the Jacobian of the transformation can be reduced to \(|X'X|^{1/2}\), which does not depend on any parameters in the model (Harville, 1974). By the likelihood principle, the two likelihood functions are the same.

\( Z_1 \) is a linear transformation of \( Y \) and independent of \( Z_2 \). The likelihood \( L_1(\nu) \) with respect to \( \nu \) based on \( Z_1 \) is a factor of the likelihood based on \( Y \) and is independent of \( \beta \). It is the Restricted (or residual or REML) likelihood. In particular, it can be shown that its log is

\[
l_1(\nu) = \log L_1(\nu) = -\frac{1}{2} \log |det(V)| - \frac{1}{2} \log |det(X'V^{-1}X)| - \frac{1}{2} (Y - X\hat{\beta})'V^{-1}(Y - X\hat{\beta}),
\]

where \( \hat{\beta} = Z_2(\nu) \) (Harville 1974). Maximizing \( L_1(\nu) \) or \( l_1(\nu) \) gives the REML estimator \( \hat{\nu} \) of \( \nu \), and \( Z_2(\hat{\nu}) \) is the estimate of \( \beta \).

Usually, \( B \) is not unique, but it can be shown that for any \( B \) we get the same REML likelihood as long as \( B \) makes \( E(Z_1) = 0 \) and \( Z_1 \) independent of \( Z_2 \) (Diggle, Liang and Zeger 1994). Harville (1974) also gave a Bayesian interpretation of the REML likelihood. Suppose \( \beta \) is given a flat prior \( g(\beta) \propto 1 \). Then integrating \( \beta \) out from the posterior \( L(\beta, \nu)g(\beta) \), yields the REML likelihood \( L_1(\nu) \).

Das (1979) proved the consistency and asymptotic normality of the REML estimates under some conditions. In the balanced data case, the REML estimate is identical to the ANOVA estimate (Patterson and Thomason 1971). Therefore, the REML estimate is also the best
unbiased estimator if the data are balanced. Both the ANOVA estimate and the REML estimate
take account of the loss in degrees of freedom that results from estimating \( \beta \), but the MLE
does not. Das (1979) also showed that the maximum likelihood estimate and REML estimate
are asymptotically equivalent in the sense that the normalized difference of the corresponding
estimates converges to zero in probability if the dimension \( p \) of \( \beta \) is bounded.

Simulation studies have shown that the MLE compares quite favorably if Mean Squared
Error (MSE) is the measure of performance (Hocking and Kutner 1975; Corbeil and Searle
1976). But REML estimators are generally preferable if a large or moderately large number of
degrees of freedom are required for fixed effects (Robinson 1987). And one of Tunicliffe-Wilson’s
(1989) examples shows very clearly how REML copes much more effectively with a near-singular
variance matrix than does MLE.

2.3 Bayes Estimate

In the Bayesian framework, both \( \beta \) and \( \nu \) are regarded as random variables and inferences
are made using their posterior distributions. The most important early work using Bayesian
methods for linear models was by Tiao and Box (1967), Tiao and Tan (1965, 1966) and Hill
(1965, 1967). Lindley and Smith (1972) systematically illustrated the Bayes approach for linear
models.

In model (2), suppose \( \beta \) and \( \nu \) are given the prior \( \pi(\beta, \nu) \)—usually we assume \( \beta \) and \( \nu \) are
independent with priors \( \pi_1(\beta) \) and \( \pi_2(\nu) \), respectively—then the Bayes hierarchical linear model
is as following:
\[
Y|\theta, \nu \sim N(C\theta, R)
\]
\[
\theta|\beta, \nu \sim N(A\beta, D)
\]  \hspace{1cm} (5)
\[
\beta \sim \pi_1(\beta)
\]
\[
\nu \sim \pi_2(\nu),
\]
and the posterior of \((\beta, \theta, \nu)\) given \(Y\) is

\[
p(\beta, \theta, \nu|Y) \propto f(Y|\theta, \nu) \cdot f(\theta|\beta, \nu) \cdot \pi_1(\beta) \cdot \pi_2(\nu),
\]
where \(f\) is the appropriate multinormal density.

If we are just interested in \(\beta\) and \(\nu\), we can consider the marginal posterior of \((\beta, \nu)\)

\[
p(\beta, \nu|Y) = \int p(\beta, \theta, \nu|Y)d\theta,
\]
where \(\theta\) is usually a vector and the integral is a multiple integral.

When the loss function is proportional to squared error, the Bayes estimator of a parameter that minimizes Bayes risk is the parameter’s posterior mean. However, in all but fairly simple cases, the posterior mean estimate is impossible to derive in closed form (Harville 1977). It can be shown that the practice of using posterior means may yield very inefficient estimators, especially when using improper priors (Klotz, Milton and Zacks 1969). The inefficiency appears to be caused by posteriors with heavy tails resulting from the improper prior distributions. Other characteristics of the posterior distributions such as the mode, which is insensitive to the tails of the posterior density, can give more efficient point estimators (Klotz, Milton and Zacks 1969). We can use either the mode of the marginal posterior density of a parameter or the relevant component of the mode of the joint posterior density of that parameter and
various other parameters. It seems preferable to use the marginal posterior mode (O’Hagan 1976, Harville 1977).
3 The Modality of the Likelihood and Restricted Likelihood for the One-way Random Effect Model

Consider the simplest hierarchical linear model, the one-way random effect linear model with balanced data:

\[ y_{ij} | \theta_i \sim N(\theta_i, \sigma^2), \quad i = 1, 2, \cdots, N, \quad j = 1, 2, \cdots, n, \]

\[ \theta_i \sim N(\mu, \tau^2), \quad i = 1, 2, \cdots, N, \quad (6) \]

where \( \mu, \theta_i, \sigma^2 \) and \( \tau^2 \) are unknown parameters. In terms of model (2), \( A = J_N, \beta = \mu, \theta = (\theta_1, \theta_2, \cdots, \theta_N), R = \sigma^2 I_{NN}, D = \tau^2 I_N \) and

\[ C = \begin{pmatrix} J_n & J_n & \cdots & J_n \end{pmatrix}_{N \times N}. \]

Then the marginal distribution of the \( y_{ij} \) is

\[ Y \sim N(\mu J_{NN}, V(\sigma^2, \tau^2)), \]

where \( Y = (y_{11}, y_{12}, \cdots, y_{NN})^T \), \( V(\sigma^2, \tau^2) \) is block-diagonal with \( n \times n \) blocks \( V_1 = \sigma^2 I_n + \tau^2 U_n \), \( I_k \) is a \( k \times k \) identity matrix and \( J_k \) is a \( k \times 1 \) matrix with elements 1. It is easy to show that

\[ |V| = (\sigma^2 + n\tau^2)\sigma^2(n-1) \text{ and } V^{-1} = \frac{1}{\sigma^2} I_n - \frac{\tau^2}{\sigma^2(\tau^2 + n\sigma^2)} U_n. \]

3.1 The Maximum Likelihood Estimate

The (marginal) likelihood with respect to \( (\mu, \sigma^2, \tau^2) \) is

\[
L(\mu, \sigma^2, \tau^2 | Y) = \frac{1}{(2\pi)^{Nn/2}|V|^{N/2}} \exp \left( -\frac{1}{2} \sum_{i=1}^{N} (Y_i - \mu J_n)' V^{-1} (Y_i - \mu J_n) \right) \\
= \frac{1}{(2\pi)^{Nn/2}[\sigma^2 + n\tau^2]\sigma^2(n-1)]^{N/2} \]
\[ \cdot \exp \left( -\frac{1}{2\sigma^2} \left[ S^2_W + \frac{\sigma^2}{n^2 + \tau^2} S^2_B + \frac{\sigma^2}{n^2 + \tau^2} Nn(\bar{y}_. - \mu)^2 \right] \right) \] (7)

We can prove that \( L(\mu, \sigma^2, \tau^2; Y) \) has only one mode and that the MLE of \((\mu, \sigma^2, \tau^2)\) is

\[
\begin{align*}
\hat{\mu} &= \bar{y}_. \\
\hat{\sigma}^2 &= \frac{1}{N(n-1)} S^2_W \\
\hat{\tau}^2 &= \frac{1}{N} S^2_B - \frac{1}{nN(n-1)} S^2_W
\end{align*}
\]

where \( S^2_W = \sum_{i=1}^{N} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 \) and \( S^2_B = n \sum_{i=1}^{N} (\bar{y}_i - \bar{y}_.)^2 \).

When \( S^2_B < S^2_W / (n - 1) \), i.e. \( \hat{\tau}^2 < 0 \), we usually use

\[
\begin{align*}
\hat{\sigma}^2 &= \frac{1}{N} S^2_W \\
\hat{\tau}^2 &= 0
\end{align*}
\]

as the maximum likelihood estimate of \((\sigma^2, \tau^2)\) (Searle, casella and McCulloch 1992).

### 3.2 The REML Estimate

Integrating \( \mu \) out from likelihood (7), we get the REML likelihood

\[ L_1(\sigma^2, \tau^2) = K(\sigma^2, \tau^2) \cdot \exp \left( -\frac{1}{2} \left[ \frac{1}{\sigma^2} S^2_W + \frac{1}{\sigma^2 + n\tau^2} S^2_B \right] \right), \]

where

\[ K(\sigma^2, \tau^2) = \frac{\sqrt{\sigma^2 + n\tau^2}}{(2\pi)^{(Nn-1)/2} (\sigma^2 + n\tau^2)^{(N-1)/2}}. \]

We can prove that \( L_1(\sigma^2, \tau^2) \) has only one mode and that the REML estimator of \((\sigma^2, \tau^2)\) is

\[
\begin{align*}
\hat{\sigma}^2 &= \frac{1}{N(n-1)} S^2_W \\
\hat{\tau}^2 &= \frac{1}{N - \tau^2} S^2_B / n - \frac{1}{nN(n-1)} S^2_W
\end{align*}
\]

and \( \mu \) is also estimated by \( \bar{y}_. \). This is the same as the analysis-of-variance estimator for balanced data. Also, if \( \hat{\tau}^2 < 0 \), we use

\[
\begin{align*}
\hat{\sigma}^2 &= \frac{1}{N-n-1} S^2_W \\
\hat{\tau}^2 &= 0
\end{align*}
\]

as the REML estimate of \((\sigma^2, \tau^2)\) (Searle, casella and McCulloch 1992).
4 The Modality of the Joint Posterior for the One-Way Random Effect Model

In this section, we consider the Bayesian analysis of the one-way random effect model with conjugate priors. For a Bayesian analysis, every unknown parameter should have a prior. In model (6), give a flat prior to $\mu$ and Inverse Gamma priors to $\sigma^2$ and $\tau^2$; then the model becomes

\[ y_{ij} \sim N(\theta_i, \sigma^2), i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, n; \]
\[ \theta_i \sim N(\mu, \tau^2), i = 1, 2, \ldots, N; \]
\[ \sigma^2 \sim IG(\alpha, \lambda); \]
\[ \tau^2 \sim IG(a, b); \]
\[ \pi(\mu) \propto 1, \]

where $\alpha > 0$, $\lambda > 0$, $a > 0$, $b > 0$ are given, and the inverse gammas are parameterized so that $E(\sigma^2) = \frac{\lambda}{\alpha-1}$, $E(\tau^2) = \frac{b}{a-1}$.

We begin by completely characterizing the modality of the joint posterior in sections 4.1 and 4.2. Section 4.3 gives a more intuitive discussion of sections 4.1 and 4.2, and section 4.4 gives some examples.

4.1 The Modality of the Joint Posterior

The joint posterior of $(\theta_1, \theta_2, \ldots, \theta_N, \mu, \sigma^2, \tau^2)$ corresponding to model (8) is

\[
p(\theta_1, \theta_2, \ldots, \theta_N, \mu, \sigma^2, \tau^2 | Y) \propto \frac{1}{(\sigma^2)^{\frac{nN}{2}}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{n} (y_{ij} - \theta_i)^2 \right] \frac{1}{(\tau^2)^{\frac{N}{2}}} \exp \left[ -\frac{1}{2\tau^2} \sum_{i=1}^{N} (\theta_i - \mu)^2 \right] \\
\cdot \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp(-\frac{\lambda \sigma^2}{\sigma^2}) \frac{1}{(\tau^2)^{\frac{n}{2}}} \exp(-\frac{b \tau^2}{\tau^2}) \\
= \frac{1}{(\sigma^2)^{\frac{nN}{2} + \alpha + 1}} \exp \left\{ -\frac{1}{2\sigma^2} \left[ 2\lambda + S^2_W + n \sum_{i=1}^{N} (\bar{y}_i - \theta_i)^2 \right] \right\}
\]
\begin{equation}
\frac{1}{(\tau^2)^{a+1}} \exp \left\{ -\frac{1}{2\tau^2} \left[ 2b + \sum_{i=1}^{N} (\theta_i - \mu)^2 \right] \right\} \tag{9}
\end{equation}

Let's see how many modes this posterior has.

Let

\[
\ell(\theta_1, \ldots, \theta_N, \mu, \sigma^2, \tau^2 | Y) = \log(p) = -\left( \frac{N}{2} + \alpha + 1 \right) \log \sigma^2 - \frac{1}{2\sigma^2} \left[ 2\lambda + S_W^2 + n \sum_{i=1}^{N} (\bar{y}_i - \theta_i)^2 \right]
\]

\[
-\left( \frac{N}{2} + a + 1 \right) \log \tau^2 - \frac{1}{2\tau^2} \left[ 2b + \sum_{i=1}^{N} (\theta_i - \mu)^2 \right] + c \tag{10}
\]

be the log posterior, we simply denote it by \( \ell \); where \( c \) is a constant not related to the parameters.

Then \( \ell \) and the posterior (9) have the same modes. From now, we mostly discuss the modality of \( \ell \). Consider

\[
\frac{\partial \ell}{\partial \theta_i} = \frac{n}{\sigma^2}(\bar{y}_i - \theta_i) - \frac{1}{\tau^2}(\theta_i - \mu) = \frac{n}{\sigma^2}\bar{y}_i + \frac{1}{\tau^2}\mu - \frac{n\tau^2 + \sigma^2}{\sigma^2\tau^2}\theta_i, \quad i = 1, 2, \ldots, N
\]

\[
\frac{\partial \ell}{\partial \mu} = \frac{1}{\tau^2} \sum_{i=1}^{N} (\theta_i - \mu).
\]

Let \( \frac{\partial \ell}{\partial \theta_i} = 0 \) and \( \frac{\partial \ell}{\partial \mu} = 0 \); then we get the only critical point \((\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N, \hat{\mu})\) of \( \ell \) given \( \sigma^2 \) and \( \tau^2 \), namely,

\[
\hat{\mu} = \bar{y}.
\]

\[
\hat{\theta}_i(\sigma^2, \tau^2) = \frac{n\tau^2}{n\tau^2 + \sigma^2}\bar{y}_i + \frac{\sigma^2}{n\tau^2 + \sigma^2}\bar{y}_i, \quad i = 1, 2, \ldots, N \tag{11}
\]

**Lemma 1** Given \( \sigma^2 > 0 \) and \( \tau^2 > 0 \), \((\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N, \hat{\mu})\) maximizes the log posterior \( \ell \).

For the proof, see Appendix A.

Replacing \((\theta_1, \theta_2, \cdots, \theta_N, \mu)\) by \((\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_N, \hat{\mu})\) in (10), we get

\[
\ell_1(\sigma^2, \tau^2) = -\left( \frac{Nn}{2} + \alpha + 1 \right) \log \sigma^2 - \frac{1}{2\sigma^2} \left[ 2\lambda + S_W^2 + n \sum_{i=1}^{N} (\bar{y}_i - \hat{\theta}_i)^2 \right]
\]
\[-\frac{N}{2} + a + 1) \log \tau^2 - \frac{1}{2\tau^2} \left[ 2b + \sum_{i=1}^{N} (\theta_i - \bar{\theta})^2 \right] \]

\[= -\frac{Nn}{2} + \alpha + 1) \log \sigma^2 - \frac{1}{2\sigma^2} \left[ 2\lambda + S_{W}^2 + \left( \frac{\sigma^2}{\nu \tau^2 + \sigma^2} \right)^2 S_{B}^2 \right] \]

\[-\frac{N}{2} + a + 1) \log \tau^2 - \frac{1}{2\tau^2} \left[ 2b + \left( \frac{\nu \tau^2}{\sigma^2 + \nu \tau^2} \right)^2 S_{B}^2 / n \right] , \]

the profile log posterior with respect to \( \sigma^2 \) and \( \tau^2 \), where \( S_{W}^2 \) and \( S_{B}^2 \) are the same as in section

3. If \((\sigma^2, \tau^2)\) maximizes \(\ell_1(\sigma^2, \tau^2)\), then

\( \hat{\theta}_1(\sigma^2, \tau^2), \hat{\theta}_2(\sigma^2, \tau^2), \ldots, \hat{\theta}_N(\sigma^2, \tau^2), \hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2 \)

maximizes \(\ell\). Conversely, if \((\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N, \hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2)\) maximizes \(\ell\), then \((\sigma^2, \tau^2)\) is the mode of \(\ell_1(\sigma^2, \tau^2)\) and satisfies (11).

Now, we maximize \(\ell_1(\sigma^2, \tau^2)\) with the restriction \(\frac{\sigma^2}{\nu \tau^2 + \sigma^2} = \phi\), where \(\phi\) is in \((0, 1)\); we borrow this technique from O'Hagan (1985). This is equivalent to maximizing

\[\ell_2(\sigma^2, \tau^2, \phi) = -\left(\frac{Nn}{2} + \alpha + 1\right) \log \sigma^2 - \frac{1}{2\sigma^2} \left[ 2\lambda + S_{W}^2 + \phi^2 S_{B}^2 \right] \]

\[-\frac{N}{2} + a + 1) \log \tau^2 - \frac{1}{2\tau^2} \left[ 2b + (1 - \phi)^2 S_{B}^2 / n \right] \]

with the restriction \(\frac{\sigma^2}{\nu \tau^2 + \sigma^2} = \phi\). From now, we simply denote \(\ell_1(\sigma^2, \tau^2)\) and \(\ell_2(\sigma^2, \tau^2, \phi)\) by \(\ell_1\) and \(\ell_2\), respectively.

We use Lagrange’s method to find the critical points of \(\ell_2\) with the restriction \(\frac{\sigma^2}{\nu \tau^2 + \sigma^2} = \phi\). Since \(\phi\) is a free variable, maximizing \(\ell_2\) with respect to \((\sigma^2, \tau^2, \phi)\) under the restriction is equivalent to maximizing \(\ell_1\) with respect to \((\sigma^2, \tau^2)\) without the restriction. Let \(G(\phi, \sigma^2, \tau^2) = \frac{\sigma^2}{\nu \tau^2 + \sigma^2} - \phi\), and consider solving

\[\frac{\partial \ell_2}{\partial \sigma^2} + \zeta \frac{\partial G}{\partial \sigma^2} = 0 \]

\[\frac{\partial \ell_2}{\partial \tau^2} + \zeta \frac{\partial G}{\partial \tau^2} = 0 \]

\[\frac{\partial \ell_2}{\partial \phi} + \zeta \frac{\partial G}{\partial \phi} = 0 \]

(13)
with the restriction \( G(\phi, \sigma^2, \tau^2) = 0 \).

Because \( \frac{\partial \ell_1}{\partial \phi} \equiv 0 \) with the restriction and \( \frac{\partial G}{\partial \phi} = -1 \), we get \( \zeta = 0 \). So we only need to consider

\[
\begin{cases}
\frac{\partial \ell_2}{\partial \sigma^2} = 0 \\
\frac{\partial \ell_2}{\partial \tau^2} = 0 \\
G = 0,
\end{cases}
\]

where

\[
\frac{\partial \ell_2}{\partial \sigma^2} = -\frac{1}{\sigma^2} \left( \frac{Nn}{2} + \alpha + 1 \right) + \frac{1}{2\alpha} [2\lambda + S_W^2 + \phi^2 S_B^2] \\
\frac{\partial \ell_2}{\partial \tau^2} = -\frac{1}{\tau^2} \left( \frac{N}{2} + a + 1 \right) + \frac{1}{2\tau} [2b + (1 - \phi)^2 S_B^2 / n].
\]

Letting \( \frac{\partial \ell_2}{\partial \sigma^2} = \frac{\partial \ell_2}{\partial \tau^2} = 0 \), we get

\[
\sigma^2 = \frac{1}{Nn + 2\alpha + 2} \left[ 2\lambda + S_W^2 + \phi^2 S_B^2 \right] \\
\tau^2 = \frac{1}{N + 2a + 2} \left[ 2b + (1 - \phi)^2 S_B^2 / n \right].
\]

Denote \( Nn + 2\alpha + 2 \) and \( N + 2a + 2 \) by \( e \) and \( f \), respectively; from (15), (16) and \( G = 0 \), that is \( \phi = \frac{\sigma^2}{n\tau^2 + \sigma^2} \), we have

\[
\frac{1}{e} \left[ 2\lambda + S_W^2 + \phi^2 S_B^2 \right] = \frac{n\phi}{f} \left[ 2b + (1 - \phi)^2 S_B^2 / n \right] + \frac{\phi}{e} \left[ 2\lambda + S_W^2 + \phi^2 S_B^2 \right].
\]

Simplifying (17) yields

\[
(e + f) S_B^2 \phi^3 - S_B^2 (2e + f) \phi^2 + (2\lambda f + f S_W^2 + 2neb + eS_B^2) \phi - f(2\lambda + S_W^2) = 0,
\]

a cubic polynomial in \( \phi \). Denote the left part of (18) by \( f(\phi) \); then \( f'(\phi) = 3(e + f) S_B^2 \phi^2 - 2(2e + f) S_B^2 \phi + [(2\lambda + S_W^2) f + 2neb + eS_B^2] \). \( f(\phi) = 0 \) has at most three real solutions. Therefore, the posterior has at most three modes.

Notice that
(a) \( f(0) = -f(2\lambda + S_B^2) < 0; \)

(b) \( f(1) = 2nеб > 0; \)

(c) \( f'(0) = (2\lambda + S_W^2)f + 2nеб + eS_B^2 > 0; \)

(d) \( f'(1) = fS_B^2 + (2\lambda + S_W^2)f + 2nеб > 0; \)

(e) If \( \Delta > 0 \), then \( f'() = 0 \) has two solutions \( \phi_1 \) and \( \phi_2 \), where

\[
\begin{align*}
\Delta &= 4(2e + f)^2 - 12(e + f) \left( \frac{2\lambda + 2nеб}{S_B^2} + \frac{f(N(n-1))}{(N(n-1))^2} + e \right) \\
\phi_1 &= \frac{2(2e+f) - \sqrt{\Delta}}{6(e+f)} \\
\phi_2 &= \frac{2(2e+f) + \sqrt{\Delta}}{6(e+f)}
\end{align*}
\]

and \( F = \frac{S_B^2/(N-1)}{S_W^2/(n-1)} \) is the usual F-statistic for the one-way balanced ANOVA. If \( \Delta = 0 \), \( f'() \) has a single solution \( \frac{2e+f}{5(e+f)} \).

Now we can use the geometric method to characterize the solutions of \( f() = 0 \). Notice that \( f() \) is a cubic polynomial and \( f'() \) is a quadratic polynomial with positive coefficient for \( \phi^2 \). From points (a), (b), (c), (d) and (e) above, we can draw these conclusions:

1. From points (a) and (b), there is at least one \( \hat{\phi} \) in \((0, 1)\) such that \( f(\hat{\phi}) = 0 \). Therefore, if \( f() \) has only one real solution \( \hat{\phi} \), then \( \hat{\phi} \in (0, 1) \).

[Figure 1: The plot of \( f() \) for \( \Delta < 0 \)]
2. If Δ ≥ 0, then 0 < φ₁ ≤ φ₂ < 1. To see this, notice that 0 < φ₁ < 1 by the definition (19) and φ₁ ≤ φ₂. Now f'(φ) < 0 for all φ ∈ (φ₁, φ₂) because f'(φ) is a quadratic polynomial and the coefficient of φ² is positive. If φ₂ > 1, then 1 ∈ (φ₁, φ₂) and f'(1) < 0, which contradicts point (d) above. Therefore, φ₂ < 1, i.e. 0 < φ₁ ≤ φ₂ < 1.

3. If Δ < 0, φ₁ and φ₂ are not real numbers and f'(φ) > 0 for all φ ∈ (0,1), i.e. f(φ) is monotone. Thus f(φ) = 0 has only one solution ̂φ in (0, 1) by points (a) and (b) above. In this case, the plot of f(φ) looks like the plot in Figure 1.

![Plot of f(φ) for Δ = 0](image)

Figure 2: The plot of f(φ) for Δ = 0

4. If Δ = 0, then φ₁ = φ₂, i.e. f'(φ) > 0 for all φ ∈ (0,1) except at φ₁ = φ₂, at which f'(φ₁) = 0. Therefore, f(φ) = 0 has only one real solution ̂φ in (0, 1) because f(φ) is a
cubic polynomial. In this case, the plot of $f(\phi)$ looks like one of plots in Figure 2.

5. If $\Delta > 0$, then $\phi_1 < \phi_2$. There are three cases as follows.

![Figure 3: The plot of $f(\phi)$ for $\Delta > 0$](image)

(a) If $f(\phi_1) \cdot f(\phi_2) > 0$, then $f(\phi_1)$ and $f(\phi_2)$ are on the same side of the line $f = 0$, therefore, $f(\phi) = 0$ has only one real solution $\hat{\phi}$ in $(0, 1)$. In this case, the plot of
\( f(\phi) \) looks like (a) or (b) in Figure 3.

(b) If \( f(\phi_1) \cdot f(\phi_2) = 0 \), then either \( f(\phi_1) = 0 \) or \( f(\phi_2) = 0 \), but not both because \( f'(\phi) < 0 \) for all \( \phi \in (\phi_1, \phi_2) \). Therefore, \( f(\phi) = 0 \) has two different real solutions. Actually, by the properties of polynomials, \( f(\phi) = 0 \) has three real solutions \( \hat{\phi}^{(1)}, \hat{\phi}^{(2)}, \hat{\phi}^{(3)} \) satisfying \( 0 < \hat{\phi}^{(1)} = \phi_1 < \hat{\phi}^{(2)} < \phi_2 < \hat{\phi}^{(3)} < 1 \) or \( 0 < \hat{\phi}^{(1)} < \phi_1 < \hat{\phi}^{(2)} = \phi_2 = \hat{\phi}^{(3)} < 1 \), i.e. two of them are the same and the third is different. In this case, the plot of \( f(\phi) \) looks like (c) or (d) in Figure 3.

(c) Finally, if \( f(\phi_1) \cdot f(\phi_2) < 0 \), we must have \( f(\phi_1) > 0 \) and \( f(\phi_2) < 0 \). Because \( 0 < \phi_1 < \phi_2 < 1 \) and \( f(0) < 0 \), \( f(\phi_1) > 0 \), \( f(\phi_2) < 0 \) and \( f(1) > 0 \), it follows that \( f(\phi) = 0 \) has three real solutions \( \hat{\phi}^{(1)} < \hat{\phi}^{(2)} < \hat{\phi}^{(3)} \) satisfying \( 0 < \hat{\phi}^{(1)} < \phi_1 < \hat{\phi}^{(2)} < \phi_2 < \hat{\phi}^{(3)} < 1 \). The plot of \( f(\phi) \) looks like (e) in Figure 3.

From the above, we know that all the real solutions of \( f(\phi) = 0 \) are in \((0, 1)\).

Define

\[
q = K_1 + K_2 \frac{2\lambda + S_W^2}{S_B^2} + K_3 \frac{b}{S_B^2},
\]

\[
p = L_1 + L_2 \frac{2\lambda + S_W^2}{S_B^2} + L_3 \frac{b}{S_B^2},
\]

where

\[
K_1 = \frac{2e^3 + 3e^2 f - 3ef^2 - 2f^3}{27(e + f)^3}, \quad K_2 = \frac{(2 + 2f)f}{3(e + f)^2}, \quad K_3 = \frac{2(2e + f)ne}{3(e + f)^2}
\]

and

\[
L_1 = \frac{e^2 + ef + f^2}{3(e + f)^2}, \quad L_2 = \frac{2ne}{e + f}, \quad L_3 = \frac{f}{e + f}.
\]

Notice that the \( K' \)s and \( L' \)s depend only on \( N, n, a \) and \( \alpha \). Then we can prove the following theorem:
Theorem 4.1 For model (8), if $\Delta > 0$ and $f(\phi_1) \cdot f(\phi_2) < 0$ then the log posterior $\ell$ has two modes; otherwise, $\ell$ has only one mode, where $\Delta$, $\phi_1$ and $\phi_2$ are defined in (19). Specifically:

1. If $\Delta > 0$ and $f(\phi_1)f(\phi_2) < 0$, then $\ell_1$ has three critical points $(\hat{\sigma}_k^2, \hat{\tau}_k^2)$, $k = 1, 2, 3$, two of which maximize $\ell_1$, where

\[
\hat{\sigma}_k^2 = \frac{1}{Nn + 2a + 2} \left[ 2\lambda + S_W^2 + (\hat{\phi}^{(k)})^2 S_B^2 \right], \quad k = 1, 2, 3
\]

\[
\hat{\tau}_k^2 = \frac{1}{N + 2a + 2} \left[ 2b + (1 - \hat{\phi}^{(k)})^2 S_B^2 / n \right], \quad k = 1, 2, 3
\]

\[
\hat{\phi}^{(1)} = -\sqrt{\frac{-p}{3}} \left[ \cos \psi + \sqrt{3} \sin \psi \right] + \frac{2e + f}{3(e + f)}
\]

\[
\hat{\phi}^{(2)} = -\sqrt{\frac{-p}{3}} \left[ \cos \psi - \sqrt{3} \sin \psi \right] + \frac{2e + f}{3(e + f)}
\]

\[
\hat{\phi}^{(3)} = \frac{2}{3} \sqrt{\frac{-p}{3}} \cos \psi + \frac{2e + f}{3(e + f)};
\]

and $(\hat{\sigma}_1^2, \hat{\tau}_1^2)$ and $(\hat{\sigma}_3^2, \hat{\tau}_3^2)$ maximize $\ell_1$, where $\psi = \frac{1}{3} \arccos \left( \frac{3a}{2p} \sqrt{\frac{-3}{p}} \right)$.

2. If $\Delta > 0$ and $f(\phi_1)f(\phi_2) = 0$, then $\ell_1$ has two critical points $(\tilde{\sigma}_k^2, \tilde{\tau}_k^2)$, $k = 1, 2$, where $\tilde{\sigma}_k^2$ and $\tilde{\tau}_k^2$ are defined as above, $\{\hat{\phi}^{(1)}, \hat{\phi}^{(2)}\} = \{-2\sqrt{q/2} + 2e + f \sqrt{q/3}, \sqrt{q/2} + 2e + f \sqrt{q/3}\}$ as a set and only $(\tilde{\sigma}_2^2, \tilde{\tau}_2^2)$ corresponding to $\hat{\phi} = -2\sqrt{q/2} + \frac{2e + f}{3(e + f)}$ maximizes $\ell_1$.

3. If $\Delta \leq 0$ or $\Delta > 0$ with $f(\phi_1)f(\phi_2) > 0$, then $\ell_1$ has only one critical point $(\check{\sigma}^2, \check{\tau}^2)$ that maximizes $\ell_1$, where

\[
\check{\sigma}^2 = \frac{1}{Nn + 2a + 2} \left[ 2\lambda + S_W^2 + \hat{\phi}^2 S_B^2 \right]
\]

\[
\check{\tau}^2 = \frac{1}{N + 2a + 2} \left[ 2b + (1 - \hat{\phi}^2) S_B^2 / n \right]
\]

\[
\check{\phi} = \sqrt{\frac{-q}{2}} + \sqrt{\left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3} + \sqrt{\frac{-q}{2} - \sqrt{\left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3} + \frac{2e + f}{3(e + f)}}.
\]

If $(\check{\sigma}_k^2, \check{\tau}_k^2)$ maximizes $\ell_1$, then $(\hat{\phi}_1^{(k)}, \hat{\phi}_2^{(k)}, \cdots, \hat{\phi}_N^{(k)}, \hat{\mu}, \check{\sigma}_k^2, \check{\tau}_k^2)$ is a mode of $\ell$ and all the modes of $\ell$ have this form, where

\[
\hat{\mu} = \bar{y},
\]

\[
\hat{\phi}_i^{(k)} = (1 - \hat{\phi}^{(k)}) \bar{y}_i + \hat{\phi}^{(k)} \bar{y}_i, i = 1, 2, \cdots, N.
\]

For the proof, see Appendix B.
4.2 One-Way Random Effect Model with $\tau^2$ Having an Improper Prior

Improper priors are common in Bayesian analysis because sometimes we do not have an exact prior and in some cases the improper prior makes the Bayes approach have good frequentist properties.

**Theorem 4.2** If $\tau^2$ has an improper prior $f(\tau^2) \propto \frac{1}{\tau^{\alpha+1}}$ in model (8), where $\alpha \geq -1$, then $\ell$ has a spine: $\ell$ goes to infinity at all points

$$(\theta_1, \ldots, \theta_N, \mu, \sigma^2, \tau^2) = (\theta, \ldots, \theta, \theta, \sigma^2, 0)$$

for arbitrary $\theta$ and $\sigma^2 > 0$. This spine covers a two-dimensional subspace of the $(N + 3)$-dimensional parameter space. If $\Delta_1 = e^2 - 4(e + N)N\left(\frac{2\lambda}{S_\lambda^2} + \frac{N(n-1)}{N-1}\phi\right) \geq 0$ then $\ell$ also has a finite mode at $(\hat{\theta}_1, \ldots, \hat{\theta}_N, \hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2)$, where

$$\hat{\mu} = \bar{y},$$
$$\hat{\theta}_i = \hat{\phi}\bar{y}_i + (1 - \hat{\phi})\bar{y}_i, \quad i = 1, 2, \ldots, N$$
$$\hat{\sigma}^2 = \left[2\lambda + S_W^2 + \hat{\phi}S_B^2\right]/e$$
$$\hat{\tau}^2 = (1 - \hat{\phi})^2 S_B^2/ef$$
$$\hat{\phi} = (e - \sqrt{\Delta_1})/2(e + f)$$

and $e = nN + 2\alpha + 2$, $f = N + 2a + 2$.

The proof is the similar to Theorem 4.1; see Appendix C.

When $\alpha = \alpha = -1$ and $\lambda = 0$, the theorem reduces to the case of O'Hagan (1985). Actually, this theorem is meaningful only when $\alpha < 0$ or $\lambda > 0$ because the posterior corresponding to model (8) is improper if $a \geq 0$ when $b = 0$ and $\lambda = 0$ (Hobert and Casella 1996). (It is nonsense to discuss the modality of the posterior if the posterior is improper.)
4.3 Intuitive Explanation of Theorem 4.1; Sensitivity Analysis of Modes

From Theorem 4.1, when $\Delta \leq 0$, $\ell$ has only one mode. When $\Delta > 0$, $\ell$ may have two modes. When $\lambda$ or $b$ is large, $\Delta$ is relatively small. When $S_B^2$ is large, $\Delta$ is relatively large. Therefore:

1. When $b$ or $\lambda$ is large for fixed $a$ and $\alpha$, it is more likely that $\ell$ has only one mode.

2. When $S_B^2$ is large relative to $S_W^2$, it is more likely that $\ell$ has two modes.

We can intuitively explain this as follows. The information about $\sigma^2$ and $\tau^2$ in the posterior comes from both the data and the priors. If $S_B^2$ is very large relative to $S_W^2$, the data have strong information that $\tau^2$ is large relative to $\sigma^2$, while if $S_W^2$ is very large relative to $S_B^2$, the data have strong information that $\sigma^2$ is large relative to $\tau^2$. For the priors $\sigma^2 \sim IG(\alpha, \lambda)$ and $\tau^2 \sim IG(a, b)$, $E(\sigma^2) = \frac{\lambda}{\alpha - 1}$, $Var(\sigma^2) = \frac{\lambda^2}{(\alpha - 1)^2(\alpha - 2)}$, and $E(\tau^2) = \frac{b}{\alpha - 1}$, $Var(\tau^2) = \frac{b^2}{(\alpha - 1)^2(\alpha - 2)}$.

Because these priors are conjugate, we can interpret $\alpha$ and $a$ as the number of observations that the respective priors are “worth”. If $\lambda$ is small or $\alpha$ is large, the prior provides strong information that $\sigma^2$ takes a small value. If $\lambda$ is large or $\alpha$ is small, the prior suggests a large $\sigma^2$ but with considerable uncertainty. Similar facts hold for $a$, $b$ and $\tau^2$.

Suppose the priors and the data have different information about $\sigma^2$ and $\tau^2$. If both have strong information, the posterior is more likely bimodal. If one has much stronger information than the other, then the posterior is more likely to be unimodal with the mode being determined by the one that has stronger information.

We now summarize a variety of particular cases. For this purpose, we use $\phi = \frac{\sigma^2}{n\tau^2 + \sigma^2}$; when $\sigma^2$ is relatively small, $\phi$ is close to 0; when $\tau^2$ is relatively small, $\phi$ is close to 1. From the above discussion we have the following observations:
1. If both $\alpha$ and $a$ are large, $\ell$ is more likely to have two modes; $\phi$ is close to 0 for one and $\phi$ is close to 1 for the other.

In this case the priors have strong information that both $\sigma^2$ and $\tau^2$ are small. Since $\sigma^2 + \tau^2$ is the variance of the data, when the priors force $\sigma^2$ and $\tau^2$ to be too small, the posterior has to have two modes, one with small $\tau^2$ and large $\sigma^2$, and one with small $\sigma^2$ and large $\tau^2$, i.e., one with $\phi$ near to 1 and one with $\phi$ close to 0.

2. If $\alpha$ is small and $a$ is large, $\ell$ is more likely to have one mode, for which $\phi$ is close to 1.

When $\alpha$ is very small, the prior has strong information that $\tau^2$ is small and it dominates the posterior. The posterior is more likely to have only one mode, which has a small $\tau^2$.

3. If $\alpha$ is large and $a$ is small, $\ell$ is more likely to have one mode, for which $\phi$ is close to 0.

This is similar to case 2.

4. If both $\alpha$ and $a$ are small, $\ell$ is more likely to have one mode. Whether $\phi$ is close to 0 or close to 1 or in the middle depends on $S_B^2$ and $S_W^2$.

In this case, the priors do not provide very strong information and it is more likely that the data dominate the posterior; therefore, the posterior is unimodal with the mode indicated by data.

5. If both $\lambda$ and $b$ are large, $\ell$ is more likely to have one mode. Whether $\phi$ is close to 0 or close to 1 or in the middle depends on $S_B^2$ and $S_W^2$.

6. If $\lambda$ is large and $b$ is small, $\ell$ is more likely to have one mode, for which $\phi$ is close to 1.

7. If $\lambda$ is small and $b$ is large, $\ell$ is more likely to have one mode, for which $\phi$ is close to 0.
8. If both $\lambda$ and $b$ are small, $\ell$ is more likely to have two modes; $\phi$ is close to 0 for one and $\phi$ is close to 1 for the other.

9. If $S_B^2$ is large relative to $S_W^2$, then $\ell$ is more likely to have two modes.

From the above discussion we see that the modality is sensitive to both the data and the parameters in the priors. The above is an intuitive explanation about the relationship among the modality, the prior parameters and the data. The actual relationship is somewhat more complicated.

When $\lambda$ or $b$ gets large, $\Delta$ becomes negative, so $\ell$ will be unimodal by Theorem 4.1. Actually, even before $\Delta$ becomes non-positive, we can make one mode disappear by increasing $b$ or $\lambda$. We can prove the following theorem.

**Theorem 4.3** For model (8) and the corresponding log posterior $\ell$, if $\ell$ has two modes then

1. If $b$ or $\alpha$ increases and all else is fixed, the right mode (in $\phi$) moves left until it disappears (before $\Delta$ becomes non-negative); the other mode remains and also moves left;

2. If $\lambda$ or $\alpha$ increases and all else is fixed, the left mode (in $\phi$) moves right until it disappears (before $\Delta$ becomes non-negative); the right mode remains and also moves right.

3. For any fixed $a, b, \alpha, \lambda, n, N$ and $S_W^2$, $\ell$ can be made bimodal by increasing $S_B^2$.

The proof is in Appendix D.

From the theorem, we see that no choice of prior guarantees a unimodal posterior and the data can not be made to dominate the posterior by changing $S_B^2$. This theorem just shows the relationship between modality and parameter values. It is probably unwise to use this theorem to make the posterior unimodal.
4.4 Examples

In this subsection, we give several examples to illustrate Theorem 4.1 and Theorem 4.3. Substituting $\sigma^2$ and $\tau^2$ in (15) and (16) for $\sigma^2$ and $\tau^2$ in (12), we get a continuous function of $\phi$

$$
\ell(\phi) = -\frac{e}{2} [\log(2\lambda + S_B^2 + \phi^2 S_W^2) - \log(e) + 1] - \frac{f}{2} [\log(2b + (1 - \phi)^2 S_B^2/n) - \log(f) + 1].
$$

We can prove that $\ell(\phi)$ and $\ell_1$ have the same modes. A mode $\hat{\phi}$ of $\ell(\phi)$ gives $(\sigma^2, \tau^2)$ that maximize $\ell_1$, and if $(\sigma^2, \tau^2)$ maximizes $\ell_1$ then $\hat{\phi} = \sigma^2/(n\tau^2 + \sigma^2)$ is a mode of $\ell(\phi)$. That is, $\ell(\phi)$ and $\ell_1$ have the same modes with respect to $\phi$. Therefore, we use $\ell(\phi)$ to display the modality of the posterior.

**Example 1:** The effect of changing $\lambda$.

Suppose $N = 100$, $n = 2$, $S_B^2 = 21$, $S_W^2 = 1$, $a = 13$, $b = 0.2$, $\alpha = 3$, and $\lambda = 1$ in model (8). Then $\Delta = 1.056 \times 10^8 > 0$, $\phi_1 = 0.3$, $\phi_2 = 0.78$, $f(\phi_1) \cdot f(\phi_2) = 253.92 \times (-149.60) = -37986.43 < 0$. Thus $f(\phi)$ should have three solutions according to Theorem 4.1. Using the formulae in Theorem 4.1, we get $\hat{\phi}^{(1)} = 0.1$, $\hat{\phi}^{(2)} = 0.58$, $\hat{\phi}^{(3)} = 0.94$, $\ell(\hat{\phi}^{(1)}) = 436.40$, $\ell(\hat{\phi}^{(2)}) = 405.53$ and $\ell(\hat{\phi}^{(3)}) = 431.16$, where $\hat{\phi}^{(1)}$ and $\hat{\phi}^{(3)}$ are the two modes of $\ell(\phi)$. This is consistent with the plots of $f(\phi)$ and $\ell(\phi)$ in Figure 4 (a) and (e). Considering $\lambda = 2, 3$ and 10 separately, we get the values of $\Delta$, $\phi_1$, $\phi_2$, $f(\phi_1)$, $f(\phi_2)$, $\hat{\phi}^{(1)}$, $\hat{\phi}^{(2)}$, $\hat{\phi}^{(3)}$, $\ell(\hat{\phi}^{(1)})$, $\ell(\hat{\phi}^{(2)})$ and $\ell(\hat{\phi}^{(3)})$ shown in Table 2.

From Table 2 we see that when $\lambda = 1$ or 2, $\Delta > 0$ and $f(\phi_1) f(\phi_2) < 0$, so $f(\phi) = 0$ has three solutions and $\hat{\phi}^{(1)}$ and $\hat{\phi}^{(3)}$ are the maximizing points. When $\lambda$ increases, the modes move to the right, that is $\hat{\phi}^{(1)}$ and $\hat{\phi}^{(3)}$ get larger. For $\lambda$ large enough
Table 2: Values of $\Delta$, $\phi$ and $f(\phi)$ for different values of $\lambda$

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\Delta$</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$f(\phi_1)$</th>
<th>$f(\phi_2)$</th>
<th>$\phi^{(1)}$</th>
<th>$\phi^{(2)}$</th>
<th>$\phi^{(3)}$</th>
<th>$\ell(\phi^{(1)})$</th>
<th>$\ell(\phi^{(2)})$</th>
<th>$\ell(\phi^{(3)})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.056 x 10^8</td>
<td>0.30</td>
<td>0.78</td>
<td>-149.60</td>
<td>0.10</td>
<td>0.58</td>
<td>0.94</td>
<td>436.40</td>
<td>405.53</td>
<td>431.16</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8.39 x 10^7</td>
<td>0.32</td>
<td>0.76</td>
<td>-208.62</td>
<td>0.20</td>
<td>0.47</td>
<td>0.94</td>
<td>388.48</td>
<td>384.63</td>
<td>421.94</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6.223 x 10^7</td>
<td>0.35</td>
<td>0.73</td>
<td>-274.82</td>
<td>-</td>
<td>-</td>
<td>0.95</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-8.95 x 10^7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.97</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(for example, $\lambda \geq 3$), the left mode disappears and $\ell(\phi)$ just has one mode left. The plots of $f(\phi)$ and $\ell(\phi)$ on Figure 4 show this change. We also see that $\ell(\phi)$ loses one mode before $\Delta$ becomes negative (Figure 4 (c) and (g)).

Now we intuitively explain the results of this example. Consider the data: $N = 100$, $n = 2$, $S_H^2 = 21$, $S_W^2 = 1$, so $F = \frac{S_H^2/N-1}{S_W^2/(n-1)} = 21.21$, which would usually be considered evidence that the groups differ. The MLEs of $\sigma^2$ and $\tau^2$ are 0.01 and 0.1, respectively, that is, the data favour a relatively large $\tau^2$ and small $\sigma^2$. On the other hand, a priori $E(\tau^2) = \frac{\lambda}{\alpha-1} = 0.017$, $E(\sigma^2) = \frac{\lambda}{\alpha-1} = 0.5$ when $\lambda = 1$, which is opposite to the information in the data about $\sigma^2$ and $\tau^2$. Because $a = 13$, the prior for $\tau^2$ is “worth” 13 observations without error on $\theta$. The prior prefers relatively large $\sigma^2$, but the information is not strong. Because $N = 100$, $n = 2$, the data provide 100 observations on the $\theta_i$ with non-trivial error ($y_i \sim N(\mu, \tau^2 + \sigma^2/n)$).

When $\lambda = 1$, neither the data nor the priors can dominate the posterior, so the posterior has two modes. With the increment of $\lambda$, the prior prefers larger $\sigma^2$ and the information about $\tau^2$ from the data gets weaker. When $\lambda$ is large enough (such as $\geq 3$), the priors dominate the posterior and then the posterior has only one mode, which has a large $\sigma^2$ and a small $\tau^2$, i.e. a large $\phi$ (close to 1), which seems plainly
contrary to the data.

**Example 2:** The effect of changing $b$.

In Example 1, let $\lambda = 1$, and leave the others unchanged. Then $\ell(\phi)$ is bimodal (see Figure 5 (a) and (e)) and it fits the conditions of case 1 in Theorem 4.1, that is $\Delta > 0$ and $f(\phi_1) \cdot f(\phi_2) < 0$. Let $b$ take the values 0.2, 0.4, 2, 8 sequentially (see Table 3); then $\ell(\phi)$ changes as stated in Theorem 4.3, that is, with the increment of $b$, the modes move to the left. After $b \geq 0.5$, the right mode disappears and only one mode remains. Table 3 and Figure 5 show the relationship among $\Delta$, $f(\phi_1)$, $f(\phi_1)$ and the modality and how the modality changes as $b$ changes.

<table>
<thead>
<tr>
<th>$b$</th>
<th>$\Delta$</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$f(\phi_1)$</th>
<th>$f(\phi_2)$</th>
<th>$\hat{\phi}_1^{(1)}$</th>
<th>$\hat{\phi}_2^{(2)}$</th>
<th>$\hat{\phi}_3^{(3)}$</th>
<th>$I(\hat{\phi}_1)$</th>
<th>$I(\hat{\phi}_2)$</th>
<th>$I(\hat{\phi}_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>$1.056 \times 10^8$</td>
<td>0.30</td>
<td>0.78</td>
<td>253.92</td>
<td>-149.60</td>
<td>0.10</td>
<td>0.58</td>
<td>0.94</td>
<td>436.40</td>
<td>405.53</td>
<td>431.16</td>
</tr>
<tr>
<td>0.4</td>
<td>$9.149 \times 10^7$</td>
<td>0.31</td>
<td>0.77</td>
<td>304.71</td>
<td>-20.79</td>
<td>0.09</td>
<td>0.70</td>
<td>0.83</td>
<td>433.60</td>
<td>392.95</td>
<td>393.86</td>
</tr>
<tr>
<td>2</td>
<td>$-2.12 \times 10^7$</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.07</td>
<td>—</td>
<td>—</td>
<td>415.32</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>8</td>
<td>$-4.44 \times 10^8$</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.04</td>
<td>—</td>
<td>—</td>
<td>374.52</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Intuitively, at $b = 0.2$, the data and the priors provide opposite information about $\sigma^2$ and $\tau^2$ and neither can dominate the posterior, so the posterior has two modes. Increasing $b$ increases both the prior mean and variance of $\tau^2$, so, the priors provide weaker information about $\tau^2$. After $b \geq 0.5$, the data dominates the posterior and the posterior has just one mode, the one favored by the data. After $b > 8$, the priors also prefer large $\tau^2$ and small $\sigma^2$. The posterior and the data provide consistent information about $\sigma^2$ and $\tau^2$, therefore, the posterior is unimodal and consistent.
with the apparent message in the data and priors.

**Example 3:** The effect of changing $\alpha$ and $a$.

If we begin with the same values as in example 1 and example 2, the modality of the posterior changes with the change of $\alpha$ and $a$ as stated in Theorem 4.3. Table 4, Table 5, Figure 6 and Figure 7 show the relationship among $\Delta$, $f(\phi_1)$, $f(\phi_1)$ and how the modality changes as $\alpha$ and $a$ change.

**Table 4:** Values of $\Delta$, $\phi$ and $f(\phi)$ for different values of $\alpha$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\Delta$</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$f(\phi_1)$</th>
<th>$f(\phi_2)$</th>
<th>$\phi^{[1]}$</th>
<th>$\phi^{[2]}$</th>
<th>$\phi^{[3]}$</th>
<th>$I(\phi^{[1]})$</th>
<th>$I(\phi^{[2]})$</th>
<th>$I(\phi^{[3]})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>$1.056 \times 10^8$</td>
<td>0.30</td>
<td>0.78</td>
<td>253.92</td>
<td>-149.60</td>
<td>0.10</td>
<td>0.58</td>
<td>0.94</td>
<td>436.40</td>
<td>405.53</td>
<td>431.16</td>
</tr>
<tr>
<td>20</td>
<td>$1.33 \times 10^8$</td>
<td>0.30</td>
<td>0.80</td>
<td>367.08</td>
<td>-103.37</td>
<td>0.08</td>
<td>0.65</td>
<td>0.92</td>
<td>508.83</td>
<td>456.87</td>
<td>471.32</td>
</tr>
<tr>
<td>80</td>
<td>$2.586 \times 10^8$</td>
<td>0.32</td>
<td>0.84</td>
<td>769.24</td>
<td>41.83</td>
<td>0.05</td>
<td>—</td>
<td>—</td>
<td>783.37</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>150</td>
<td>$4.621 \times 10^8$</td>
<td>0.33</td>
<td>0.87</td>
<td>1240.82</td>
<td>189.98</td>
<td>0.04</td>
<td>—</td>
<td>—</td>
<td>1130.01</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

**Table 5:** Values of $\Delta$, $\phi$ and $f(\phi)$ for different values of $a$

<table>
<thead>
<tr>
<th>$a$</th>
<th>$\Delta$</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$f(\phi_1)$</th>
<th>$f(\phi_2)$</th>
<th>$\phi^{[1]}$</th>
<th>$\phi^{[2]}$</th>
<th>$\phi^{[3]}$</th>
<th>$I(\phi^{[1]})$</th>
<th>$I(\phi^{[2]})$</th>
<th>$I(\phi^{[3]})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>$1.056 \times 10^8$</td>
<td>0.30</td>
<td>0.78</td>
<td>253.92</td>
<td>-149.60</td>
<td>0.10</td>
<td>0.58</td>
<td>0.94</td>
<td>436.40</td>
<td>405.53</td>
<td>431.16</td>
</tr>
<tr>
<td>30</td>
<td>$1.212 \times 10^8$</td>
<td>0.30</td>
<td>0.76</td>
<td>138.77</td>
<td>-270.61</td>
<td>0.14</td>
<td>0.47</td>
<td>0.95</td>
<td>484.48</td>
<td>472.83</td>
<td>530.04</td>
</tr>
<tr>
<td>60</td>
<td>$1.545 \times 10^8$</td>
<td>0.30</td>
<td>0.72</td>
<td>-60.72</td>
<td>-496.90</td>
<td>—</td>
<td>—</td>
<td>0.97</td>
<td>—</td>
<td>—</td>
<td>713.94</td>
</tr>
<tr>
<td>100</td>
<td>$2.102 \times 10^8$</td>
<td>0.24</td>
<td>0.69</td>
<td>-320.64</td>
<td>-812.60</td>
<td>—</td>
<td>—</td>
<td>0.98</td>
<td>—</td>
<td>—</td>
<td>972.36</td>
</tr>
</tbody>
</table>

**Example 4:** The effect of changing $S_B^2$.

Let $N = 100$, $n = 2$, $S_B^2 = 6$, $S_W^2 = 1$, $a = 13$, $b = 0.2$, $\alpha = 3$ and $\lambda = 1$. Then $\Delta = -892109 < 0$ and the posterior is unimodal (see the plot of $\ell(\phi)$ in Figure 8 (a) and Table 6). As $S_B^2$ increases, $\Delta$ becomes positive, but $f(\phi_2) \cdot f(\phi_2) > 0$ for
some $S_B^2$ yielding positive $\Delta$, e.g. $S_B^2 = 8$. As $S_B^2$ continues to increase, $f(\phi_2) \cdot f(\phi_2)$ becomes negative and the posterior becomes bimodal. Table 6 and Figure 8 show how $\Delta$, $f(\phi_1) \cdot f(\phi_2)$ and the modes change as $S_B^2$ changes.

**Table 6: Values of $\Delta$, $\phi$ and $f(\phi)$ for different values of $S_B^2$**

<table>
<thead>
<tr>
<th>$S_B^2$</th>
<th>$\Delta$</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$f(\phi_1)$</th>
<th>$f(\phi_2)$</th>
<th>$\phi(1)$</th>
<th>$\phi(2)$</th>
<th>$\phi(3)$</th>
<th>$\text{I}(\phi(1))$</th>
<th>$\text{I}(\phi(2))$</th>
<th>$\text{I}(\phi(3))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>-892109</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.80</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>4331930</td>
<td>0.41</td>
<td>0.67</td>
<td>-22.41</td>
<td>-45.52</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.84</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>$1.232 \times 10^7$</td>
<td>0.37</td>
<td>0.71</td>
<td>14.74</td>
<td>-56.16</td>
<td>0.28</td>
<td>0.47</td>
<td>0.87</td>
<td>489.09</td>
<td>488.22</td>
<td>498.79</td>
</tr>
<tr>
<td>200</td>
<td>$1.34 \times 10^{10}$</td>
<td>0.25</td>
<td>0.83</td>
<td>4404.13</td>
<td>-1928.2</td>
<td>0.01</td>
<td>0.62</td>
<td>0.99</td>
<td>288.99</td>
<td>69.42</td>
<td>204.37</td>
</tr>
</tbody>
</table>

Intuitively, When $S_B^2 = 6$, $F = \frac{S_B^2/(N-1)}{S_W/N(n-1)} = \frac{6/99}{1/100} \approx 6$. The $F$ value is large, but the priors dominate the posterior (they have stronger information about $\sigma^2$ and $\tau^2$ than the data), and the posterior has only the mode favored by the priors (large $\sigma^2$ and small $\tau^2$). With the increment of $S_B^2$, the data also has strong information about $\sigma^2$ and $\tau^2$ that is opposite to the priors. So the posterior becomes bimodal. One mode is preferred by the data (large $\tau^2$, small $\sigma^2$, therefore, $\phi$ is close to 0), and the other is preferred by the priors (large $\sigma^2$, small $\tau^2$, therefore, $\phi$ is close to 1).
5 Future Work

1. The Modality of the Marginal Posterior for $\sigma^2$ and $\tau^2$.

2. The Boundaries for $\alpha$, $\lambda$, $a$, $b$, $S_B^2$ to change the number of modes.

3. Sensitivity Analysis: other directions, how the modality changes with $S_{W}^2$ and how the modes move with the change of $S_{B}^2$ and $S_{W}^2$.

4. The Modality of the likelihood, restricted likelihood and posterior: Regression Model

   Consider the following model and characterize the modes as in sections 4.

   $$ y_{ij} \sim N(\theta_i + x_{ij}\beta, \sigma^2), i = 1, 2, ..., N, j = 1, 2, ..., n; $$
   $$ \theta_i \sim N(\gamma_i, \tau^2), i = 1, 2, ..., N; $$
   $$ \sigma^2 \sim IG(\alpha, \lambda); $$
   $$ \tau^2 \sim IG(a, b). $$

5. The Modality of the likelihood, restricted likelihood and posterior: Random Regression Model

   Consider the following model and characterize the modes as in sections 4.

   $$ y_{ij} \sim N(\alpha_{0i} + \alpha_{1i}x_{i}, \sigma^2), i = 1, 2, ..., N, j = 1, 2, ..., n $$

   $$ \begin{pmatrix} \alpha_{0i} \\ \alpha_{1i} \end{pmatrix} \sim N(A, \Sigma) $$

   $$ \sigma^2 \sim IG(\alpha, \lambda) $$

   $$ \Sigma \sim Wishart(\Omega, p). $$
6 References


7 Appendix

7.1 Appendix A: The Proof of Lemma 1

Proof: It is easy to see that \( \frac{\partial^2 t}{\partial \theta_i^2} = -\frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} \) for all \( i \neq j \) and \( \frac{\partial^2 t}{\partial \theta_i \partial \theta_j} = 0 \), for all \( i \neq j \).

\[
\frac{\partial^2 t}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 t}{\partial \theta_i \partial \theta_j} = \frac{1}{\tau^2}.
\]

Consider the determinant

\[
D = \begin{vmatrix}
\frac{\partial^2 t}{\partial \theta_1^2} & \frac{\partial^2 t}{\partial \theta_1 \partial \theta_2} & \ldots & \frac{\partial^2 t}{\partial \theta_1 \partial \theta_N} & \frac{\partial^2 t}{\partial \theta_1 \partial \mu} \\
\frac{\partial^2 t}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 t}{\partial \theta_2^2} & \ldots & \frac{\partial^2 t}{\partial \theta_2 \partial \theta_N} & \frac{\partial^2 t}{\partial \theta_2 \partial \mu} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\frac{\partial^2 t}{\partial \theta_N \partial \theta_1} & \frac{\partial^2 t}{\partial \theta_N \partial \theta_2} & \ldots & \frac{\partial^2 t}{\partial \theta_N \partial \theta_N} & \frac{\partial^2 t}{\partial \theta_N \partial \mu} \\
\frac{\partial^2 t}{\partial \mu \partial \theta_1} & \frac{\partial^2 t}{\partial \mu \partial \theta_2} & \ldots & \frac{\partial^2 t}{\partial \mu \partial \theta_N} & \frac{\partial^2 t}{\partial \mu \partial \mu}
\end{vmatrix} = \begin{vmatrix}
-\frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} & 0 & \ldots & 0 & 1/\tau^2 \\
0 & -\frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} & \ldots & 0 & 1/\tau^2 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & -\frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} & 1/\tau^2 \\
1/\tau^2 & 1/\tau^2 & \ldots & 1/\tau^2 & -N/\tau^2
\end{vmatrix}
\]

\( D \) is a \((N + 1)\) dimensional determinant. Let \( D_i \) denote the left upper \( i \times i \) subdeterminant; if \( D_i < 0 \) for all the odd \( i \) and \( D_i > 0 \) for all the even \( i \) with \( \sigma^2 > 0 \) and \( \tau^2 > 0 \), then \((\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N, \hat{\mu})\) maximizes \( \ell \) given \( \sigma^2 > 0 \) and \( \tau^2 > 0 \).

It is easy to see that \( D_i = \left( -\frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^i \) for \( 1 \leq i \leq N \). Therefore, \( D_i < 0 \) for odd \( i \) and \( D_i > 0 \) for even \( i \) if \( i < N + 1 \). Now, consider \( i = N + 1 \), i.e. \( D_i = D \). If \( D < 0 \) when \( N \) is even and \( D > 0 \) when \( N \) is odd, then the conclusion of Lemma 1 holds.

Denote \( D \) as \( M^{(N)} \), where \( N \) means that the dimension of determinant \( D \) is \( N + 1 \). For \( N = 1 \), \( M^{(N)} = \frac{1}{\tau^4} \left( N \frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} - 1 \right) > 0 \). For \( N > 1 \), use the Laplace expansion,

\[
M^{(N)} = -\frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} M^{(N-1)} + (-1)^{N+2} \frac{1}{\tau^2} \frac{1}{\tau^2} \left( -1 \right)^{N-1} \frac{1}{\tau^2} \frac{1}{\tau^2} \left( -\frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-1}
\]

\[
= (-1)^N \frac{1}{\tau^4} \left( \frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-1} - \frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} M^{(N-1)}
\]

\[
= (-1)^N \frac{1}{\tau^4} \left( \frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-1} - \frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} \left[ (-1)^{N-1} \frac{1}{\tau^4} \left( \frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-2} - \frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} M^{(N-2)} \right]
\]

\[
= 2(-1)^N \frac{1}{\tau^4} \left( \frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-1} + \left( \frac{n\tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^2 M^{(N-2)}
\]
When $N$ is odd ($i = N + 1$ is even)

$$M^{(N)} = (N - 1)(-1)^N \frac{1}{\tau^4} \left( \frac{n \tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-1} + \left( \frac{n \tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-1} M^{(1)}$$

$$= \left( \frac{n \tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-1} \left[ M^{(1)} - \frac{N - 1}{\tau^4} \right]$$

$$= \left( \frac{n \tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-1} \left[ 1 + \frac{1}{\tau^4} \left( \frac{N n \tau^2 + \sigma^2}{\sigma^2} - 1 \right) - \frac{N - 1}{\tau^4} \right]$$

$$= \frac{N n}{\sigma^2 \tau^2} \left( \frac{n \tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-1} > 0$$

Similarly, when $N$ is even ($i = N + 1$ is odd) $M^{(N)} = -\frac{N n}{\sigma^2 \tau^2} \left( \frac{n \tau^2 + \sigma^2}{\sigma^2 \tau^2} \right)^{N-1} < 0$.

So, the only critical point $(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_N, \hat{\mu})$ maximizes $\ell$ given $\sigma^2 > 0$ and $\tau^2 > 0$. □

7.2 Appendix B: The Proof of Theorem 4.1

Prepare for the proof of Theorem 4.1.

Let $\phi = \frac{\sigma^2}{n \tau^2 + \sigma^2}$ and $\tau^2 = \tau_1^2$, a one-to-one map from the domain of $(\sigma^2, \tau^2)$, $\{ \langle \sigma^2, \tau^2 \rangle; \sigma^2 > 0, \tau^2 > 0 \}$, to the domain of $(\tau^2, \phi)$, $\{ \langle \tau^2, \phi \rangle; \tau^2 > 0, 0 < \phi < 1 \}$. Then $\ell_1(\sigma^2, \tau^2)$ is equivalent to

$$\ell_1(\tau^2, \phi) = -\frac{e}{2} \log \left( \frac{\phi \tau^2}{1 - \phi} \right) - \frac{1 - \phi}{2n \phi \tau^2} (2\lambda + S_W^2 + \phi^2 S_B^2) - \frac{1}{2} \log \tau^2 - \frac{1}{2\tau^2} (2b + (1 - \phi)^2 S_B^2 / n).$$

We can prove that $\tau^2 = \frac{1}{e + f} \left[ \frac{1 - \phi}{n} [2\lambda + S_W^2 + \phi^2 S_B^2] + [2b + (1 - \phi)^2 S_B^2 / n] \right]$ maximizes $\ell_1(\tau^2, \phi)$ given any $\phi \neq 0$. Substituting $\tau^2$ for $\tau^2$, we get the profile log posterior

$$\ell_1(\phi) = \frac{e}{2} \log(n \phi) + \frac{e}{2} \log(1 - \phi) - \frac{e + f}{2} \log(e + f) - \frac{e + f}{2} \log[(1 - \phi)(2\lambda + S_W^2) + 2nb \phi + (1 - \phi) S_B^2],$$

and $\ell_1' = \frac{e}{2} \alpha(\phi) \cdot f(\phi)$, where $\alpha(\phi) > 0$ for all $\phi \in (0, 1)$. This function $\ell_1(\phi)$ has the same modes as the log posterior with respect to $\phi$ and the modes are the solutions of $f(\phi) = 0$. Since $\ell_1(\phi)$ is a continuous function in $\phi$ and there are at most three critical points, therefore, $\ell_1(\phi)$ has
at most two modes. If \( f(\phi) \) has only two different solutions, then \( \ell_1(\phi) \) has at most one mode, and so does the log posterior.

**The proof of Theorem 4.1:** By Lemma 1 and the discussion in section 4.1, if a solution \((\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi})\) of (13) is a mode of \( \ell_2 \) with restriction \( \phi = \frac{\sigma^2}{n\tau^2+\sigma^2} \), then \((\hat{\sigma}^2, \hat{\tau}^2)\) is a mode of \( \ell_1 \).

From the equivalence of (13) and (14), all the \((\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi})\)'s are critical points of \( \ell_2 \). If \((\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi})\) maximizes \( \ell_2 \) then \((\hat{\sigma}^2, \hat{\tau}^2)\) is a mode of \( \ell_1 \). Now, we check which \((\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi})\)'s maximize \( \ell_2 \). If

\[
\frac{\partial^2 \ell_2}{\partial (\sigma^2)^2} < 0,
\begin{vmatrix}
\frac{\partial^2 \ell_2}{\partial (\sigma^2) \partial \phi} & \frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \tau^2} & \frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \phi} \\
\frac{\partial^2 \ell_2}{\partial \tau^2 \partial \sigma^2} & \frac{\partial^2 \ell_2}{\partial \tau^2 \partial \phi} & \frac{\partial^2 \ell_2}{\partial \tau^2 \partial \phi} \\
\frac{\partial^2 \ell_2}{\partial \phi \partial \sigma^2} & \frac{\partial^2 \ell_2}{\partial \phi \partial \tau^2} & \frac{\partial^2 \ell_2}{\partial \phi \partial \phi}
\end{vmatrix}
> 0,
\begin{vmatrix}
\frac{\partial^2 \ell_2}{\partial (\sigma^2)^2} & \frac{\partial^2 \ell_2}{\partial (\sigma^2) \partial \phi} & \frac{\partial^2 \ell_2}{\partial \phi \partial \sigma^2} \\
\frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \tau^2} & \frac{\partial^2 \ell_2}{\partial \tau^2 \partial \sigma^2} & \frac{\partial^2 \ell_2}{\partial \tau^2 \partial \phi} \\
\frac{\partial^2 \ell_2}{\partial \phi \partial \tau^2} & \frac{\partial^2 \ell_2}{\partial \phi \partial \phi} & \frac{\partial^2 \ell_2}{\partial \phi \partial \phi}
\end{vmatrix}
< 0
\]

at a critical point \((\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi})\), then \((\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi})\) maximizes \( \ell_2 \). First,

\[
\frac{\partial^2 \ell_2}{\partial (\sigma^2)^2} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = -\frac{e}{2\sigma^4}, \quad \frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \tau^2} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = 0, \quad \frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \phi} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = \frac{\hat{\phi}}{\sigma^2} S^2_{B};
\]

\[
\frac{\partial^2 \ell_2}{\partial \tau^2 \partial \sigma^2} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = 0, \quad \frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \phi} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = -\frac{f}{2\tau^4}, \quad \frac{\partial^2 \ell_2}{\partial \tau^2 \partial \phi} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = -\frac{1-\hat{\phi}}{n\tau^2+\sigma^2} S^2_{B};
\]

\[
\frac{\partial^2 \ell_2}{\partial \phi \partial \sigma^2} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = \frac{\hat{\sigma}}{\sigma^2} S^2_{B}, \quad \frac{\partial^2 \ell_2}{\partial \phi \partial \tau^2} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = -\frac{1-\hat{\phi}}{n\tau^2+\sigma^2} S^2_{B}; \quad \frac{\partial^2 \ell_2}{\partial \phi \partial \phi} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = -\frac{S^2_{B}}{n\tau^2\phi}.
\]

Therefore, \( \frac{\partial^2 \ell_2}{\partial (\sigma^2)^2} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = -\frac{e}{2\sigma^4} < 0 \) and

\[
\begin{vmatrix}
\frac{\partial^2 \ell_2}{\partial (\sigma^2)^2} & \frac{\partial^2 \ell_2}{\partial (\sigma^2) \partial \phi} & \frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \sigma^2} \\
\frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \tau^2} & \frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \phi} & \frac{\partial^2 \ell_2}{\partial \phi \partial \sigma^2} \\
\frac{\partial^2 \ell_2}{\partial \phi \partial \sigma^2} & \frac{\partial^2 \ell_2}{\partial \phi \partial \phi} & \frac{\partial^2 \ell_2}{\partial \phi \partial \phi}
\end{vmatrix} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = \begin{vmatrix}
-\frac{e}{2\sigma^4} & 0 & -\frac{f}{2\tau^4} \\
0 & -\frac{f}{2\tau^4} & 0
\end{vmatrix} = \frac{ef S^2_{B}}{4n\tau^2 \phi > 0}.
\]

Now, we need to check if

\[
0 > \begin{vmatrix}
\frac{\partial^2 \ell_2}{\partial (\sigma^2)^2} & \frac{\partial^2 \ell_2}{\partial (\sigma^2) \partial \phi} & \frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \sigma^2} \\
\frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \tau^2} & \frac{\partial^2 \ell_2}{\partial \sigma^2 \partial \phi} & \frac{\partial^2 \ell_2}{\partial \phi \partial \sigma^2} \\
\frac{\partial^2 \ell_2}{\partial \phi \partial \sigma^2} & \frac{\partial^2 \ell_2}{\partial \phi \partial \phi} & \frac{\partial^2 \ell_2}{\partial \phi \partial \phi}
\end{vmatrix} \bigg|_{\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}} = -\frac{ef S^2_{B}}{4n\sigma^2 \phi + \frac{f\hat{\phi}^2 S^2_{B}}{2n\tau^2 \phi^2} + \frac{e(1-\hat{\phi})^2 S^2_{B}}{2n\tau^2 \phi^2}}.
\]
\[ g(\hat{\phi}) = -e f \sigma^4 \hat{\phi}^4 \left( -\frac{\dot{\phi}}{\phi} \sigma^2 + 2 f \hat{\phi}(1 - \hat{\phi}) \sigma^4 S_B^2 + 2 e \hat{\phi}(1 - \hat{\phi})^2 \sigma^4 S_B^2 \right) = \frac{S_B}{4n^2r^2b^2 + \sigma^2} g(\hat{\phi}). \]

where \( g(\hat{\phi}) = -e f n \sigma^4 \hat{\phi}^2 + 2n^2 f \hat{\phi}^4 \sigma^2 S_B^2 + 2 e \hat{\phi}(1 - \hat{\phi})^2 \sigma^4 S_B^2 \). We need to determine if \( g(\hat{\phi}) < 0 \).

From equation (18), we know that \( -(e + f) S_B^2 \hat{\phi}^2 + (2e + f) S_B^2 \hat{\phi}^2 - [f(2\lambda + S_W^2) + 2neb + eS_B^2] \hat{\phi} = -f(2\lambda + S_W^2) \). Therefore, we have

\[ g(\hat{\phi}) = -(1 - \hat{\phi})^2 \left\{ 3(e + f) S_B^2 \hat{\phi}^2 - 2(e + f) S_B^2 \phi + [(2\lambda + S_W^2)f + 2neb + eS_B^2] \right\} \]

and \( 0 < \hat{\phi} < 1 \). We consider three cases.

**Case 1.** If \( \Delta > 0 \) and \( f(\phi_1)f(\phi_2) < 0 \), the three solutions \( \hat{\phi}^{(1)} < \hat{\phi}^{(2)} < \hat{\phi}^{(3)} \) of \( f(\phi) = 0 \) satisfy \( f'(\hat{\phi}^{(1)}) > 0 \), \( f'(\hat{\phi}^{(2)}) < 0 \) and \( f'(\hat{\phi}^{(3)}) > 0 \). Therefore, \( g(\hat{\phi}^{(1)}) < 0 \), \( g(\hat{\phi}^{(2)}) > 0 \), \( g(\hat{\phi}^{(3)}) < 0 \), i.e. \( \hat{\phi}^{(1)} \) and \( \hat{\phi}^{(3)} \) maximize \( \ell_2 \) with the corresponding \( (\hat{\sigma}^2, \hat{\tau}^2) \). Therefore, the corresponding \( (\hat{\sigma}^2, \hat{\tau}^2) \) maximize \( \ell_1 \). By the discussion before the proof, we know that the log posterior has at most two modes. We already have two modes, corresponding to \( \hat{\phi}^{(1)} \) and \( \hat{\phi}^{(3)} \), so, \( (\hat{\sigma}^2, \hat{\tau}^2) \) corresponding to \( \hat{\phi}^{(2)} \) is not a mode of \( \ell_1 \).

**Case 2.** If \( \Delta > 0 \) and \( f(\phi_1)f(\phi_2) = 0 \), then \( f(\phi) = 0 \) has two real solutions \( \hat{\phi}^{(1)} < \hat{\phi}^{(2)} \) with \( \hat{\phi}^{(1)} = \phi_1 \) or \( \hat{\phi}^{(2)} = \phi_2 \). If \( \hat{\phi}^{(1)} = \phi_1 \), then \( f'(\phi_2(\hat{\phi}^{(2)}) > 0 \). Therefore, \( g(\hat{\phi}^{(2)}) < 0 \) and \( \hat{\phi}^{(2)} \) maximizes \( \ell_2 \) with \( (\hat{\sigma}^2, \hat{\tau}^2) \), and as in case 1, \( \phi^{(1)} \) could not maximize \( \ell(\phi) \). Thus, \( (\hat{\sigma}^2, \hat{\tau}^2, \hat{\phi}^{(2)}) \)
is the only mode of $\ell_2$. Analogously, if $\hat{\phi}^{(2)} = \phi_2$, then $(\hat{\phi}^{2}, \hat{\tau}^{2}, \hat{\phi}^{(1)})$ is the only mode of $\ell_2$.

**Case 3.** If $\Delta \leq 0$ or $\Delta > 0$ with $f(\phi_1)f(\phi_2) > 0$ for $\phi \in (0, 1)$, then $f(\phi) = 0$ has only one real solution $\hat{\phi}$. If $\Delta \leq 0$, $f'(\phi) > 0$ (I have not finished the proof for case that both $\Delta = 0$ and $f'(\phi) = 0$). Thus, $g(\phi) < 0$ is always true and the only solution $\hat{\phi}$ of $f(\phi) = 0$ maximizes $\ell_2$ with $(\hat{\phi}^{2}, \hat{\tau}^{2})$. If $\Delta > 0$ and $f(\phi_1)f(\phi_2) > 0$, then the only solution $\hat{\phi}$ of $f(\phi) = 0$ is less than $\phi_1$ or greater than $\phi_2$. By the properties of $f'(\phi)$, $f'(\hat{\phi}) > 0$, therefore, $g(\hat{\phi}) < 0$ and the only solution $\hat{\phi}$ maximizes $\ell_2$ with $(\hat{\phi}^{2}, \hat{\tau}^{2})$.

Now, we derive the formulae for the $\hat{\phi}$'s in the theorem, the formulae for $\hat{\phi}^{2}$ and $\hat{\tau}^{2}$ are being given in (15) and (16).

Let $a_1 = (e + f)S_B^2$, $b_1 = -(2e + f)S_B^2$, $c_1 = 2f\lambda + fS_W^2 + 2neb + eS_B^2$, $d_1 = -f(2\lambda + S_W^2)$. Then we can rewrite $f(\phi) = 0$ as

$$a_1\phi^3 + b_1\phi^2 + c_1\phi + d_1 = 0. \quad (22)$$

Divide (22) by $a_1$ and let $\phi = y - \frac{b_1}{3a_1}$, (22) becomes

$$y^3 + py + q = 0, \quad (23)$$

where

$$q = K_1 + K_2 \frac{2\lambda + S_W^2}{S_B^2} + K_3 \frac{b}{S_B^2},$$
$$p = L_1 + L_2 \frac{2\lambda + S_W^2}{S_B^2} + L_3 \frac{b}{S_B^2},$$

where

$$K_1 = \frac{2e^3 + 3e^2f - 3ef^2 - 2f^3}{27(e + f)^3}, \quad K_2 = -\frac{(2 + 2f)f}{3(e + f)^2}, \quad K_3 = \frac{2(2e + f)ne}{3(e + f)^2}$$

and

$$L_1 = -\frac{e^2 + ef + f^2}{3(e + f)^2}, \quad L_2 = \frac{2ne}{e + f}, \quad L_3 = \frac{f}{e + f}.$$
If (23) has a solution \( y_1 \), then (22) has a solution

\[
\dot{y} = y_1 - \frac{b_1}{3a_1} = y_1 + \frac{2e + f}{3(e + f)}.
\]

(24)

Let \( \Delta_0 = \left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3 \); we know that

1. When \( \Delta_0 < 0 \), (23) has three different real solutions

\[
\begin{align*}
   y_1 &= \sqrt[3]{-\frac{q}{2} + \sqrt{\left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3}}; \\
   y_2 &= \omega \sqrt[3]{-\frac{q}{2} + \sqrt{\left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3}} + \omega^2 \sqrt[3]{-\frac{q}{2} - \sqrt{\left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3}}; \\
   y_3 &= \omega^2 \sqrt[3]{-\frac{q}{2} + \sqrt{\left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3}} + \omega \sqrt[3]{-\frac{q}{2} - \sqrt{\left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3}};
\end{align*}
\]

where \( \omega = \frac{-1 + i\sqrt{3}}{2}, \omega^2 = \frac{-1 - i\sqrt{3}}{2} \).

2. When \( \Delta_0 = 0 \) and \( \left( \frac{q}{2} \right)^2 = -\left( \frac{p}{3} \right)^3 \neq 0 \), then (23) has three real solutions; two of them are the same and the third is different. One is \( y_1 = -2\sqrt[3]{2} \) and the other two are \( y_2 = y_3 = \sqrt[3]{\sqrt[3]{2}} \).

3. When \( \Delta_0 = 0 \) and \( p = q = 0 \), then (23) has three real solutions and all are the same. The solutions are \( y_1 = y_2 = y_3 = 0 \).

4. When \( \Delta_0 > 0 \), (23) has only one real solution. The only real solution is

\[
y_1 = \sqrt[3]{-\frac{q}{2} + \sqrt{\left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\left( \frac{q}{2} \right)^2 + \left( \frac{p}{3} \right)^3}}
\]

Notice that (23) is just the cubic polynomial equation \( f(\phi) = 0 \) in the theorem. Combine 3 and 4, it will be case 3 in the theorem, and case 1 and case 2 in the theorem also correspond to 1 and 2 here. We only need to prove that the solutions here can be written in the forms in (20) for case 1 and case 2.

**Case 1.** In this case, we have these solutions, but it is not clear which \( y_1 \) corresponds to which \( \phi^{(i)} \). We now derive this using a few tricks to simplify the expressions for \( y_1 \).
Since $\Delta_0 < 0$, we have $p < 0$. Let $a_2 = -\frac{q}{3}$, $b_2 = |\Delta_0|^{1/2} = \sqrt{-\left(\frac{q}{3}\right)^2 - \left(\frac{p}{3}\right)^3}$, then
\[
a_2^2 + b_2^2 = \left(-\frac{p}{3}\right)^3, \quad \frac{a_2}{\sqrt{a_2^2 + b_2^2}} = \frac{3q}{2p} \sqrt{-\frac{3}{p}}.
\] (25)

Consider
\[
\sqrt[3]{-\frac{q}{2} + \sqrt{(\frac{q}{2})^2 + \left(\frac{p}{3}\right)^3}} = \sqrt[3]{-\frac{q}{2} + i\sqrt{-\left(\frac{q}{3}\right)^2 - \left(\frac{p}{3}\right)^3}} = \sqrt[3]{a_2 + b_2^2} = 
\]
\[
\left[\sqrt{a_2^2 + b_2^2} \cdot e^{i \arccos \frac{a_2}{\sqrt{a_2^2 + b_2^2}}}\right]^{1/3} = (-\frac{p}{3})^{1/2} \cdot e^{i \frac{1}{3} \arccos \frac{a_2}{\sqrt{a_2^2 + b_2^2}}}.
\] (26)

Similarly,
\[
\sqrt[3]{-\frac{q}{2} - \sqrt{(\frac{q}{2})^2 + \left(\frac{p}{3}\right)^3}} = (-\frac{p}{3})^{1/2} \cdot e^{i \frac{1}{3} \arccos \frac{a_2}{\sqrt{a_2^2 + b_2^2}}},
\] (27)

where $0 \leq \arccos \frac{a_2}{\sqrt{a_2^2 + b_2^2}} \leq \pi$. We used arccos is because $b_2$ is positive.

Notice, we used $\sqrt{(\frac{q}{2})^2 + \left(\frac{p}{3}\right)^3} = i\sqrt{-\left(\frac{q}{3}\right)^2 - \left(\frac{p}{3}\right)^3}$ and $\left[\sqrt{a_2^2 + b_2^2} \cdot e^{i \arccos \frac{a_2}{\sqrt{a_2^2 + b_2^2}}}\right]^{1/3} = e^{i \frac{1}{3} \arccos \frac{a_2}{\sqrt{a_2^2 + b_2^2}}}$.

Actually, they can also be $\sqrt{(\frac{q}{2})^2 + \left(\frac{p}{3}\right)^3} = -i\sqrt{-\left(\frac{q}{3}\right)^2 - \left(\frac{p}{3}\right)^3}$, $\left[\sqrt{a_2^2 + b_2^2} \cdot e^{i \arccos \frac{a_2}{\sqrt{a_2^2 + b_2^2}}}\right]^{1/3} = e^{i \frac{1}{3} \arccos \frac{a_2}{\sqrt{a_2^2 + b_2^2}} + \frac{2\pi}{3}}$

or $\left[\sqrt{a_2^2 + b_2^2} \cdot e^{i \arccos \frac{a_2}{\sqrt{a_2^2 + b_2^2}} + \frac{4\pi}{3}}\right]^{1/3} = e^{i \frac{1}{3} \arccos \frac{a_2}{\sqrt{a_2^2 + b_2^2}} + \frac{2\pi}{3}}$. But as long as we use the same one for $y_1$, $y_2$ and $y_3$, $\{y_1, y_2, y_3\}$ will be the same as a set.

Let $\psi = \frac{1}{3} \arccos \left(\frac{3q}{2p} \sqrt{-\frac{q}{3}}\right)$, from (25), (26) and (27), we have
\[
y_1 = \sqrt[3]{-\frac{q}{2} + \sqrt{(\frac{q}{2})^2 + \left(\frac{p}{3}\right)^3}} + \sqrt[3]{-\frac{q}{2} - \sqrt{(\frac{q}{2})^2 + \left(\frac{p}{3}\right)^3}} = \sqrt[3]{\sqrt[3]{\frac{p}{3} \cdot e^{i \psi} + \sqrt[3]{\frac{p}{3} \cdot e^{-i \psi}}}} = \sqrt[3]{\sqrt[3]{\frac{p}{3} \cdot \cos \psi + i \cdot \sin \psi} + \sqrt[3]{\frac{p}{3} \cdot \cos \psi - i \cdot \sin \psi}} = 2\sqrt[3]{\frac{p}{3} \cos \psi}.
\]

Similarly,
\[
y_2 = -\frac{1 + i\sqrt{3}}{2} \sqrt[3]{-\frac{q}{2} + \sqrt{(\frac{q}{2})^2 + \left(\frac{p}{3}\right)^3}} - \frac{1 - i\sqrt{3}}{2} \sqrt[3]{-\frac{q}{2} - \sqrt{(\frac{q}{2})^2 + \left(\frac{p}{3}\right)^3}} = \frac{1}{2} \left(\sqrt[3]{\frac{p}{3} \cdot e^{i \psi} + \sqrt[3]{\frac{p}{3} \cdot e^{-i \psi}}} - \sqrt[3]{\frac{p}{3} \cdot \cos \psi + i \cdot \sin \psi} - \sqrt[3]{\frac{p}{3} \cdot \cos \psi - i \cdot \sin \psi}\right)
\]
\[ y = \frac{-1 + i\sqrt{3}}{2}\sqrt{-\frac{p}{3} + \frac{i\sqrt{3}}{2}i\cdot \sin\psi} + \frac{-1 - i\sqrt{3}}{2}\sqrt{-\frac{p}{3} - \frac{i\sqrt{3}}{2}i\cdot \sin\psi} \]
\[ y_3 = \frac{-1 - i\sqrt{3}}{2}\sqrt{-\frac{q}{2} + \sqrt{\frac{q}{2}^2 + \left(\frac{p}{3}\right)^2} + \frac{-1 + i\sqrt{3}}{2}\sqrt{-\frac{q}{2} - \sqrt{\frac{q}{2}^2 + \left(\frac{p}{3}\right)^2}}. \]

Since \(0 \leq \arccos \frac{a}{\sqrt{a_3^2 + b_3^2}} \leq \pi, 0 \leq \psi \leq \frac{\pi}{2},\) and \(\sin\psi, \cos\psi\) and \(-3\) are positive. So, \(y_2 \leq y_3\) is always true. Now, we compare \(y_1\) and \(y_3\).

Consider \(y_1 - y_3 = (-\frac{p}{3})^{1/2} [3\cos\psi - \sqrt{3}\sin\psi].\) Since \(3\cos(\frac{\pi}{2}) - \sqrt{3}\sin(\frac{\pi}{2}) = 3\frac{1}{2} - 3\frac{\sqrt{2}}{2} = 0\) and \(\frac{d}{d\psi}(3\cos\psi - \sqrt{3}\sin\psi) = -3\sin\psi - \sqrt{3}\cos\psi < 0, y_1 - y_3 is always positive for \(\psi \in [0, \frac{\pi}{2}].\) Therefore, we have \(y_1 > y_3 > y_2.\) By \((24), (25)\) and the assumption \(\hat{\phi}^{(1)} < \hat{\phi}^{(2)} < \hat{\phi}^{(3)},\) we get the formulae in \((20).\)

**Case 2.** When \(\Delta_0 = 0\) and \((\frac{q}{2})^2 = -(-\frac{p}{3})^3 \neq 0, y_2\) is always equal to \(y_3.\) So, \(\phi^{(1)} = y_1 + \frac{2e + f}{3(e + f)} = -2\sqrt{\frac{2}{3}} + \frac{2e + f}{3(e + f)}\) corresponds to the only mode but \(\phi^{(2)} = y_2 + \frac{2e + f}{3(e + f)} = \sqrt{\frac{2}{3}} + \frac{2e + f}{3(e + f)}\) does not.

### 7.3 Appendix C: The Proof of Theorem 4.2

**Proof:** From \((9),\) when \(b = 0\) the posterior becomes

\[
p(\theta_1, \theta_2, \ldots, \theta_N, \mu, \sigma^2, \tau^2) \propto \frac{1}{(\sigma^2)^{\frac{N\mu + \alpha}{2} + \alpha + 1}} \exp \left\{ -\frac{1}{2\sigma^2} \left[ 2\lambda + \sum_{i=1}^{N} \sum_{j=1}^{N} (y_{ij} - \bar{y}_i)^2 + n \sum_{i=1}^{N} (\bar{y}_i - \bar{\theta}_1)^2 \right] \right\} \]  
\[
\times \frac{1}{(\tau^2)^{\frac{N\mu + \alpha}{2} + \alpha + 1}} \exp \left\{ -\frac{1}{2\tau^2} \sum_{i=1}^{N} (\theta_i - \mu)^2 \right\} \quad (28) \]

From \((28),\) when \(\theta_1 = \theta_2 = \cdots = \theta_N = \mu, p(\theta_1, \theta_2, \cdots, \theta_N, \mu, \sigma^2, \tau^2) \rightarrow \infty\) as \(\tau^2 \rightarrow 0\) for any \(\infty > \sigma^2 > 0.\) Otherwise (some \(\theta_i \neq \mu), p(\theta_1, \theta_2, \cdots, \theta_N, \mu, \sigma^2, \tau^2) \rightarrow 0\) as \(\tau^2 \rightarrow 0.\) Therefore, the log posterior \(\ell,\) has a spine at \((\theta, \theta, \cdots, \theta, \theta, \sigma^2, 0)\) for any \(\theta \) and \(\sigma^2 > 0.\) It is clear that the
spine covers a two-dimensional subspace of the \((N + 3)\)-dimensional parameter space. Now, we prove that the posterior may have another mode when \(\tau^2 > 0\).

When \(b = 0\), \(0 < \sigma^2 < \infty\) and \(0 < \tau^2 < \infty\), simplying (17) gives

\[
(f + e)S_B^2\phi^2 - eS_B^2\phi + f(2\lambda + S_W^2) = 0.
\]

(29)

Denote the left part of equation (29) by \(f_1(\phi)\), then the solutions of \(f_1(\phi) = 0\) correspond to the critical values of \(\ell_2\).

Let \(\Delta_1 = e^2 - 4(e + f)f \left(\frac{2\lambda}{S_B^2} + \frac{N(n-1)}{(N-1)F}\right)\), where \(F = \frac{S_B^2/(N-1)}{S_W} = \frac{S_B^2/(N-1)}{S_W}\) is the usual \(F\) statistic for balanced ANOVA, then \(f_1(\phi) = 0\) has two real solutions

\[
\begin{align*}
\hat{\phi}^{(1)} &= \frac{e - \sqrt{\Delta}}{2(e + f)} \\
\hat{\phi}^{(2)} &= \frac{e + \sqrt{\Delta}}{2(e + f)}
\end{align*}
\]

when \(\Delta_1 \geq 0\). If \(\Delta_1 = 0\), then \(\hat{\phi}^{(1)} = \hat{\phi}^{(2)}\). If \(\Delta_1 < 0\), then \(f(\phi) = 0\) has no real solution, therefore, the posterior does not have another mode.

Now we prove that the posterior has another mode when \(\delta_1 \geq 0\). To prove that a solution \(\hat{\phi}\) of \(f_1(\phi) = 0\) corresponds a mode of \(\ell_2\), as in the proof of Theorem 4.1, we need to prove that \(g(\hat{\phi}) < 0\) in (21). From (29), we have \(-f(2\lambda + S_W^2) = (f + e)S_B^2\hat{\phi}^2 - eS_B^2\hat{\phi}\). So \(g(\hat{\phi}) = (1 - \hat{\phi})\sigma^2 S_B^2 \left[-e + (3e + 2f)\hat{\phi} - 2(e + f)\hat{\phi}^2\right]\). It is easy to check that

\[
g(\hat{\phi}) \begin{cases} 
> 0 & \text{if } \frac{e}{2(e + f)} < \hat{\phi} < 1 \\
< 0 & \text{if } 0 \leq \hat{\phi} < \frac{e}{2(e + f)} \\
= 0 & \text{if } \hat{\phi} = \frac{e}{2(e + f)} \text{ or } \hat{\phi} = 1
\end{cases}
\]

It is clear that \(0 < \hat{\phi}^{(1)} < \frac{e}{2(e + f)} < \hat{\phi}^{(2)} < 1\) when \(\Delta_1 > 0\), i.e. \(g(\hat{\phi}^{(1)}) < 0\) and \(g(\hat{\phi}^{(2)}) > 0\). Therefore, \(\hat{\phi}^{(1)} = \frac{e - \sqrt{\Delta}}{2(e + f)}\) corresponds to a mode of \(\ell_2\) but \(\hat{\phi}^{(2)}\) does not (the proof for \(\hat{\phi}^{(2)}\) is similar to the proof of Theorem 4.1).
Now, we consider $\Delta_1 = 0$. At this case, $\hat{\phi} = \hat{\phi}^{(1)} = \hat{\phi}^{(2)} = \frac{e}{2(e+f)}$ and $g(\hat{\phi}) = 0$. We could not decide if $\hat{\phi}$ corresponds a maximal point of $\ell_1$ or not. As the proof of Theorem 4.1, we consider the profile $\ell(\phi)$ of $\ell_2$ with respect to $\phi$, of which the modes are corresponding to the modes of $\ell_2$ with respect to $\phi$. Now $(b = 0)$,

$$\ell(\phi) = \frac{e}{2} \left[ \log(2\lambda + S_W^2 + S_B^2 \phi^2) - \log(e) + 1 \right] - \frac{f}{2} \left[ \log((1 - \phi)^2 S_B^2/n) - \log(f) + 1 \right];$$

$$\ell'(\phi) = -\frac{e S_B^2 \phi}{2\lambda + S_W^2 + S_B^2 \phi^2} + \frac{f}{1 - \phi};$$

$$\ell''(\phi) = -\frac{e S_B^2 (2\lambda + S_W^2 - S_B^2 \phi^2)}{(2\lambda + S_W^2 + S_B^2 \phi^2)^2} + \frac{f}{(1 - \phi)^2}$$

$$= \frac{f [2\lambda + S_W^2 + S_B^2 \phi^2]^2 - (1 - \phi)^2 e S_B^2 (2\lambda + S_W^2 - S_B^2 \phi^2)}{[2\lambda + S_W^2 + S_B^2 \phi^2]^2 (1 - \phi)^2}$$

We can prove that $\hat{\phi} = \frac{e}{2(e+f)}$ is a critical value of $\ell(\phi)$ when $\Delta_1 = 0$. If $\ell''(\hat{\phi}) < 0$, then $\hat{\phi}$ is a mode of $\ell(\phi)$. Consider the numerator of $\ell''(\phi)$, it can be written as

$$\left\{ S_B^2 \phi^2 - \frac{e}{f + e} S_B^2 \phi + \frac{1}{f + e} [ (f - e)(2\lambda + S_W^2) + \frac{f e}{f + e} S_B^2 ] \right\} f_1(\phi) + g_1(\phi),$$

where

$$g_1(\phi) = \frac{e}{f + e} S_B^2 \left[ (e + 2f)(2\lambda + S_W^2) + \frac{f e}{f + e} S_B^2 \right] \phi + \frac{2\lambda + S_W^2}{f + e} \left[ 2f e (2\lambda + S_W^2) - e(f + e) S_B^2 - \frac{f e}{f + e} S_B^2 \right].$$

It is clear that $\ell''(\phi) < 0$ if only if $g_1(\phi) < 0$ because $f_1(\hat{\phi}) = 0$. We can check that

$$g_1(\hat{\phi}) = -\frac{e^2 f (2\lambda + S_W^2) S_B^2}{(f + e)^2} < 0.$$  

Therefore, the only critical point $\hat{\phi} = \frac{e}{2(e+f)}$ corresponds the finite mode of $\ell_2$. □

7.4 Appendix D: The Proof of Theorem 4.3

Proof:

1. First, we prove the conclusion for $b$.

Rewrite $f(\phi)$ as

$$f(\phi) = \left( \frac{1}{3} \phi - \frac{2e + f}{9(e + f)} \right) \hat{f}'(\phi) - \frac{S_B^2 \Delta}{18(e + f)} \phi - \frac{7e + 8f}{9(e + f)} (2\lambda + S_W^2) f + \frac{2e + f}{9(e + f)} (2neb + eS_B^2),$$

(30)
where
\[
\Delta = 4(2e + f)^2 - 12(e + f) \left( \frac{2f + 2neb}{S_B^2} + fN(n-1) (N-1)F + e \right) \\
= 4(2e + f)^2 - \frac{12(e + f)}{S_B^2} \left[ 2f + 2neb + fS_W^2 + eS_B^2 \right] > 0
\]

since \( \ell \) has two modes. Recall that \( f'(\phi_1) = f'(\phi_2) = 0, \phi_1 = \frac{2(2e + f) - \sqrt{\Delta}}{6(e + f)}, \phi_2 = \frac{2(2e + f) + \sqrt{\Delta}}{6(e + f)} \), \( f(\phi_1) > 0 \) and \( f(\phi_2) < 0 \); we can show that
\[
f(\phi_1) = -\frac{S_B^2 \Delta}{18(e + f)} - \frac{2(2e + f)^3}{9(e + f)} \frac{e + 2f}{3(e + f)} (2\lambda + S_W^2) F + \frac{2e + f}{3(e + f)} (2neb + eS_B^2) + \frac{S_B^2 \Delta^2}{108(e + f)^2} > 0, \quad (31)
\]
\[
f(\phi_2) = -\frac{2(2e + f)^3}{27(e + f)^2} - \frac{e + 2f}{3(e + f)} (2\lambda + S_W^2) F + \frac{2e + f}{3(e + f)} (2neb + eS_B^2) - \frac{S_B^2 \Delta^2}{108(e + f)^2} < 0, \quad (32)
\]
\[
\Delta = 4(2e + f)^2 - \frac{12(e + f)}{S_B^2} \left[ 2f + 2neb + fS_W^2 \right] - 12(e + f) e
\]
\[
= 4e^2 + 4e f + 4f^2 - \frac{12(e + f)}{S_B^2} \left[ 2f + 2neb + fS_W^2 \right]
\]
\[
\leq 4(e + f)^2. \quad (33)
\]

Regarding \( f(\phi_1) \), \( f(\phi_2) \) as functions of \( b \) with all the other parameters fixed, denote \( f(\phi_1) \) and \( f(\phi_2) \) by \( h_1(b) \) and \( h_2(b) \), respectively. Then
\[
\frac{dh_1}{db} = \frac{2ne(2e + f)}{3(e + f)} - \frac{ne\Delta^2}{3(e + f)} \geq 0, \quad \frac{2ne(2e + f)}{3(e + f)} - \frac{2ne(e + f)}{3(e + f)} = 0
\]
\[
\frac{dh_2}{db} = \frac{2ne(2e + f)}{3(e + f)} + \frac{ne\Delta^2}{3(e + f)} > 0.
\]

by (34). So, when \( b \) goes up, \( f(\phi_1) \) will remain positive. But
\[
f(\phi_2) \geq -\frac{2(2e + f)^3}{27(e + f)^2} - \frac{e + 2f}{3(e + f)} (2\lambda + S_W^2) F + \frac{2e + f}{3(e + f)} (2neb + eS_B^2) - \frac{S_B^2 \Delta^2}{108(e + f)^2} 8(e + f)^3 = c_1 + c_2 b
\]
by (34), where \( c_1 = -\frac{2(2e + f)^3}{27(e + f)^2} - \frac{e + 2f}{3(e + f)} (2\lambda + S_W^2) F + \frac{2e + f}{3(e + f)} (2neb + eS_B^2) - \frac{S_B^2 \Delta^2}{108(e + f)^2} 8(e + f)^3 < 0, \)
\( c_2 = \frac{2ne(2e + f)}{3(e + f)} > 0 \) and \( c_1, c_2 \) are not functions of \( b \). When \( b > -c_1/c_2, \) then \( f(\phi_2) > 0. \)

Therefore, increasing \( b \) we can get \( f(\phi_1)f(\phi_2) > 0, \) so \( \ell \) has only one mode left.

Now, let’s see why \( \ell(\phi) \) becomes unimodal before \( \Delta \) changes to negative. Because \( f(\phi_1) \)
keeps positive and \( f(\phi_2) \) can change to positive, when \( f(\phi_2) \) becomes little bit greater than 0,
\( f(\phi_1) \cdot f(\phi_2) > 0 \), but \( f(\phi_1) > f(\phi_2) \), thus \( \Delta > 0 \). Therefore, with the increase of \( b, \ell \) becomes unimodal before \( \Delta \) changes to negative.

We can simply write \( f(\phi) \) as \( f(\phi) = \frac{2}{3} \phi^3 n b - \frac{S^2_{\phi_1,\Delta} \phi}{18 \phi_2, f} \phi + t \), where \( t \) is not a function of \( b \). Since \( \Delta \) decreases with \( b \) increases, \( f(\phi) \) increases with \( b \) increases. Therefore, \( \phi^{(1)} \) and \( \phi^{(3)} \) move left with the increase of \( b \) because \( f(\phi) \) increases with \( b \) (see figure 9). Since \( f(\phi_2) \) goes positive from negative, \( \phi^{(2)} \) and \( \phi^{(3)} \) diminish as \( b \) increases. Therefore, only one mode at \( \phi^{(1)} \) left.

Now, we prove the conclusion for \( \alpha \). From (31) we see that \( f(\phi_1) \) can be written as

\[
\begin{align*}
2\alpha \cdot t_1(\alpha) + \frac{n^{(n+2)} + 2a + 4\alpha + 6}{(n+1)(2\alpha + 4\alpha + 6)} (2\lambda + S^2_W) f,
\end{align*}
\]

where \( t_1(\alpha) \rightarrow \frac{4}{3} nb + \frac{2}{27} S^2_B + \frac{S^2_B}{108} \left( 4 - \frac{4nb}{S^2_B} \right)^{\frac{3}{2}} \alpha \rightarrow \infty \). So, when \( \alpha \) is large enough, \( f(\phi_1) \geq 2\alpha \left[ \frac{4}{3} nb + \frac{2}{27} S^2_B + \frac{S^2_B}{108} \left( 4 - \frac{4nb}{S^2_B} \right)^{\frac{3}{2}} - \frac{3}{2} \right] + (\frac{1}{3} - \delta) (2\lambda + S^2_W) f \) for a very small positive \( \delta \) and similarly \( f(\phi_2) \geq 2\alpha \left[ \frac{4}{3} nb + \frac{2}{27} S^2_B - \frac{S^2_B}{108} \left( 4 - \frac{4nb}{S^2_B} \right)^{\frac{3}{2}} + \frac{3}{2} \right] + (\frac{1}{3} - \delta) (2\lambda + S^2_W) f \). Since \( \Delta \) can be written as \( \Delta = (2\alpha)^2 t_2(\alpha) \) and \( t_2(\alpha) \rightarrow 4 - \frac{4nb}{S^2_B} \) as \( \alpha \rightarrow \infty \). So when \( \alpha \) is large enough \( \Delta \) will be negative if \( \frac{4nb}{S^2_B} > 4, \ell \) has only one mode. If

\[
\frac{4nb}{S^2_B} \leq 4\text{ then } f(\phi_1) \geq 2\alpha \left[ \frac{4}{3} nb + \frac{2}{27} S^2_B + \frac{S^2_B}{108} \left( 4 - \frac{4nb}{S^2_B} \right)^{\frac{3}{2}} - \frac{3}{2} \right] + (\frac{1}{3} - \delta) (2\lambda + S^2_W) f > 0, \text{ and } f(\phi_2) \geq 2\alpha \left[ \frac{4}{3} nb + \frac{2}{27} S^2_B - \frac{S^2_B}{108} \left( 4 - \frac{4nb}{S^2_B} \right)^{\frac{3}{2}} + \frac{3}{2} \right] + (\frac{1}{3} - \delta) (2\lambda + S^2_W) f = \left( \frac{8}{3} nb - 2\delta \right) \alpha + (\frac{1}{3} - \delta) (2\lambda + S^2_W) f > 0
\]

when \( \alpha \) is large enough since \( \delta \) is very small. Both \( f(\phi_1) \) and \( f(\phi_2) \) are positive when \( \alpha \) is large enough. Therefore, \( \ell \) has only one mode.

We can write \( f(\phi) \) as \( f(\phi) = 2(S^2_B \phi^2 - 2S^2_B \phi + 2nb + S^2_B) \phi \alpha + c \), where \( c \) is not a function of \( \alpha \). Consider \( S^2_B \phi^2 - 2S^2_B \phi + 2nb + S^2_B \) as a quadratic polynomial with respect to \( \phi \). Since

\[
(-2S^2_B)^2 - 4 \cdot S^2_B \cdot (2nb + S^2_B) = -8nbS^2_B < 0 \text{ and } S^2_B > 0, \text{ so, } S^2_B \phi^2 - 2S^2_B \phi + 2nb + S^2_B > 0 \text{ for all } \phi, \text{ i.e. } f(\phi) \text{ increases with } \alpha \text{. As the case of } b, \phi^{(1)} \text{ and } \phi^{(3)} \text{ move left with the increase of } \alpha \text{. Because } f(\phi_2) \text{ becomes positive from negative, } \phi^{(2)} \text{ and } \phi^{(3)} \text{ diminish when } \alpha \text{ is large. Only} \]
the mode at $\hat{\phi}^{(1)}$ left.

2. First, we prove the conclusion for $\lambda$. Similarly, consider $f(\phi_1)$ and $f(\phi_2)$ as functions of $\lambda$ with all the other parameters fixed, denote $f(\phi_1)$ and $f(\phi_2)$ by $h_1(\lambda)$ and $h_2(\lambda)$, respectively, then

$$\frac{dh_1}{d\lambda} = -2f \frac{e + 2f}{3(e + f)} - \frac{f \Delta^\frac{1}{2}}{3(e + f)} \leq 0$$

$$\frac{dh_2}{d\lambda} = -2f \frac{e + 2f}{3(e + f)} + \frac{f \Delta^\frac{1}{2}}{3(e + f)} \leq -2f \frac{e + 2f}{3(e + f)} + \frac{f}{3(e + f)} \cdot 2(e + f) = -\frac{2f^2}{3(e + f)} < 0.$$

by (34). So, when $\lambda$ goes up $f(\phi_2)$ stays negative. As for $h_2(b)$, when $\lambda$ increases, $h_1(\lambda) = f(\phi_1)$ becomes negative. Therefore, when $\lambda$ increases we can get $\Delta > 0$ with $f(\phi_1)f(\phi_2) > 0$, and only one mode is left.

Rewrite $f(\phi)$ as $f(\phi) = t_3 - 2f(1 - \phi)\lambda$, where $t_3 = (e + f)S_B^2 \phi^3 - S_B^2 (2e + f)\phi^2 + (f S_W^2 + 2neb + eS_B^2)\phi - S_W^2 f$ is not a function of $\lambda$, we can see that $f(\phi)$ decreases with the increase of $\lambda$ when $\phi < 1$. By the similar argument as that for $b$, $\hat{\phi}^{(1)}$ and $\hat{\phi}^{(3)}$ move right with the increase of $\lambda$. Because $f(\phi_1)$ goes to negative from positive with the increase of $\lambda$, $\hat{\phi}^{(1)}$ and $\hat{\phi}^{(2)}$ diminish with the increase of $\lambda$. Only one mode at $\hat{\phi}^{(3)}$ left.

Similar to $a$, when $a$ is large enough,

$$f(\phi_1) \leq 2a \left[-\frac{2}{27}S_B^2 - \frac{2}{3}(S_W^2 + 2\lambda) + \frac{S_W^2}{108} \left[4 - \frac{12}{S_B^2}(2\lambda + S_W^2)\right]^\frac{1}{2} + \delta \right] + \left(\frac{1}{3} + \delta\right)(2neb + eS_B^2)$$

and

$$f(\phi_2) \leq 2a \left[-\frac{2}{27}S_B^2 - \frac{2}{3}(S_W^2 + 2\lambda) - \frac{S_W^2}{108} \left[4 - \frac{12}{S_B^2}(2\lambda + S_W^2)\right]^\frac{1}{2} + \delta \right] + \left(\frac{1}{3} + \delta\right)(2neb + eS_B^2).$$

If $\frac{12}{S_B^2}(2\lambda + S_W^2) > 4$ then $\Delta$ is negative, when $a$ is large enough, and $\lambda$ has only one mode. If $\frac{12}{S_B^2}(2\lambda + S_W^2) \leq 4$, then $f(\phi_1) \leq 2a \left[-\frac{2}{27}S_B^2 - \frac{2}{3}(S_W^2 + 2\lambda) + \frac{S_W^2}{108}4\left[\lambda + 3\delta\right]^\frac{1}{2} + \left(\frac{1}{3} + \delta\right)(2neb + eS_B^2) \right] = -\frac{2}{3}a(S_W^2 + 2\lambda + 3\delta) + \left(\frac{1}{3} + \delta\right)(2neb + eS_B^2) < 0$ and $f(\phi_2) < f(\phi_1) < 0$ when $a$ is large enough. Both $f(\phi_1)$ and $f(\phi_2)$ are negative, so $f(\phi_1)f(\phi_2) > 0$, and $\lambda$ has only one mode.

Similarly, $f(\phi) = 2(S_B^2 \phi^2 + 2\lambda + S_W^2)(\phi - 1)a + c$, where $c$ is not a function of $a$. $2(S_B^2 \phi^2 + 2\lambda + S_W^2)$ is not a function of $\lambda$.
When $S_W^2$ overight with the increase of $a$, diminish when $S_W^2$ from $a$. From $S_W^2$, if $f(\phi_1)$, $f(\phi_2)$ and $f(S_W^2)$, respectively, as function of $S_W^2$ with all other variables fixed. Then

$$
\frac{d\phi}{dS_W^2} = \frac{-2(2e + f)^3}{27(e + f)^2} + \frac{(2e + f)e}{3(e + f)} + \frac{\Delta^2}{108(e + f)^2} + \frac{\Delta^f}{6(e + f)S_W^2} \left[ 2f \lambda + 2neb + f S_W^2 \right]
$$

When $S_B^2$ is large enough,

$$
\Delta > (2e + f)^2
$$

and

$$
\Delta > (e + 2f)^2
$$

by (33).

From (36),

$$
\frac{d\phi}{dS_B^2} = \frac{-2(2e + f)^3}{27(e + f)^2} + \frac{(2e + f)e}{3(e + f)} + \frac{(e + 2f)^3}{108(e + f)^2} + \frac{(e + 2f)}{6(e + f)S_B^2} \left[ 2f \lambda + 2neb + f S_W^2 \right] 
> \frac{-8(2e + f)^3 + 36(2e + f)(e + f)e + (e + 2f)^3}{108(e + f)^2} 
= \frac{(e + 2f)e^2}{12(e + f)^2} > 0.
$$

Similarly, from (35),

$$
\frac{d\phi}{dS_B^2} < \frac{-2(2e + f)^3}{27(e + f)^2} + \frac{(2e + f)e}{3(e + f)} - \frac{(2e + f)^3}{108(e + f)^2} - \frac{(2e + f)}{6(e + f)S_B^2} \left[ 2f \lambda + 2neb + f S_W^2 \right] 
< \frac{-8(2e + f)^3 + 3e(2e + f)(e + f)e - (2e + f)^3}{108(e + f)^2} 
= \frac{-f^2(2e + f)}{12(e + f)^2} < 0
$$

50
Therefore, if \( f(\phi_1) \cdot f(\phi_2) > 0 \) and \( f(\phi_2) > 0 \), then as \( S_B^2 \) increases \( f(\phi_1) \) remains positive, \( f(\phi_2) \) becomes negative and \( \Delta \) becomes positive when \( S_B^2 \) is large enough. Similarly, if \( f(\phi_1) \cdot f(\phi_2) > 0 \) and \( f(\phi_2) < 0 \), then \( f(\phi_2) \) keeps negative when \( S_B^2 \) increases, and \( f(\phi_1) \) and \( \Delta \) become positive when \( S_B^2 \) is large enough. In either case, \( \ell \) will have two modes. This means that no prior parameters can guarantee unimodality in complete generality. \( \square \)
Figure 4: The plot of $f(\phi)$ and $\ell(\phi)$ for different $\lambda$
Figure 5: The plot of $f(\phi)$ and $\ell(\phi)$ for different $b$
Figure 6: The plot of $f(\phi)$ and $\ell(\phi)$ for different $\alpha$
Figure 7: The plot of $f(\phi)$ and $\ell(\phi)$ for different $a$
Figure 8: The plot of $f(\phi)$ and $\ell(\phi)$ for different $S^2_B$.
Figure 9: The plot to show how $\hat{\phi}^{(1)}$, $\hat{\phi}^{(s)}$ move as $b$ increases