Chapter 1

Parametric Prediction Region based on LMM

1.1 Introduction

The concept of *tolerance region* is of central importance in quality control. A tolerance region is a prediction set for an unobserved or future observation, which takes account of both the random nature of this observation and the uncertainty about its distribution. Parameters of the latter distribution are estimated using past data, where it is desired to account for the statistical error resulting from simply plugging in the estimated values.

Convential tolerance regions take the uncertainty of estimated parameters into account in one of two ways. Either the region captures the future observation a fraction of $1-\alpha$ times on average over both future and past observations (the $(1-\alpha)$-*expectation tolerance region*), or it captures the future observation with probability at least $1-\delta$ with $1-\alpha$ confidence over past observations (the (δ,α)-*tolerance region*). (See Section 1.2 for precise definitions.) The second way appears to be preferred in the pharmaceutical industry. The “on average" and “confidence" can refer to the sampling distribution of the data in a frequentist sense, but can also refer to a posterior distribution in the Bayesian statistical framework. The main focus on the present paper is the second, but we do relate it to the frequentist setup.

Frequentist tolerance regions have been well studied in the literature. A general reference is the book [28], especially for the situation that the data are i.i.d. For the linear mixed model (LMM), the paper [51] provides an elegant solution to build one- and two-sided (δ,α)-tolerance intervals, and includes a comprehensive review of the literature. One purpose of the present paper is to provide a Bayesian approach for the general LMM.

The Bayesian formulation of tolerance regions dates back to at least 1964, but the subsequent literature is relatively small. In his paper [1] Aitchison derived Bayesian (δ, α)-tolerance regions from a decision-theoretic framework, and contrasted them to the frequentist counterparts. In [2] he extended his discussion to $(1-\alpha)$-tolerance regions, which are a natural way to build Bayesian prediction intervals.

A one-sided (δ, α)-tolerance interval for a univariate future observation is usually easy to compute, but two-sided tolerance intervals pose challenges, both conceptually and computationally. There are at least two common approaches: intersecting two one-sided tolerance intervals, or fixing one degree of freedom of the interval (e.g. the midpoint of the interval). The former approach is identical to specifying probability masses in the two tails of the distribution of the future variable separately and gives a valid construction, in view of Bonferroni's inequality, but it yields longer intervals than necessary. (They are called "anti-conservative" in the pharmaceutical industry in reference to the customers, whereas statisticians might use the term "conservative".) See [21] for an example application. The second approach, fixing one degree of freedom, is the conventional choice, especially in the frequentist framework, but requires untangling the dependence of the interval on the unknown true parameter. Solutions are often not available in analytical form and computationally more challenging. Wolfinger in [67] proposed an algorithm to derive a two-sided Bayesian interval for a future normal variate, which was refined by Krishnamoorthy and Mathew [28]. Their algorithms have been widely adopted in practice, and also in other literature (e.g.[27], [57], [56]).

Among other contributions to the literature we mention the following: [35] used the empirical Bayes method to construct a one-sided tolerance interval given an i.i.d. sample from a normal distribution; [22] derived a Bayesian tolerance interval that contains a proportion of observations with a specified confidence; [14] and [68] focused on the sample size needed to attain a certain accuracy; [57] and [56] allowed data from the unbalanced one-way random effects model and the balanced two-factor nested random effects model; [36] discussed probability matching priors (PMP) in the one-sided case to ensure second-order frequentist validity; [41] extended this to the two-sided case; [42] incorporated it to a balanced one-way random effects model, and evaluated its performance against the frequentist method MLS in [28].

Although the PMP approach has merit when the sample size is small, it is analytically demanding even when data are i.i.d., and it seems difficult to extend to the general LMM setting. The algorithms of Wolfinger [67] and Krishnamoorthy and Mathew [28] can be extended to LMM, but they overly simplify the target function during optimization and may result in less satisfactory performance.

In this paper we propose a computationally efficient solution for the general case that the future observation possesses a normal distribution. We show that this is easy to implement given any data model for which a sample from the posterior distribution is available. We investigate when the shortest interval is centered at the posterior mean of the parameter. We discuss the interval in particular for the linear mixed model, and within this context show its good performance by simulation. We illustrate

the method on an example that is representative for pharmaceutical applications. Finally we also prove that the Bayesian interval has frequentist validity in the case of large samples.

1.2 Definitions and setup

Given are observed data X, with a distribution P_θ depending on a parameter θ, and future unobserved "data" Z , with a distribution Q_θ depending on the same parameter θ. In both cases the sample space is arbitrary. A tolerance region is a set $\mathscr{R}(X)$ in the sample space of Z that captures Z with a "prescribed probability". It will typically be constructed using the observation X to overcome the problem that θ, and hence the law of Z, is unknown. There are various ways to make the "prescribed probability" precise, and these can be divided into frequentist and Bayesian definitions. The probability statement will refer to both X and Z, and is fixed by one or two parameters α and δ, which are typically chosen small, e.g. 5%.

The parameter θ will typically be chosen to identify the distribution of Z. The distribution of X may also depend on unknown "nuisance" parameters. For simplicity of notation we do not make this explicit in the following. We shall use the notation P or P_θ for general probability statements, which may be reduced to P_θ or Q_θ if the event involves only X or Z.

1.2.1 Frequentist definitions

The most common frequentist definition is the (δ,α)-tolerance region. For a set R abbreviate $Q_\theta(R) = \mathrm{P}_\theta(Z \in R)$. Then $\mathscr{R}(X)$ is an (δ,α)-*tolerance region* if

$$P_\theta\Big(x\colon Q_\theta\Big(\mathscr{R}(x)\Big) \geqslant 1-\delta\Big) \geqslant 1-\alpha, \qquad \forall\theta. \tag{1.1}$$

If we let $Q_\theta\Big(\mathscr{R}(X)\Big)$ denote the probability or coverage of $\mathscr{R}(X)$ under Q_θ, for X held fixed, then we can also write the display in the shorter form $\mathrm{P}_\theta\Big(Q_\theta\Big(\mathscr{R}(X)\Big) \geqslant 1-\delta\Big) \geqslant 1-\alpha$, where the outer probability P_θ refers to X, and the inequality must hold for all possible values of the parameter θ. The latter reminds us of the definition of confidence sets, and indeed it can be seen that $\mathscr{R}(X)$ *is a frequentist* (δ,α)-tolerance region *if and only if the set* $\mathscr{C}(X) = \{\theta\colon Q_\theta\Big(\mathscr{R}(X)\Big) \geqslant 1-\delta\}$ *is a confidence set for* θ *of confidence level* $1-\alpha$.

An alternative is the α-*expectation tolerance region*, which requires that

$$\int Q_\theta\Big(\mathscr{R}(x)\Big)\, dP_\theta(x) \geqslant 1-\alpha, \qquad \forall\theta. \tag{1.2}$$

With the notational convention as before, the display can be written in the shorter form $\mathrm{E}_\theta Q_\theta\Big(\mathscr{R}(X)\Big) \geqslant 1-\alpha$, which is again required for all possible parameter values.

Both definitions have the form of requiring that $\mathrm{E}_\theta \ell\Big[Q_\theta\Big(\mathscr{R}(X)\Big)\Big] \geqslant 1-\alpha$, for all θ, and some given loss function ℓ. In the two cases this loss function is given by $\ell(q) = 1\{q \geqslant 1-\delta\}$ for (1.1), and $\ell(q) = q$ for (1.2), respectively, where $1\{\cdot\}$ is the indicator function.

1.2.2 Bayesian definitions

In the Bayesian setup the parameter θ is generated from a prior distribution Π, and the densities p_θ and q_θ are the conditional densities of X and Z given θ, respectively. To proceed, it is necessary to make further assumptions that fix the joint law of (θ, X, Z). The typical assumption is that X and Z are independent given θ.

A natural Bayesian approach is to refer to the predictive distribution of Z, and define a tolerance region $\mathscr{R}(X)$ to be a set such that $\mathrm{P}\Big(Z \in \mathscr{R}(X) | X\Big) \geqslant 1-\alpha$, i.e. a credible set in the posterior law of Z given X. The inequality can be written in terms of the posterior distribution $\Pi(\cdot|X)$ of θ given X as

$$\int \mathrm{P}\Big(Z \in \mathscr{R}(X) | X, \theta\Big)\, d\Pi(\theta|X) \geqslant 1-\alpha.$$

Under the conditional independence assumption this becomes

$$\int Q_\theta\Big(\mathscr{R}(X)\Big)\, d\Pi(\theta|X) \geqslant 1-\alpha. \tag{1.3}$$

This is like a frequentist α-expectation tolerance region (1.2), but with the expectation with respect to X under P_θ replaced by the expectation with respect to θ under the posterior distribution.

An alternative, derived from a utility analysis by Aitchison [1], is the *Bayesian* (δ,α)*-tolerance region*, which is a set $\mathscr{R}(X)$ such that

$$\Pi\Big(\theta : Q_\theta\Big(\mathscr{R}(X)\Big) \geqslant 1-\delta \,\Big|\, X\Big) \geqslant 1-\alpha. \tag{1.4}$$

This may be compared to (1.1). We can also say that $\mathscr{R}(X)$ *is a Bayesian* (δ,α)-tolerance region *if and only if the set* $\mathscr{C}(X) = \{\theta : Q_\theta\Big(\mathscr{R}(X)\Big) \geqslant 1-\delta\}$ *is a credible set at level* $1-\alpha$.

Both types of Bayesian regions satisfy $\int \ell\Big[Q_\theta\Big(\mathscr{R}(X)\Big)\Big]\, d\Pi(\theta|X) \geqslant 1-\alpha$, for the appropriate loss function ℓ. Solving the region $\mathscr{R}(X)$ from such an equation may seem daunting, but good approximations may be easy to obtain using stochastic simulation. This is true even for complicated data models, as long as one is able to implement an MCMC procedure for generating a sample from the posterior distribution given X. We make this concrete in Section 1.3.1 for a Gaussian variable Z, and illustrate this in Section 1.4 for an unbalanced linear mixed model (LMM).

1.2.3 Comparison

The frequentist and Bayesian definitions differ in the usual way in that the frequentist probability in (1.2) and (1.1) refers to the possible values of x in the sample space, whereas the Bayesian probability in (1.3) and (1.4) conditions on the observed value of X and refers to the distribution of the parameter.

As is the case for credible sets versus confidence sets, the Bayesian approach may feel more natural.

An advantage of Bayesian tolerance set is that while their form (1.4) is determined by the future variable Z, through the prediction problem Q_θ, the model for the data X enters only through the posterior distribution $\Pi(\theta \in \cdot | X)$. If in the former the dependence on the parameter θ is not too complicated, then the problem is solvable for even complicated data models. In contrast, the frequentist problem permits explicit solutions only in very special cases, although approximations and asymptotic expansions may extend their use (see [51]).

Neither the frequentist nor the Bayesian formulation restrains the shape of the region $\mathscr{R}(x)$. One may prescribe a fixed form and/or seek to optimize the shape with respect to an additional criterion, such as the volume of the region. The Bayesian formulation is again easier to apply, as the optimization will be given the data X. In the case of frequentist region it may be necessary to optimize an expected quantity instead.

In general the two approaches give difference tolerance regions, but the difference may disappear in the large sample limit. The requirements of the frequentist and Bayesian *tolerance regions* $\mathscr{R}(X)$ *for loss function* ℓ *and level* α, can be given symmetric formulations, as:

$$\mathrm{E}\Big(\ell\Big[Q_\theta\Big(\mathscr{R}(X)\Big)\Big] | \theta\Big) \geqslant 1-\alpha, \qquad \forall \theta, \tag{1.5}$$

$$\mathrm{E}\Big(\ell\Big[Q_\theta\Big(\mathscr{R}(X)\Big)\Big] | X\Big) \geqslant 1-\alpha. \tag{1.6}$$

In the first the expectation is taken with respect to X, which gives an integral over x with respect to the density p_θ. In the second the expectation is relative to θ, which leads to an integral relative to the posterior distribution given X. The integral of the first relative to the prior is identical to the integral of the second relative to the Bayesian marginal distribution of X, but there is no reason that a Bayesian tolerance region would also be a frequentist tolerance region.

However, Bayesian and frequentist inference typically merge if the informativeness in the data tends to a limit. For instance, this is true for regular parametric models in the sense that Bayesian credible sets are frequentist confidence sets, in the limit, with corresponding levels. The prior is then washed out and the Bayesian credible sets are equivalent to confidence sets based on the maximum likelihood estimator. This equivalence extends to tolerance regions, under some conditions. We defer a discussion to Section 1.5.

1.2.4 One-sided and two-sided tolerance intervals

For a one-dimensional future variable Z it is natural to choose $\mathscr{R}(x)$ an interval in the real line. The endpoints of such an interval are referred to as *tolerance limits*.

The single finite tolerance limit of a Bayesian *one-sided interval* is determined by meeting the (δ,α) or α tolerance criterion. The pair of tolerance limits of a Bayesian *two-sided interval* might be optimized to give an interval of minimal length, next to requiring that the tolerance criterion is met.

One-sided tolerance limits possess a straightforward interpretation and implementation. In particular, the (δ,α)-type has a simple description in terms of confidence intervals and posterior quantiles:

- $(-\infty,U(X)]$ is a frequentist (δ,α)-tolerance interval if and only if it is a $(1-\alpha)$-confidence interval for the induced parameter $Q_\theta^{-1}(1-\delta)$; it is a Bayesian (δ,α)-tolerance interval if $U(X)$ is the $(1-\alpha)$-quantile of the posterior distribution of $Q_\theta^{-1}(1-\delta)$ given X.
- $[L(X),\infty)$ is a frequentist (δ,α)-tolerance interval if and only if it is a $(1-\alpha)$-confidence interval for $Q_{\theta+}^{-1}(\delta)$; it is a Bayesian (δ,α)-tolerance interval if $L(X)$ is the $(1-\alpha)$-quantile of the posterior distribution of $Q_{\theta+}^{-1}(\delta)$.

Here $Q_\theta^{-1}(u)=\inf\{z\colon Q_\theta(-\infty,z]\geq u\}$ is the usual quantile function of Z, and $Q_{\theta+}^{-1}(u)$ is the right-continuous version of this quantile function. (The distinction between the two is usually irrelevant, and linked to the arbitrary convention of including the boundary point in the tolerance intervals.) The assertions follow by inverting the inequality $Q_\theta\Big(\mathscr{R}(X)\Big)\geqslant 1-\delta$, using the fact that for a cumulative distribution function Q and its quantile functions: $Q^{-1}(u)\leqslant x$ if and only if $u\leqslant Q(x)$, and $Q_+^{-1}(u)<x$ if and only if $u<Q(x-)$, for $u\in[0,1]$ and $x\in\mathbb{R}$.

A valid two-sided interval might be constructed as the intersection of two one-sided intervals, each at half of the error rate, but this will be conservative and lead to needlessly wide intervals. It makes good sense to try and minimize the length of the interval. We consider this in the next section for the case that the future observation Z is univariate Gaussian.

1.3 Normally distributed future observation

Consider the case that the future observation Z is univariate Gaussian with mean ν and variance τ^2. Thus the parameter is $\theta=(\nu,\tau)$, and $Z\sim Q_\theta=N(\nu,\tau^2)$. The probability that the future observation is captured within a candidate tolerance interval $[L,U]$ is

$$Q_\theta[L,U]=\Phi\left(\frac{U-\nu}{\tau}\right)-\Phi\left(\frac{L-\nu}{\tau}\right).$$

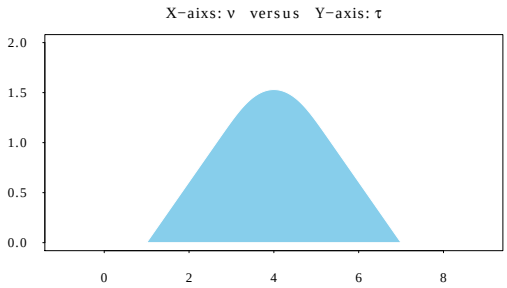

Fig. 1.1 The set $G_{A,B,\delta}$ of pairs (ν,τ) such that $\Phi\Big((U-\nu)/\tau\Big)-\Phi\Big((L-\nu)/\tau\Big)\geqslant 1-\delta$, for $A=4$, $B=3$, $\delta=0.1$. The number A is its horizontal point of symmetry and B is the half-length of its base. The base of the set (the line segment at height $\tau=0$) corresponds to the tolerance interval $[L,U]$.

It is convenient to parametrize the interval $[L,U]$ by its midpoint $A=(L+U)/2$ and half length $B=(U-L)/2$. For given $[L,U]$, or $\{A,B\}$, and $\delta\in(0,1)$, define the set

$$G_{A,B,\delta}=\{\theta=(\nu,\tau)\colon Q_\theta[L,U]\geqslant 1-\delta\Big\}.$$

For given $[L,U]$ the set $G_{A,B,\delta}$ is shaped as in Figure 1.1. It is symmetric about the vertical line $\nu=(L+U)/2$ and intersects the horizontal axis $\tau=0$ in the interval $[L,U]$. Changing A moves the set $G_{A,B,\delta}$ horizontally, while changing B changes its shape, with bigger B making the set both wider and taller. Although we use the normal distribution as our example, similarly shaped sets and conclusions would be obtainable for other unimodal symmetric distributions.

It follows that $\mathscr{R}(X)=\Big[L(X),U(X)\Big]$ satisfies inequality (1.1) and hence is a frequentist (δ,α)-tolerance interval for Z if and only if

$$P_\theta\Big(x\colon\theta\in G_{A(x),B(x),\delta}\Big)\geqslant 1-\alpha,\qquad\forall\theta.$$

In other words $[L(X),U(X)]$ is an (δ,α)-tolerance interval for Z if and only if $G_{A(X),B(X),\delta}$ is an $(1-\alpha)$-confidence region for $\theta=(\nu,\tau)$. Setting a joint confidence interval for location and dispersion is a familiar problem, but here the shape is restrained to the form $G_{A,B,\delta}$ and the focus will be on minimizing $B=B(X)$ (in some average sense). Solutions wil depend on the type of data X. Standard solutions are available in closed form for the simplest models, and more generally as approximations.

Similarly $\mathscr{R}(X)=\Big[L(X),U(X)\Big]$ satisfies inequality (1.4) and hence is a (δ,α)-Bayesian tolerance interval for Z if

$$\Pi\Big(\theta\colon\theta\in G_{A(X),B(X),\delta}\big|X\Big)\geqslant 1-\alpha. \tag{1.7}$$

It is natural to choose $A(X)$ and $B(X)$ to satisfy this inequality in such a way that $B(X)$ is minimal. In the resulting optimization problem the posterior distribution $\Pi(\theta\in\cdot\,|X)$ is a fixed probability distribution on the upper half plane and optimization entails shifting and scaling the shape shown in Figure 1.1 in a position such that it captures posterior mass at least $1-\alpha$, meanwhile minimizing

its width. In Section 1.3.1 we show how to achieve this given a large sample from the posterior distribution.

The data X determines the posterior distribution, but does not enter the optimization problem. The parameter θ may not be the full parameter characterising the distribution of X, but our computational strategy will work as long as θ is a function of this full parameter. For instance, if X follows a linear regression model with predictor "time" and Z is an observation at time 0, then ν will be a function the regression intercept; if X follows a random-effects model, then typically ν will depend on the fixed effects and τ^2 will be a specific linear combination of the variance components, depending on practical interests.

Frequentist methods typically choose $A(X)$ equal to a standard estimator of ν. One might guess that the Bayesian solution will be to take $A(X)$ equal to the posterior mean $\mathrm{E}(\nu|X)$ of ν. This would be convenient as it would reduce the optimization of (A,B) to the problem of only optimizing B. However, the posterior mean does not necessarily give the minimal length interval. The following lemma gives a sufficient condition.

Lemma 1.3.1. *Suppose that the conditional distribution of* ν *given* (X,τ) *is unimodal and symmetric with decreasing density to the right of its mode and has mean* $\mathrm{E}(\nu|X,\tau)$ *that is free of* τ*. Then the shortest* (δ,α)*-Bayesian tolerance interval* $[L,U]$ *for a future variable* $Z \sim N(\nu,\tau^2)$ *is centered at the posterior mean* $\mathrm{E}(\nu|X)$*.*

Proof. We can decompose the probability on the left side of (1.7) as

$$\Pi\Big(G_{A,B,\delta}|X\Big) = \int \Pi\Big(\nu \in (G_{A,B,\delta})_\tau|X,\tau\Big)\, d\Pi(\tau|X),$$

where $(G_{A,B,\delta})_\tau$ is the section of $G_{A,B,\delta}$ at height τ. By the unimodality and monotonicity the integrand is maximized over A for every τ and a given B by choosing $A = \mathrm{E}(\nu|X,\tau)$. If this does not depend on τ, then this common maximizer A will maximize the whole expression. Since we need to determine A and B so that the expression is at least $1-\alpha$, maximizing it over A will give the minimal B. By assumption $A = \mathrm{E}(\nu|X,\tau) = \mathrm{E}(\nu|X)$. ∎

The condition of the preceding lemma is not unreasonable, but depends on the prior, as illustrated in the following simple example. In Section 1.4 we show that for a LMM the condition is approximately satisfied. Then choosing A equal to the posterior mean is a fast computational shortcut that may perform almost as well as the optimal solution.

Example 1.3.1. The simplest possible data model is to let $X = (X_1,\ldots,X_n)$ be a random sample from the $N(\nu,\tau^2)$-distribution. This example was already discussed by Aitchison [1].

For the standard priors $\tau^2 \sim IG(\alpha_0, \beta_0)$ and $\nu|\,\tau \sim N(a, \tau^2/b)$, the conditional posterior distribution of ν given τ is normal with mean

$$\mathrm{E}\Big(\nu|\,\tau, X\Big) = \frac{ba + n\bar{X}}{b+n}.$$

Since this is independent of τ. the preceding discussion shows that the shortest (δ, α)-tolerance interval is centred at this posterior mean.

Choosing the prior variance $\mathrm{var}(\nu|\,\tau)$ proportional to τ^2, which is customary, is crucial for this finding. For instance, if we set the prior ν to be independent of τ, say $\nu|\,\tau \sim N(a, b^{-1})$, then the conditional posterior mean changes to

$$\mathrm{E}\Big(\nu|\,\tau, X\Big) = \frac{ba\tau^2 + n\bar{X}}{b\tau^2 + n}.$$

This is $\bar{X}$ if $\tau = 0$ and shrinks to the prior mean a as $\tau \to \infty$. For illustration, let $a = 0$, $b = 0.1$, $\alpha_0 = \beta_0 = 0.01$. Given data with $n = 3$, $\bar{x} = 10$, $s^2 = 1$, we approximated the posterior distribution of (ν, τ) given X by a Gibbs sampler, using the full conditional posteriors, where $s^2 = \sum (x_i - \bar{x})^2/(n-1)$:

$$\begin{aligned}
\nu|\,\tau, X &\sim N\left(\frac{ba\tau^2 + n\bar{X}}{b\tau^2 + n}, \frac{\tau^2}{b\tau^2 + n}\right), \\
\tau^2|\,\nu, X &\sim IG\left(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{(n-1)s^2 + n\bar{X}^2 + n\nu^2 - 2n\nu\bar{X}}{2}\right).
\end{aligned}$$

The contour plots of the posterior distribution in the left panel of Figure 1.2 show dependence between ν and τ given X, and ensuing functional dependence of $\mathrm{E}(\nu|\,\tau, X)$ on τ. Using Algorithm 2, as explained in Section 1.3.1, we computed the shortest tolerance interval (i.e. the smallest $\hat{B}$) for every possible location A of the interval, for A in a neighbourhood of the posterior mean $\mathrm{E}(\nu|\,X)$. The right panel of Figure 1.2 shows $\hat{B}$ as a function of $\hat{A}$. The minimum value is *not* taken at the location of the posterior mean $\mathrm{E}(\nu|\,X)$, which is indicated by a dashed line.

Admittedly the data in this example has been tweaked to illustrate the principle. Inspection of the vertical scale for $\hat{B}$ shows that the global minimal length of a tolerance interval is only slightly smaller than the length of the interval centred at the posterior mean of ν. In Section 1.4 we show, by theoretical derivation, a similar phenomenon for linear mixed models, and in Section 1.5 we study this approximation in a large sample context.

1.3.1 Computational strategy

In this section we elaborate on the computation of the two-sided (δ, α) Bayesian tolerance interval for a normally distributed univariate future variable Z, as discussed in Section 1.3. We also compare our approach to the one taken in [67] and [28]. We assume given a large sample of values $\theta_j = (\nu_j, \tau_j)$

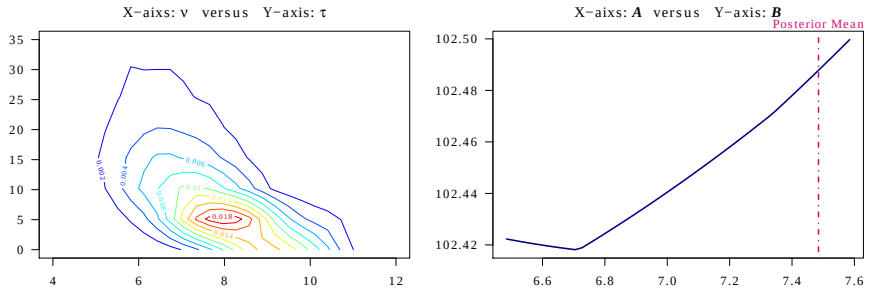

Fig. 1.2 Left panel: density-level contourplots of MCMC approximation to realization of posterior distribution $\nu,\tau|X$ in Example 1.3.1 (prior parameters: $a=0$, $b=0.1$, $\alpha_0=\beta_0=0.01$; data $\bar{x}=10$, $s^2=1$, $n=3$). Right panel: corresponding half-lengths B (vertical axis) of the $(\delta=0.05,\alpha=0.1)$-tolerance interval centered at A (horizontal axis); the minimal length is not taken at the posterior mean $\mathrm{E}(\nu|X)$, whose location is indicated by the abscissa of the dotted line.

from the posterior distribution of $\theta=(\nu,\tau)$ given X. This could be the result of an MCMC run of a sampler for the posterior distribution, or, depending on the data model, of using an analytic formula for the posterior distribution. We shall use the sample values (ν_j,τ_j) to approximate expectations under the posterior distribution, whence they need not be independent, and values from a (burnt-in) MCMC run will indeed qualify. Possible dependence together with sample size will determine the error due to simulation.

The idea is to replace the posterior distribution in (1.4) by the empirical distribution of the values $\{\theta_j\}_{j=1}^J$. For a given interval $\mathscr{R}(X)=[L,U]$ we can (in theory) compute the values $Q_{\theta_j}[L,U]$ and next search for the interval $[L,U]$ of minimal length $U-L$ such that

$$\frac{1}{J}\#\Big(Q_{\theta_j}[L,U]\geqslant 1-\delta\Big)\doteq 1-\alpha,$$

where $\doteq$ means approximately equal, yielding a slightly conservative or anti-conservative solution in case exact equality is not attainable due to discretization.

Algorithm 1 was proposed by Wolfinger [67] and later refined (or corrected) by Krishnamoorthy and Mathew [28], Chapter 11. This algorithm has a convenient graphical representation and has been widely adopted in practice. The idea is to compute for every θ_j the quantiles $L_j=Q_{\theta_j}^{-1}(\delta/2)$ and $U_j=Q_{\theta_j}^{-1}(1-\delta/2)$, yielding intervals $[L_j,U_j]$ with $Q_{\theta_j}[L_j,U_j]\geqslant 1-\delta$, and next setting the tolerance interval $[L,U]$ equal to an interval that is symmetric about the posterior mean and contains a fraction $1-\alpha$ of the intervals $[L_j,U_j]$ (Krishnamoorthy and Mathew, referred to as KM), or is contained in a fraction α of these intervals (Wolfinger, W). The graphical interpretation is to plot the points (U_j,L_j) in the x-y-plane and search for a point (U,L) on the line $y+x=2\hat{\nu}$, for $\hat{\nu}$ the posterior mean or some other useful estimator, such that a fraction $1-\alpha$ of the points are in the left-upper quadrant relative to the point $[L,U]$ (see Figure 11.1 in [28] for an example). The KM method results in an interval that is more confident (KM) than the prescribed level $1-\alpha$, and appears not to optimize the length of the interval.

Algorithm 1: WKM solution for two-sided tolerance interval

Data: Given $\alpha, \delta, \{(\nu_j, \tau_j)\}_{j=1}^J$

Let $\hat{A} = \sum_j \nu_j / J$;

Calculate two quantiles sequences: $\{L_j \equiv Q^{-1}_{\nu_j,\tau_j}(\frac{\delta}{2})\}_{j=1}^J$ and $\{U_j \equiv Q^{-1}_{\nu_j,\tau_j}(1-\frac{\delta}{2})\}_{j=1}^J$;

Find a point $(\hat{L}, \hat{U})$ such that $\hat{L} + \hat{U} = 2\hat{A}$ satisfying one of the following ;

(W) $\arg\min_{\hat{L},\hat{U}} \left| \frac{\#S}{J} - \alpha \right|$, where $S = \{(L_j, U_j) : L_j \leq \hat{L}, U_j \geq \hat{U}\}$;

(KM) $\arg\min_{\hat{L},\hat{U}} \left| \frac{\#S}{J} - 1 + \alpha \right|$, where $S = \{(L_j, U_j) : L_j \geq \hat{L}, U_j \leq \hat{U}\}$;

Result: two-sided tolerance interval $[\hat{L}, \hat{U}]$

Here we propose another algorithm that directly utilizes (1.4). We seek to minimize B under the constraint that the interval $[L,U] = [A-B, A+B]$ satisfies (1.4). This takes two steps: for fixed A we optimize over B; next we perform a grid search over A. Because given A, the optimizer over B will yield equality in (1.4), $\hat{B}$ will be the solution to

$$\Pi\left[\Phi\left(\frac{A+B-\nu}{\tau}\right) - \Phi\left(\frac{A-B-\nu}{\tau}\right) \geq 1-\delta \,|\, X\right] = 1-\alpha. \tag{1.8}$$

The posterior mean $\mathrm{E}(\nu|X)$ will typically be close to the optimal solution for A, and is a good starting point for this parameter. As a fast approximation we may also set A to this value and omit the grid search. The posterior mean can be approximated by the average of the sample values ν_j.

In practice we replace the posterior distribution in equation (1.8) by an average over the sample values (ν_j, τ_j). Given $\hat{A}$ we approximate $\hat{B}$ by the $(1-\alpha)$-quantile of the points g_j computed as the solutions to

$$Q_{\nu_j,\tau_j}[\hat{A}-g_j, \hat{A}+g_j] \equiv \Phi\left(\frac{\hat{A}+g_j-\nu_j}{\tau_j}\right) - \Phi\left(\frac{\hat{A}-g_j-\nu_j}{\tau_j}\right) = 1-\delta. \tag{1.9}$$

The motivation for this procedure is that $(\nu_j, \tau_j) \in G_{A,B,\delta}$ if and only if $Q_{\nu_j,\tau_j}[A-B, A+B] \geqslant 1-\delta$, whence precisely the points (ν_j, τ_j) with $g_j \leqslant \hat{B}$ satisfy $Q_{\nu_j,\tau_j}[\hat{A}-\hat{B}, \hat{A}+\hat{B}] \geqslant 1-\delta$ and hence are inside the set $G_{\hat{A},\hat{B},\delta}$, whereas the other points are outside this set. This makes the posterior mass of the set equal to $1-\alpha$ up to simulation error.

Given $\hat{A}$ and (ν_j, τ_j), the function $g \mapsto Q_{\nu_j,\tau_j}[\hat{A}-g, \hat{A}+g]$ in (1.9) is increasing, from the value 0 when $g = 0$ to 1 as $g \to \infty$ (see Figure 1.3). The solutions g_j to each equation (1.9) can be found fast by a Newton-Raphson algorithm, with some caution on choosing the initial value for g_j (the algorithm will diverge if the initial value is chosen in the domain where $Q_{\nu_j,\tau_j}[\hat{A}-g_j, \hat{A}+g_j]$ is very close to 1). An appropriate algorithm is listed in Algorithm 2. Note that the (middle) expression in (1.9) does not change if $\varepsilon := \hat{A} - \nu_j$ is replaced by $-\varepsilon$.

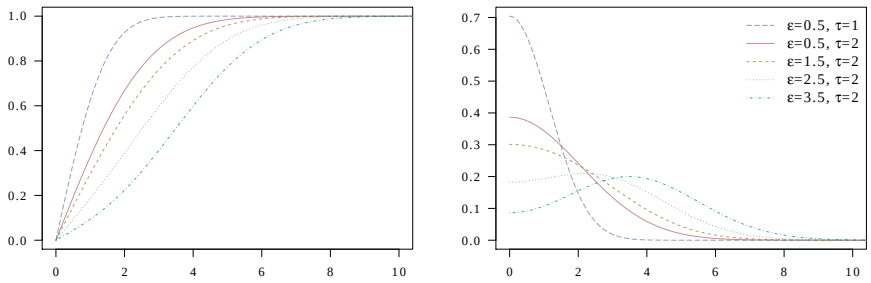

Fig. 1.3 Left panel: plot of the curves $g \mapsto \Phi(\frac{\varepsilon+g}{\tau}) - \Phi(\frac{\varepsilon-g}{\tau})$ for various settings of ε and τ. Right panel: derivatives of these curves.

Algorithm 2: Proposed solution for two-sided tolerance interval

Data: Given $\alpha, \delta, \{(\nu_j, \tau_j)\}_{j=1}^{J}$

Let $\hat{A} = \sum_j \nu_j / J$;

for $j = 1, 2, ..., J$ **do**

  Solve equation (1.9) by a Newton-Raphson algorithm as follows;

  if $|\hat{A} - \nu_j| < \tau_j$ **then**

    $g_0 = |\hat{A} - \nu_j| + \tau_j$

  else

    $g_0 = |\hat{A} - \nu_j|$

  set initial value for g_j at g_0;

  while $\omega > 0.0001$ **do**

    let $q_{\nu_j,\tau_j}[\hat{A} - g_j, \hat{A} + g_j]$ be the first-order derivative of $Q_{\nu_j,\tau_j}[\hat{A} - g_j, \hat{A} + g_j]$;

    $g_j = g_j - \frac{Q_{\nu_j,\tau_j}[\hat{A}-g_j,\hat{A}+g_j]-1+\delta}{q_{\nu_j,\tau_j}[\hat{A}-g_j,\hat{A}+g_j]}$;

    $\omega = Q_{\nu_j,\tau_j}[\hat{A} - g_j, \hat{A} + g_j] - 1 + \delta$;

The above loop results in $\{g_j\}_{j=1}^{J}$, and let $\hat{B}$ be its $(1-\alpha)th$ sample quantile;

Result: two-sided tolerance interval $[\hat{A} - \hat{B}, \hat{A} + \hat{B}]$

1.4 Data from a linear mixed model

In this section we apply the preceding to a model that is representative for practice in pharmaceutical quality control: the *linear mixed model* (LMM). We assume that the data X are acquired in an LMM design, and that the future variable Z is defined in terms of the same LMM. We concentrate attention to the two-sided (δ,α)-tolerance interval.

In the LMM we observe a vector $X = U\beta + V\gamma + e$, for known (deterministic) matrices U and V of covariates, a vector of fixed effects parameters β, an unobserved random effect vector γ, and an error vector e. Assume that γ and e are independent, with

$$\gamma \sim N(0,D), \qquad e \sim N(0,\sigma^2 I). \tag{1.10}$$

Then the data X follows a $N(U\beta,C)$-distribution, for $C = VDV^T + \sigma^2 I$, and the full parameter is (β,D,σ^2).

Consider predicting a new observation $Z = u^T\beta + v^T\gamma' + e'$ with given fixed and random effects coefficients u and v and *newly generated* random effect vector γ' and error e', with $\gamma' \sim \gamma$ and $e' \sim N(0,\sigma^2)$. Thus γ' is assumed equal in distribution to, but independent of γ, and similarly for e'. This target for prediction is reasonable in many contexts, but sometimes another choice, in particular for γ', may be more relevant. Typically γ will carry a group structure matched by a block structure in V. The vector v will then have nonzero coordinates corresponding to a single group. The target corresponds to setting the distribution Q_θ of Z equal to $N(\nu,\tau^2)$, with

$$\nu = u^T\beta, \qquad \tau^2 = v^T D v + \sigma^2.$$

The "prediction" parameter $\theta = (\nu,\tau)$ is of smaller dimension than the full parameter" (β,D,σ^2) governing the distribution of the data X, whence part of the latter full parameter should be considered a nuisance parameter. To set a Bayesian tolerance interval we need a posterior distribution of θ given the data X. This will typically be inferred from a posterior distribution of the full parameter, resulting from a prior distribution on (β,D,σ^2).

The (conditional) posterior distribution for a conditional prior $\beta|D,\sigma^2 \sim N(0,\Lambda)$, where Λ may depend on (D,σ^2), satisfies

$$\beta|D,\sigma^2,X \sim N\Big((U^TC^{-1}U+\Lambda^{-1})^{-1}U^TC^{-1}X, (U^TC^{-1}U+\Lambda^{-1})^{-1}\Big).$$

In general $\mathrm{E}(\nu|D,\sigma^2,X) = u^T\mathrm{E}(\beta|D,\sigma^2,X)$ will depend on D and σ^2 (hidden in C) and hence typically also on τ^2. Therefore Lemma 1.3.1 does not apply, and there appears to be no reason that a shortest tolerance interval would be centered at the posterior mean of ν. To obtain the shortest interval Algorithm 2 should be augmented with a search on possible centerings A.

As in the standard *i.i.d.* model in Example 1.3.1, the dependence on σ^2 can be removed by choosing the variances Λ and D proportional to σ^2. If $D = \sigma^2 D_0$ and $\Lambda = \sigma^2 \Lambda_0$, then C will be $\sigma^2(VD_0V^T + I) =: \sigma^2 C_0$ and the conditional posterior mean of β will be $(U^T C_0^{-1} U + \Lambda_0^{-1})^{-1} U^T C_0^{-1} X$. However, the dependence on D_0 (through C_0) remains, in general.

Letting the prior covariance matrix Λ tend to infinity corresponds to the noninformative prior. If all other quantities are fixed and $\Lambda \to \infty$, then

$$\mathrm{E}(\beta | D, \sigma^2, X) \to (U^T C^{-1} U)^{-1} U^T C^{-1} X.$$

The limit is the maximum likelihood estimator of β in the model where C is known. Since this is still dependent on C (and hence D and σ^2), it seems that for both the Bayesian and frequentist tolerance intervals the two parameters ν and τ cannot be separated in general. The choice $\Lambda = \lambda (U^T C^{-1} U)^{-1}$ leads to $\lambda/(1+\lambda)$ times the maximum likelihood estimator.

1.4.1 Approximations to the conditional posterior mean

For special designs the dependence of $\mathrm{E}(\beta | D, \sigma^2, X)$ on (D, σ^2) is only mild and can be quantified. We discuss some examples.

Example 1.4.1 (One-way random effects)**.** Suppose X is a vector with coordinates $X_{ik} = \beta + \gamma_i + e_{ik}$, for $i = 1, \ldots, m$ and $k = 1, \ldots, n$, ordered as $(X_{11}, \ldots, X_{1n}, X_{21}, \ldots, X_{mn})$, where $\beta \in \mathbb{R}$ and $\gamma = (\gamma_1, \ldots, \gamma_m)^T$ with i.i.d. $\gamma_i \sim N(0, d^2)$, so that $D = d^2 I_m$, for I_m the $(m \times m)$-identity matrix. As prior on β we choose a one-dimensional normal distribution $N(0, \lambda^2)$.

The matrix U is the mn-vector 1_{mn} with all coordinates equal to 1, while V is the $(mn \times m)$-matrix with $i$the column having 1s in rows $(i-1)n+1$ to in and 0s in the other rows. Then $V^T V = nI_m$, and $U^T V = n 1_m$, and it can be verified that $C 1_{mn} = (nd^2 + \sigma^2) 1_{mn}$ and hence $C^{-1} U = (nd^2 + \sigma^2)^{-1} 1_{mn}$. The coefficient vector of $\mathrm{E}(\beta | D, \sigma^2, X)$ is

$$(U^T C^{-1} U + \lambda^{-2})^{-1} U^T C = \left(mn + \frac{nd^2 + \sigma^2}{\lambda^2} \right)^{-1} 1_{mn}^T.$$

For $\lambda = \infty$, this is free of d^2 and σ^2, while for finite, fixed λ and $m, n \to \infty$, the coefficient vector is $(mn)^{-1} \Big(1 + O(d^2/(m\lambda^2)) + O(\sigma^2/(mn\lambda^2)) \Big)$.

The dependence on d and σ can be removed by choosing $d = d_0 \sigma$ and $\lambda = \lambda_0 \sqrt{n d_0^2 + 1}\, \sigma$.

Example 1.4.2. Suppose $X_{ik} = u_{ik}^T \beta + v_{ik}^T \gamma_i + e_{ik}$, ordered as in the preceding example, but now with observed covariates $u_{ik} \in \mathbb{R}^p$ and $v_{ik} \in \mathbb{R}^q$, fixed effects parameter $\beta \in \mathbb{R}^p$ and i.i.d. random effects

$\gamma_i \sim N_q(0, D_q)$. The corresponding matrices U and V are

$$U = \begin{pmatrix} U_1 \\ \vdots \\ U_m \end{pmatrix}, \quad V = \begin{pmatrix} V_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_m \end{pmatrix}, \quad U_i = \begin{pmatrix} u_{i1}^T \\ \vdots \\ u_{in}^T \end{pmatrix}, \quad V_i = \begin{pmatrix} v_{i1}^T \\ \vdots \\ v_{in}^T \end{pmatrix}.$$

Then $C = VDV^T + \sigma^2 I$ is an $(mn \times mn)$-block-diagonal matrix with blocks $V_i D_q V_i^T + \sigma^2 I_n$, and

$$\begin{aligned} U^T C^{-1} U &= \sum_{i=1}^m U_i^T (V_i D_q V_i^T + \sigma^2 I_n)^{-1} U_i, \\ U^T C^{-1} &= \Big(U_1^T (V_1 D_q V_1^T + \sigma^2 I_n)^{-1}, \ldots, U_m^T (V_m D_q V_m^T + \sigma^2 I_n)^{-1} \Big). \end{aligned}$$

The matrices $U_i^T V_i$ and $V_i^T V_i$ are of dimensions $p \times q$ and $q \times q$, and are sums over the n observations (for $k = 1, \ldots, n$) per group i, as defined by the random effect γ_i. In convential asymptotics we could view them as n times a matrix of fixed order. Then, for $D_0 = \sigma^{-2} D_q$,

$$\begin{aligned} (V_i D_0 V_i^T + I)^{-1} &= I - V_i (D_0^{-1} + V_i^T V_i)^{-1} V_i^T \\ &= I - V_i (V_i^T V_i)^{-1} \Big[D_0^{-1} (V_i^T V_i)^{-1} + I \Big]^{-1} V_i^T \\ &= I - V_i (V_i^T V_i)^{-1} \Big[I - D_0^{-1} (V_i^T V_i)^{-1} + (D_0^{-1} (V_i^T V_i)^{-1})^2 + \cdots \Big] V_i^T \\ &= P_{V_i^\perp} + V_i (V_i^T V_i)^{-1} D_0^{-1} (V_i^T V_i)^{-1} V_i^T \\ &\qquad - \sum_{k=2}^{\infty} (-1)^k V_i (V_i^T V_i)^{-1} (D_0^{-1} (V_i^T V_i)^{-1})^k V_i^T, \end{aligned} \tag{1.11}$$

where $P_{V^\perp}$ is the projection on the orthocomplement of the linear span of the columns of V. If $V_i^T V_i$ is large, then the series on the right can be neglected, as its terms contain multiple terms $(V_i^T V_i)^{-1}$. There is then still dependence on D and σ^2 in the second term, which may dominate.

In a full random effects model we shall have $u_{ik} = v_{ik}$, for every (i,k), and then $U_i = V_i$. Then $U_i^T P_{V_i^\perp} = 0$ and $U_i^T V_i (V_i^T V_i)^{-1} = I_p$. If $n^{-1} V_i^T V_i$ stabilizes as $n \to \infty$,

$$\begin{aligned} \sigma^2 U^T C^{-1} U &= \sum_{i=1}^m \sum_{k=1}^{\infty} (-1)^{k-1} D_0^{-1} ((V_i^T V_i)^{-1} D_0^{-1})^{k-1} = m \Big(D_0^{-1} + O\Big(\frac{1}{n}\Big) \Big), \\ \sigma^2 U^T C^{-1} &= \sum_{k=1}^{\infty} (-1)^{k-1} \Big(D_0^{-1} (V_i^T V_i)^{-1})^k V_i^T \Big)_{i=1}^m \\ &= \Big(D_0^{-1} (V_i^T V_i)^{-1} \Big(I + O\Big(\frac{1}{n}\Big) \Big) V_i^T \Big)_{i=1}^m. \end{aligned}$$

Thus we find that

$$(U^TC^{-1}U+\Lambda^{-1})^{-1}U^TC^{-1}$$
$$=\left(\frac{1}{m}\left(I+O\left(\frac{D_0}{n}\right)+\frac{D_0\Lambda^{-1}}{m}\right)^{-1}\left((V_i^TV_i)^{-1}\left(I+O(\frac{1}{n})\right)V_i^T\right)\right)_{i=1}^m.$$

To first order this is free of D and σ^2 with relative remainders of the order D_0/n and $D_0\Lambda^{-1}/m$.

If every random effect is matched by a fixed effect with the same covariate vector (supplying a common mean value to the random effects), but there are more fixed than random effects, then $u_{ik}^T=(v_{ik}^T,\bar{u}_{ik}^T)$, which implies $U_i=(V_i,\bar{U}_i)$ for an $(n\times(p-q))$-matrix $\bar{U}_i$. The preceding formulas must then be adapted to, where the approximations refer to ignoring the series in (1.11), for $W_i=V_i(V_i^TV_i)^{-1}$,

$$\sigma^2U^TC^{-1}U\doteq\sum_{i=1}^m\left[\begin{pmatrix}0&0\\0&\bar{U}_i^TP_{V_i^\perp}\bar{U}_i\end{pmatrix}+\begin{pmatrix}D_0^{-1}&D_0^{-1}W_i^T\bar{U}_i\\\bar{U}_i^TW_iD_0^{-1}&\bar{U}_i^TW_iD_0^{-1}W_i^T\bar{U}_i\end{pmatrix}\right],$$
$$\sigma^2U^TC^{-1}\doteq\left(\begin{pmatrix}0\\\bar{U}_i^TP_{V_i^\perp}\end{pmatrix}+\begin{pmatrix}D_0^{-1}W_i^T\\\bar{U}_i^TW_iD_0^{-1}W_i^T\end{pmatrix}\right)_{i=1}^m.$$

It is reasonable to expect that the matrices $\bar{U}_i^TW_i=\bar{U}_i^TV_i(V_i^TV_i)^{-1}$ will settle down as, as will the matrices $n^{-1}\bar{U}_i^TP_{V_i^\perp}\bar{U}_i$. Then up to lower order terms

$$(U^TC^{-1}U)^{-1}U^TC^{-1}\doteq\sum_{i=1}^m\begin{pmatrix}D_0^{-1}&D_0^{-1}W_i^T\bar{U}_i\\\bar{U}_i^TW_iD_0^{-1}&\bar{U}_i^TP_{V_i^\perp}\bar{U}_i\end{pmatrix}^{-1}\left(\begin{pmatrix}D_0^{-1}W_i^T\\U_i^TP_{V_i^\perp}\end{pmatrix}\right)_{i=1}^m.$$

Here the three appearances of D_0^{-1} in the top row cancel each other (as follows by factorizing out the block diagonal matrix with blocks D_0^{-1} and I_{p-q} from the inverse matrix), while the factor $\bar{U}_i^TW_iD_0^{-1}$ in the bottom row is a factor $1/n$ smaller in order than the matrix $\bar{U}_i^TP_{V_i^\perp}\bar{U}_i$ and hence can be set to 0 up to order $1/n$. It follows that again the matrix is free of D and σ^2 up to order $1/n$.

1.4.2 Numerical examples

We evaluate the small sample performance of our proposed algorithm via a simulation study, with data generated from a one-way random effects model. We observe $(X_{ik}:i=1,2,\ldots,m;k=1,2,\ldots,n_i)$, where X_{ik} is the k^{th} value of the i^{th} group and satisfies $X_{ik}=\nu+\gamma_i+e_{ik}$ for i.i.d. $\gamma_i\sim N(0,d^2)$ independent of the i.i.d. error $e_{ik}\sim N(0,\sigma^2)$. We are interested in the two-sided (δ,α)-tolerance interval related to a future observation $Z=\nu+\gamma+e\sim Q_\theta=N(\nu,\tau^2)$, where γ and e are independent copies of the γ_i and e_{ik}. The parameter is $\theta=(\nu,\tau^2)$, for $\tau^2=d^2+\sigma^2$.

We used 6 different parameter settings. In every setting the overall mean was set equal to $\nu = 0$, and the group variance to $d^2 = 1$. The intra-correlation $\sigma^2/(d^2+\sigma^2)$ was chosen equal to the numbers $0.1, 0.3, 0.5, 0.7, 0.9$. In every setting the number of groups was $m = 6$ and the group sizes were $(n_1, \ldots, n_6) = (2,3,4,2,3,4)$. This simulation setup is the same as in [51], and facilitates a cross-comparison against the performance of a standard frequentist solution.

We computed the $(\delta = 0.1, \alpha = 0.05)$-tolerance interval $[L,U]$ by Algorithm 2, for every of $K = 1000$ replicates of the data in each parameter setting, and computed the true coverage $Q_\theta[L,U]$ of these intervals using the true parameter θ of the simulation. We dub an interval with true coverage no less than $1-\delta$ as "qualified", and compared the empirical fraction of qualified intervals out of the K replicates to the nominal value $1-\alpha$. The procedure is considered to perform well in the frequentist sense if this empirical fraction is close to this nominal value. Here we must allow for the simulation error, which has a standard error of $\sqrt{p(1-p)/K}$, for p the true coverage, which is unknown, but hopefully close to $1-\alpha$.

For each simulated dataset the posterior distribution of (ν, d, σ^2) was approximated by a standard Gibbs sampler (with the vector of random effects γ_i added in as a fourth parameter), before utilizing Algorithm 2. Two setups of priors were deployed, both with independent priors on the three parameters ν, d, σ. The first is the vanilla setup with vague marginal prior distributions $\nu \sim N(0, 1000)$, $d^2 \sim IG(0.001, 0.001)$ and $\sigma^2 \sim IG(0.001, 0.001)$. The second uses the same prior on σ^2, but uses a t-distribution for ν given by the hierarchy $\nu|\,\sigma_0 \sim N(0, \sigma_0^2)$ and $\sigma_0^2 \sim IG(0.001, 0.001)$, and the prior on d given by the structural equation $d = |\xi|\omega$ for independent $\xi \sim N(0,1)$ and $\omega^2 \sim IG(0.001, 0.001)$. The latter specification can also be understood as over-parameterizing the distribution of the random effect two parameters instead of one, as $\gamma|\,\xi, \omega \sim N(0, \xi^2\omega^2)$. This "data augmentation" or "parameter expansion" is meant to enhance the mixing rate of the Gibbs sampler, in particular when the number of groups m is small. See [6] for a comparison of methods (including non-Bayesian methods) to fit the LMM. A more accurate approximation to the posterior should have positive impact on the subsequent tolerance interval. Finer amendments should be possible for concrete cases.

In all simulation settings the tolerance intervals were constructed both by fixing the center point A at the posterior mean $\mathrm{E}(\nu|X)$, and by seeking an optimal value of A to minimize the half length B of the interval. Thus four tolerance intervals were calculated based on each simulated dataset. To save on computation time the optimization over A was carried out only approximately. Still for 85% of the simulation cases a shorter interval was obtained than the interval at the posterior mean, in a few cases as much as 20% shorter, but in 75% of the cases no more than a few percentage points. Table 1.1 reports the quotients of the lengths.

The proportions of qualified intervals (the ones which attain true coverage $1-\delta = 0.9$) are listed in Table 1.2. They are reasonably close to the nominal value $1-\alpha = 0.95$, with deviations in both directions up to several percentage points. The performance seems to depend on the true intra-correlation. This dependence follows a similar pattern as for the high-order asymptotic solution

in [51] without correction for imbalance, and is close to their solution that includes correction when the intra-correlation is small or very big. The tolerance intervals centered at the (approximately) optimal value have shorter length and attain lower confidence, but their performance seems to surpass slightly the intervals centered at the posterior mean $\mathrm{E}(\nu|X)$.

The difference in performance of Algorithm 2 between the two prior setups is within the order of the simulation error.

Table 1.1 Interval length at optimal value of A relative to fixing A at $\mathrm{E}(\nu|X)$

	under Vanilla setup				*under Parameter Expansion setup*			
	Min	$0.25^{th}qu.$	$Median$	$0.75^{th}qu.$	Min	$0.25^{th}qu.$	$Median$	$0.75^{th}qu.$
$intra-correlation$								
0.1	0.8420	0.9958	0.9984	0.9993	0.8625	0.9969	0.9991	0.9997
0.3	0.8244	0.9943	0.9984	0.9994	0.8569	0.9969	0.9989	0.9997
0.5	0.8192	0.9905	0.9981	0.9992	0.8374	0.9954	0.9985	0.9995
0.7	0.8193	0.9897	0.9979	0.9992	0.8020	0.9947	0.9984	0.9994
0.9	0.8102	0.9889	0.9980	0.9993	0.8058	0.9925	0.9980	0.9993

Table 1.2 Approximated Confidence for $(\delta = 0.1, \alpha = 0.05)$-Bayesian tolerance interval

	under Vanilla setup		*under Parameter Expansion setup*	
	$A = \mathrm{E}(\nu|X)$	$A = Optimal$	$A = \mathrm{E}(\nu|X)$	$A = Optimal$
$intra-correlation$				
0.1	0.972	0.969	0.968	0.963
0.3	0.964	0.955	0.955	0.949
0.5	0.936	0.921	0.925	0.917
0.7	0.925	0.907	0.915	0.911
0.9	0.952	0.941	0.940	0.935

1.5 Frequentist justification of the Bayesian procedure

In this section we show that Bayesian tolerance regions are often also approximate frequentist tolerance regions, of corresponding levels. We consider an asymptotic setup, with data $X = X_n$ indexed by a parameter $n \to \infty$, in which the *Bernstein-von Mises theorem* holds. The latter theorem (see [59], Chapter 10) entails that the posterior distribution $\Pi_n(\cdot|X_n)$ of θ can be approximated by a normal distribution with deterministic covariance matrix, centered at an estimator $\hat{\theta}_n = \hat{\theta}_n(X_n)$,

$$\Pi_n(\cdot|X_n) - N\left(\hat{\theta}_n, \frac{1}{n}\Sigma_\theta\right) \xrightarrow{P} 0, \tag{1.12}$$

(in total variation norm), where the estimators $\hat{\theta}_n = \hat{\theta}_n(X_n)$ satisfy

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{\theta}{\rightsquigarrow} N(0, \Sigma_\theta). \tag{1.13}$$

For instance, under regularity conditions this is valid for X_n a vector of n i.i.d. observations from a smooth parametric model, with $\hat{\theta}_n$ the maximum likelihood estimator and Σ_θ the inverse Fisher information matrix. The Bernstein-von Mises theorem can be used to show that Bayesian and frequentist inference (testing and confidence sets) *merge* for large sample sizes. In this section we investigate this for tolerance intervals.

We shall show that Bayesian tolerance regions $\mathscr{R}_n(X_n)$ such that the functions

$$h \mapsto Q_{\hat{\theta}_n(X_n)+h/\sqrt{n}}\Big(\mathscr{R}_n(X_n)\Big), \qquad n = 1, 2, \ldots, \tag{1.14}$$

stabilize asymptotically to a *deterministic* function are asymptotically frequentist tolerance regions, for any given loss function ℓ and level α. The crux of this *stability condition* is that the randomness which enters the functions (1.14) through X_n in $\hat{\theta}_n(X_n)$ asymptotically cancels the randomness which enters through X_n in $\mathscr{R}_n(X_n)$: the Bayesian tolerance regions $\mathscr{R}_n(X_n)$ should be "asymptotically pivotal" with respect to the estimators $\hat{\theta}_n$. Some type of stability condition appears to be necessary, because the shape of a Bayesian tolerance region is left free by its definition.

An informal proof of the frequentist validity of Bayesian tolerance regions is as follows. Replacing the posterior distribution in (1.6) by its normal approximation (1.12) from the Bernstein-von Mises theorem, we find that

$$\int \ell\Big[Q_\vartheta\Big(\mathscr{R}_n(X_n)\Big)\Big]\, dN\left(\hat{\theta}_n, \frac{1}{n}\Sigma_\theta\right)(\vartheta) \doteq 1 - \alpha. \tag{1.15}$$

By the substitution $\vartheta = \hat{\theta}_n + h/\sqrt{n}$ this can be rewritten in the form

$$\int \ell\Big[Q_{\hat{\theta}_n+h/\sqrt{n}}\Big(\mathscr{R}_n(X_n)\Big)\Big]\, dN\Big(0, \Sigma_\theta\Big)(h) \doteq 1 - \alpha. \tag{1.16}$$

By the stability assumption the integrand

$$h \mapsto g_n(h; X_n) := \ell\Big[Q_{\hat{\theta}_n+h/\sqrt{n}}\Big(\mathscr{R}_n(X_n)\Big)\Big] \tag{1.17}$$

in this expression is asymptotically the same as a deterministic function $h \mapsto g_\infty(h)$. In view of (1.13) the integral in (1.16) is then approximately equal to $\mathrm{E}_\theta g_\infty\Big(\sqrt{n}(\theta - \hat{\theta}_n)\Big)$, which in turn, again by stability, is asymptotically the same as $\mathrm{E}_\theta g_n\Big(\sqrt{n}(\theta - \hat{\theta}_n); X_n\Big)$, or

$$\mathrm{E}_\theta \ell\Big[Q_{\hat{\theta}_n+\sqrt{n}(\theta-\hat{\theta}_n)/\sqrt{n}}\Big(\mathscr{R}_n(X_n)\Big)\Big] = \mathrm{E}_\theta \ell\Big[Q_\theta\Big(\mathscr{R}_n(X_n)\Big)\Big].$$

Thus the final expression, which is the frequentist level of the tolerance region $\mathscr{R}_n(X_n)$, is asymptotically equal to $1-\alpha$.

For an (δ,α)-tolerance region $\ell\Big(Q_\theta(\mathscr{R}_n(X_n)\Big)$ is the indicator of the set $\hat{G}_n = \{\theta: Q_\theta\Big(\mathscr{R}_n(X_n)\Big) \geqslant 1-\delta\}$ and the function (1.17) is the indicator of the set

$$\hat{H}_n = \sqrt{n}(\hat{G}_n - \hat{\theta}_n).$$

Thus the stability condition is that the latter sets approximate to a deterministic set, as $n \to \infty$. Condition (1.16) becomes

$$N(0,\Sigma_\theta)(\hat{H}_n) \doteq 1-\alpha. \tag{1.18}$$

This equality allows to "solve" one aspect of the sets $\hat{H}_n$; in general additional constraints will be imposed to define their shape. As the normal distribution in this display is fixed, it is not unnatural that these constraints would render the sets $\hat{H}_n$ also to become fixed, in the limit: stability is natural.

Theorem 1.5.1. *Suppose that* (1.12)–(1.13) *hold, the loss function ℓ is bounded, and suppose that there exist (deterministic) functions $f_{n,1}, f_{n,2}\colon \mathbb{R}^d \to \mathbb{R}$ with the property that $f_{n,i}(h_n) \to f_\infty(h)$ for some function f_∞ and any sequence $h_n \to h$ with limit h in a set of probability one under the normal distribution in* (1.13) *and such that*

$$f_{n,1}(h) \leqslant \ell\Big[Q_{\hat{\theta}_n + h/\sqrt{n}}\Big(\mathscr{R}_n(X_n)\Big)\Big] \leqslant f_{n,2}(h), \qquad h \in \mathbb{R}^d. \tag{1.19}$$

Then $\int \ell\Big(Q_\theta(\mathscr{R}_n(X_n))\Big)\, d\Pi(\theta|X_n) \to 1-\alpha \in (0,1)$ in probability implies that $\mathrm{E}_\theta \ell\Big(Q_\theta(\mathscr{R}_n(X_n))\Big) \to 1-\alpha$, as $n\to\infty$, for every θ.

Proof. We may assume without loss of generality that the functions $f_{n,i}$ are uniformly bounded. Then the condition $f_{n,i}(h_n) \to f_\infty(h)$ for every sequence $h_n \to h$ implies that $\mathrm{E} f_{n,i}(Y_n) \to \mathrm{E} f_\infty(Y)$, whenever the sequence of random vectors Y_n tends in distribution to the random vector Y, in view of the extended continuous mapping theorem (see [58], Theorem 1.11.1). Thus by (1.13), for $i = 1,2$,

$$\mathrm{E}_\theta f_{n,i}\Big(\sqrt{n}(\hat{\theta}_n - \theta)\Big) \to \int f_\infty\, dN(0,\Sigma_\theta).$$

By assumption the function g_n given in (1.17) is sandwiched between $f_{n,1}$ and $f_{n,2}$. Therefore $\mathrm{E}_\theta \ell\Big(Q_\theta(\mathscr{R}_n(X_n))\Big) = \mathrm{E}_\theta g_n\Big(\sqrt{n}(\hat{\theta}_n - \theta); X_n\Big)$ tends to the same limit.

By (1.12) we have

$$\int \ell\Big(Q_\theta(\mathscr{R}_n(X_n))\Big)\, d\Pi(\theta|X_n) = \int g_n(h;X_n)\, dN(0,\Sigma_\theta)(h) + o(1).$$

Again by sandwiching of $g_n(h;X_n)$ this is asymptotic to $\int f_{n,i}\, dN(0,\Sigma_\theta)$, and hence tends to $\int f_\infty\, dN(0,\Sigma_\theta)$.

■

1.5.1 Normal predictions

An (α,δ) tolerance interval $\mathscr{R}_n(X_n)=[A_n-B_n,A_n+B_n]$ for a one-dimensional Gaussian variable $Z\sim N(\nu,\tau^2)$ is the base (the section at $\tau=0$) of a set of the form

$$G_{A,B,\delta}=\left\{\theta=(\nu,\tau)\colon \Phi\left(\frac{A+B-\nu}{\tau}\right)-\Phi\left(\frac{A-B-\nu}{\tau}\right)\geqslant 1-\delta\right\}.$$

The values A_n and B_n are determined so that $\Pi(G_{A_n,B_n,\delta}|X_n)\geqslant 1-\alpha$, for $\Pi(\cdot|X_n)$ the posterior distribution of $\theta=(\nu,\tau)$, and so that the length $2B_n$ of the interval is minimal.

Under (1.12) the posterior distribution contracts (at rate $1/\sqrt{n}$) to the Dirac measure at $\hat\theta$, which tends to the true parameter under (1.13). Hence the equation $\Pi(G_{A_n,B_n,\delta}|X_n)\geqslant 1-\alpha$ forces $\hat\theta=(\hat\nu,\hat\tau)$ to be contained in $G_{A_n,B_n,\delta}$ with probability tending to one. This implies that $A_n-\hat\nu\to 0$ in probability and hence $\Phi(B_n/\hat\tau)-\Phi(-B_n/\hat\tau)\to 1-\delta$, whence $B_n/\hat\tau\to\xi_{\delta/2}$, for ξ_δ the upper δ-quantile of the standard normal distribution.

The function $\theta\mapsto \ell\Big(Q_\theta([A-B,A+B])\Big)$ corresponding to the (α,δ) tolerance interval is the indicator of the set $G_{A,B,\delta}$, and the stability condition (1.19) is that the (indicator functions) of the sets $\hat H_n=\sqrt{n}(G_{A_n,B_n,\delta}-\hat\theta)$ are asymptotically deterministic. These sets can be written $\hat H_n=\{(g,h)\colon K_n(g/\sqrt{n},h/\sqrt{n})\geqslant 0\}$, for the stochastic processes

$$K_n(g,h)=\Phi\left(\frac{A_n+B_n-\hat\nu-g}{\hat\tau+h}\right)-\Phi\left(\frac{A_n-B_n-\hat\nu-g}{\hat\tau+h}\right)-(1-\delta). \tag{1.20}$$

By a second-order Taylor expansion we see that these processes satisfy (1.21) with

$$\begin{aligned}\hat a_n=K_n(0,0)&=\Phi\left(\frac{A_n+B_n-\hat\nu}{\hat\tau}\right)-\Phi\left(\frac{A_n-B_n-\hat\nu}{\hat\tau}\right)-(1-\delta),\\ \hat b_n=\frac{\partial}{\partial h}K_n(0,0)&=\psi\left(\frac{A_n+B_n-\hat\nu}{\hat\tau}\right)\frac{1}{\hat\tau}-\psi\left(\frac{A_n-B_n-\hat\nu}{\hat\tau}\right)\frac{1}{\hat\tau},\\ V(g,h)&=h.\end{aligned}$$

Here $\psi(x)=\phi(x)x=-\phi'(x)$. (Note that the partial derivative of K_n relative to its first argument g vanishes at $(0,0)$.) Since A_n, B_n, $\hat\nu$ and $\hat\tau$ tend in probability to nontrivial limits, the conditions of Lemma 1.5.1 are satisfied and hence the sets $\hat H_n$ are asymptotically sandwiched between pairs of deterministic sets. Functions $f_{n,i}$ as in (1.19) can be constructed from these sets by letting $\varepsilon\to 0$ and $M\to\infty$ slowly with n.

It follows that the conditions of Theorem 1.19 are satisfied and hence the Bayesian (α,δ) tolerance sets are asymptotic frequentist (α,δ) tolerance sets.

The convergence $A_n-\hat\nu\to 0$ in probability means that the tolerance intervals are asymptotically centered at the (asymptotic) posterior mean. If the posterior distribution of θ is exactly normal

$N\Big((\hat\nu,\hat\tau),\Sigma_\theta\Big)$ with a diagonal covariance matrix, then $\mathrm{E}(\nu|X_n,\tau)$ is free of τ and Lemma 1.3.1 shows that the tolerance interval is centered exactly at the posterior mean. This is more generally true if the posterior distribution of ν given (τ,X_n) is normal with mean free of τ.

For a nondiagonal matrix Σ_θ this is not necessarily true, and in general the normal distribution will be an approximation only. The approximation $A_n-\hat\nu\to 0$ can then be improved to order $n^{-1/4}$, and an asymptotic expression for the half length B_n of the interval is as follows.

Theorem 1.5.2. *If the Bernstein-von Mises theorem* (1.12)–(1.13) *holds, then the Bayesian* (α,δ) *tolerance interval* $[A_n-B_n,A_n+B_n]$ *of minimal length is a frequentist* (α,δ) *tolerance interval. Its center and half length satisfy* $A_n=\hat\nu+o_P(n^{-1/4})$ *and* $B_n=\hat\tau\xi_{\delta/2}+(\xi_\alpha/\xi_{\delta/2})n^{-1/2}+o_P(n^{-1/2})$.

Proof. Since $A_n-\hat\nu\to 0$ and $B_n\to\xi_{\delta/2}$, there exist intervals I_n and J_n around ν_0 and $\xi_{\delta/2}$ that shrink to these points that contain A_n and B_n with probability tending to one. Define functions $F_n,G_n\colon I_n\times J_n\to\mathbb{R}$ by

$$\begin{aligned}F_n(A,B)&=\Pi(G_{A,B,\delta}|X_n),\\ G_n(A,B)&=N(0,\Sigma_\theta)\left\{(g,h)\colon h\leqslant\frac{\sqrt{n}K_n(0,0;A,B)}{-2\psi(\xi_{d/2})/\hat\tau}\right\},\end{aligned}$$

Here $K_n(h,g;A,B)$ is the expression on the right side of (1.20), but with (A_n,B_n) replaced by a generic (A,B). We shall show that these functions satisfy the conditions of Lemma 1.5.2 with $c_n=\hat\tau\xi_{\delta/2}$ and $\xi(\alpha)=(\xi_\alpha/\xi_{\delta/2})n^{-1/2}$, whence the theorem follows from the lemma.

In view of (1.12) we have $\sup_{A,B}\big|F_n(A,B)-N(0,\Sigma_\theta)(\hat H_{A,B,\delta})\big|\to 0$, for

$$\hat H_{A,B,\delta}=\Big\{(g,h)\colon K_n(g/\sqrt{n},h/\sqrt{n};A,B)\geqslant 0\Big\}.$$

The supremum here and in the following is taken over $(A,B)\in I_n\times J_n$. By a second-order Taylor expansion, for functions $\eta_{n,1}$ and $\eta_{n,2}$ with $\sup_{A,B}\eta_{n,i}(A,B)\to 0$ and a constant b independent of (A,B),

$$\begin{aligned}\sqrt{n}\Big|&K_n(g/\sqrt{n},h/\sqrt{n};A,B)-K_n(0,0;A,B)\\ &-\Big(-2\psi(\xi_{\delta/2})/\hat\tau+\eta_{n,1}(A,B)\Big)\frac{g}{\sqrt{n}}+\eta_{n,2}(A,B)\frac{h}{\sqrt{n}}\Big|\leqslant b\frac{g^2+h^2}{\sqrt{n}}.\end{aligned}$$

As in the proof of Lemma 1.5.1, this shows that, with $C_n=\{(g,h)\colon g^2+h^2\leqslant L_n\}$ and $L_n\to\infty$ sufficiently slowly

$$\sup_{A,B}\Big|N(0,\Sigma_\theta)(\hat H_{A,B,\delta}\cap C_n)-N(0,\Sigma_\theta)\left\{(g,h)\in C_n\colon h\leqslant\frac{\sqrt{n}K_n(0,0;A,B)}{-2\psi(\xi_{\delta/2})/\hat\tau}\right\}\Big|\to 0.$$

Because $N(0,\Sigma_\theta)(C_n^c) \to 0$, this is then also true without intersecting the sets by C_n. This finishes the proof that $\sup_{A,B}\big|F_n(A,B) - G_n(A,B)\big| \to 0$.

From the unimodality of the normal distribution it is clear that $A \mapsto G_n(A,B)$ is maximal at $A = \hat{\nu}$, for every B. It is elementary to verify by Taylor expansion that $G_n(\hat{\nu},\hat{B}) \to 1-\alpha$, for $\hat{B} = \hat{\tau}\xi_{\delta/2} + (\xi_\alpha/\xi_{\delta/2})n^{-1/2}$.

Finally, again by Taylor expansion there exist functions $\eta_{n,3}$ with $\sup_{A,B}\eta_{n,3}(A,B) \to 0$ such that

$$K_n(0,0,A,\hat{B}) \leqslant K_n(0,0,\hat{\nu},\hat{B}) + (A-\hat{\nu})^2\Big(\phi'(\hat{B}/\hat{\tau})/\hat{\tau} + \eta_{n,3}(A,B)\Big).$$

As $\phi'(x) < 0$, for $x > 0$, we obtain that $\sqrt{n}K_n(0,0,A,\hat{B})$ is strictly smaller than $\sqrt{n}K_n(0,0,\hat{\nu},\hat{B})$ if $\sqrt{n}(A-\hat{\nu})^2$ is bounded away from zero, but then

$$N(0,\Sigma_\theta)\left\{(g,h): h \leqslant \frac{\sqrt{n}K_n(0,0;A,\hat{B})}{-2\psi(\xi_{\delta/2})/\hat{\tau}}\right\}$$

is strictly smaller than its asymptotic value $1-\alpha$ at $A = \hat{\nu}$ unless $\sqrt{n}(A-\hat{\nu})^2 \to 0$. This verifies the last displayed condition of Lemma 1.5.2. ∎

Lemma 1.5.1. *Suppose that for every $M > 0$ the stochastic processes $(K_n(h): h \in \mathbb{R}^d)$ satisfy*

$$\sup_{\|h\| \leqslant M/\sqrt{n}} \sqrt{n}\big|K_n(h) - \hat{a}_n + \hat{b}_n V h\big| \xrightarrow{P} 0, \tag{1.21}$$

for random variables $\hat{a}_n$ and $\hat{b}_n > 0$ such that $\hat{b}_n^{-1}$ is bounded in probability, and a linear map $V: \mathbb{R}^d \to \mathbb{R}$. If the sets $\hat{H}_n = \{h \in \mathbb{R}^d: K_n(h/\sqrt{n}) \geqslant 0\}$ satisfy $N(0,\Sigma)(\hat{H}_n) \to 1-\alpha \in (0,1)$, then $\sqrt{n}\hat{a}_n/\hat{b}_n \to \xi_\alpha\sqrt{V\Sigma V^T}$ and for every $\varepsilon, M > 0$, with probability tending to 1,

$$\Big\{h \in \mathbb{R}^d: \|h\| \leqslant M, Vh \leqslant \xi_\alpha\sqrt{V\Sigma V^T} - \varepsilon\Big\} \subset \hat{H}_n \subset$$
$$\subset \Big\{h \in \mathbb{R}^d: Vh \leqslant \xi_\alpha\sqrt{V\Sigma V^T} + \varepsilon \text{ or } \|h\| > M\Big\}.$$

Proof. Define $\varepsilon_n(h) = \sqrt{n}\Big(K_n(h) - \hat{a}_n + \hat{b}_n V h\Big)$ and set $\hat{\varepsilon}_n = \sup_{\|h\| \leqslant M/\sqrt{n}}|\varepsilon_n(h)|$, for given M. Then by assumption $\hat{\varepsilon}_n \to 0$ in probability, and $\big|K_n(h/\sqrt{n}) - \hat{a}_n + \hat{b}_n V h/\sqrt{n}\big| \leqslant \hat{\varepsilon}_n$, for every h with $\|h\| \leqslant M$. From the latter inequality we find that

$$\|h\| \leqslant M, K_n(h/\sqrt{n}) \geqslant 0 \Rightarrow \hat{b}_n V h \leqslant \sqrt{n}\hat{a}_n + \hat{\varepsilon}_n,$$
$$\|h\| \leqslant M, \hat{b}_n V h \leqslant \sqrt{n}\hat{a}_n - \hat{\varepsilon}_n \Rightarrow K_n(h/\sqrt{n}) \geqslant 0.$$

This implies that

$$\Big\{\|h\| \leqslant M, Vh \leqslant (\sqrt{n}\hat{a}_n - \hat{\varepsilon}_n)/\hat{b}_n\Big\} \subset \hat{H}_n \subset \Big\{Vh \leqslant (\sqrt{n}\hat{a}_n + \hat{\varepsilon}_n)/\hat{b}_n \text{ or } \|h\| > M\Big\}.$$

Combining this with the fact that $N(0,\Sigma)(\hat{H}_n) \to 1-\alpha \in (0,1)$ and the fact that $Vh \sim N(0, V\Sigma V^T)$ if $h \sim N(0,\Sigma)$, we conclude that there exists $\delta_M > 0$ such that $\delta_M \to 0$ as $M \to \infty$ such that $(\sqrt{n}\hat{a}_n - \hat{\varepsilon}_n)/\hat{b}_n \leqslant \xi_{\alpha-\delta_M}\sqrt{V\Sigma V^T} + o_P(1)$ and $(\sqrt{n}\hat{a}_n + \hat{\varepsilon}_n)/\hat{b}_n \geqslant \xi_{\alpha+\delta_M}\sqrt{V\Sigma V^T} + o_P(1)$, for every M. This implies that $\sqrt{n}\hat{a}_n/\hat{b}_n = \xi_\alpha\sqrt{V\Sigma V^T} + o_P(1)$. We substitute this in the last display to obtain the result of the lemma. ∎

Lemma 1.5.2. *Let $F_n, G_n : I_n \times J_n \to \mathbb{R}$ be functions on rectangles $I_n \times J_n \subset \mathbb{R}^2$ that are nondecreasing in their second argument, such that*

$$\sup_{A,B} |F_n(A,B) - G_n(A,B)| \to 0,$$

and such that for numbers $c_n \in J_n$ and continuous functions $\xi : (0,1) \to \mathbb{R}$,

$$\sup_A G_n(A, c_n + \xi(\alpha)/\sqrt{n}) = G_n(0, c_n + \xi(\alpha)/\sqrt{n}) \to 1-\alpha, \quad \text{every } \alpha \in (0,1),$$

$$\limsup_n \sup_{A:|A|>\delta_n n^{-1/4}} G_n(A, c + \xi(\alpha)/\sqrt{n}) < 1-\alpha, \quad \text{some } \delta_n \to 0.$$

Then $B_n := \inf\big(B : \sup_A F_n(A,B) \geqslant 1-\alpha\big)$ satisfies $B_n = c_n + \xi(\alpha)/\sqrt{n} + o(n^{-1/2})$, and $\hat{A} := \operatorname{argmax}_A F_n(A, B_n)$ satisfies $\hat{A} = o(n^{-1/4})$.

Proof. The functions $\bar{F}_n$ and $\bar{G}_n$ defined by $\bar{F}_n(B) = \sup_A F_n(A,B)$ and similarly for G_n satisfy $\sup_B |\bar{F}_n(B) - \bar{G}_n(B)| \to 0$. Combined with the assumptions on G_n, this gives that $\bar{F}_n(c_n + \xi(\alpha)/\sqrt{n}) \to 1-\alpha$, for every α. The definition of B_n and monotonicity of $\bar{F}_n$ now readily give that $c_n + \xi(\alpha_2)/\sqrt{n} \leqslant B_n \leqslant c_n + \xi(\alpha_1)/\sqrt{n}$ for every $\alpha_1 < \alpha < \alpha_2$, eventually, or equivalently $\xi(\alpha_1) \leqslant \sqrt{n}(B_n - c_n) \leqslant \xi(\alpha_2)$, eventually. By the continuity of ξ it follows that $\sqrt{n}(B_n - c_n) \to \xi(\alpha)$.

By the uniform approximation of F_n by G_n, we have that

$$\sup_{|A|>\delta_n n^{-1/4}} F_n(A, B_n) = \sup_{|A|>\delta_n n^{-1/4}} G_n(A, B_n) + o(1),$$

which is strictly smaller than $1-\alpha$, eventually, by assumption. Simiarly $F_n(0, B_n) = G_n(0, B_n) + o(1) \to 1-\alpha$. It follows that the maximum of $A \mapsto F_n(A, B_n)$ is taken on the interval $(-\delta_n n^{-1/4}, \delta_n n^{-1/4})$.

∎

1.6 Annex : Extra Insights on KMW solution in Section 1.3.1

Let us rewrite (1.8) as

$$\Pi\Bigg[\overbrace{\bigg(Q_\theta(-\infty,\hat{A}+B]-1+\frac{\delta}{2}\bigg)}^{E_1}-\overbrace{\bigg(Q_\theta(-\infty,\hat{A}-B]-\frac{\delta}{2}\bigg)}^{E_2}\geq 0\,|X\Bigg]=1-\alpha.$$

The KM criterion is to seek a value of B (denoted as $\hat{B}^*$) satisfying $\Pi(E_1\geq 0,E_2\leq 0|X)=1-\alpha$. With provided $\hat{A}$ and δ, it should be clear that $\{\theta\!:E_1\geq 0,E_2\leq 0\}\subset\{\theta\!:E_1-E_2\geq 0\}$ for $\forall B>0$, and both sets expand as B grows. Therefore, the B satisfying $\Pi\Big(\{\theta\!:E_1\geq 0,E_2\leq 0\}|X\Big)=1-\alpha$ must be larger than the B satisfying $\Pi\Big(\{\theta\!:E_1-E_2\geq 0\}|X\Big)=1-\alpha$. This is identical to say, if we plug $\hat{B}^*$ into $\Pi(E_1-E_2\geq 0|X)$, it results in a total posterior mass strictly larger than $1-\alpha$.

Now, rewrite (1.8) as

$$\Pi\bigg[\Big(1-Q_\theta(-\infty,\hat{A}+B]\Big)+Q_\theta(-\infty,\hat{A}-B]\leq\delta\,|X\bigg]=1-\alpha.$$

The W criterion tries to optimize

$$\Pi\Bigg[\overbrace{\bigg(1-\frac{\delta}{2}-Q_\theta(-\infty,\hat{A}+B]\bigg)}^{E_1}+\overbrace{\bigg(Q_\theta(-\infty,\hat{A}-B]-\frac{\delta}{2}\bigg)}^{E_2}>0\,|X\Bigg]=\alpha,$$

by requiring $\Pi(E_1\geq 0,E_2\geq 0|X)=\alpha$, which is more stringent than it should be. Follow the similar reasoning, the estimate $\hat{B}^*$ fulfilling the W criterion will result in a total posterior mass strictly larger than α, and in turn less than $1-\alpha$ on its dual side.

1.7 Annex: Gibbs Sampler for one-way Random Effect model

In this annex, we work out the Gibbs sampling scheme for deriving Bayesian estimates for a one-way random effect model. We consider two common formulations of one-way the random effect model together with two choices of priors, namely:

Formulation (a) $y_{ij}=\mu_i+\varepsilon_{ij},\quad \mu_i\sim N(\mu,\tau^2)$

Formulation (b) $y_{ij}=\mu+a_i+\varepsilon_{ij},\quad a_i\sim N(0,\tau^2)$

Prior (a) $\mu\sim N(\mu_0,\sigma_0^2)$ where μ_0 and σ_0^2 are given

Prior (b) $\mu\sim N(\mu_0,\sigma_0^2)$ where μ_0 is given and $\sigma_0^2\sim IG(a_0,b_0)$

This gives four combinations, which are treated in Sections 1.7.1-1.7.4, with more precise definitions in the start of each section. To be clear, the density function of a inverse-gamma (denoted by IG) distribution is parameterized in the following way:

$$x \sim IG(\alpha,\beta), \quad f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} exp\left[\frac{-\beta}{x}\right], \quad (x>0).$$

1.7.1 Scenario 1: Formulation (a) and Prior (a)

Given data points y_{ij}, being the j^{th} measure from the i^{th} group, we assume that

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \mu_i \sim N(\mu,\tau^2), \quad \varepsilon_{ij} \sim N(0,\sigma^2)$$

where $i \in \{1,2,\cdots,m\}$, $j \in \{1,2,\cdots,n_i\}$, and any pair from $\{\mu_i\} \cup \{\varepsilon_{ij}\}$ are assumed mutually independent. To further simplify the expressions, let θ denote the vector of parameters $(\mu,\tau,\sigma)'$ that govern the distribution of y_{ij}, and let y_i be the vector $\left(y_{i1}, y_{i2}, \cdots, y_{in_i}\right)'$. Then y_i is independent to $y_{i'}$ for all $i \neq i'$ given θ. The prior of θ is set as the following:

$$p(\theta) = p(\mu)p(\tau^2)p(\sigma^2), \quad where$$

$$\mu \sim N(\mu_0, \sigma_0^2), \quad \tau^2 \sim IG(a_1,b_1), \quad \sigma^2 \sim IG(a_2,b_2)$$

$$\mu_0, \sigma_0^2, a_1, b_1, a_2, b_2 \text{ are given.}$$

The posterior can then be derived as below.

$$\begin{aligned} p(\theta \mid Data) &\propto p(\theta)p(Data \mid \theta) \\ &= p(\theta)p(y_1,\cdots,y_m \mid \mu_1,\cdots,\mu_m,\mu,\tau^2,\sigma^2) \\ &= \frac{p(\theta)p(y_1,\cdots,y_m,\mu_1,\cdots,\mu_m \mid \mu,\tau^2,\sigma^2)}{p(\mu_1,\cdots,\mu_m \mid \mu,\tau^2)} \\ &= \frac{p(\theta)\mathscr{Q}}{p(\mu_1,\cdots,\mu_m \mid \mu,\tau^2)} \propto p(\theta)\mathscr{Q} \end{aligned}$$

First, we work out the details of $\mathscr{Q}$:

$$\begin{aligned}
\mathscr{Q} &= p(y_1,\cdots,y_m \mid \mu_1,\cdots,\mu_m,\mu,\tau^2,\sigma^2)p(\mu_1,\cdots,\mu_m \mid \mu,\tau^2) \\
&= \prod_{i=1}^{m}\left((2\pi)^{-\frac{1}{2}}(\tau^2)^{-\frac{1}{2}}\exp\left[-\frac{(\mu_i-\mu)^2}{2\tau^2}\right]\prod_{j=1}^{n_i}(2\pi)^{-\frac{1}{2}}(\sigma^2)^{-\frac{1}{2}}\exp\left[-\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}\right]\right) \\
&= \prod_{i=1}^{m}\left((2\pi)^{-\frac{1}{2}}(\tau^2)^{-\frac{1}{2}}\exp\left[-\frac{(\mu_i-\mu)^2}{2\tau^2}\right](2\pi)^{-\frac{n_i}{2}}(\sigma^2)^{-\frac{n_i}{2}}\exp\left[-\frac{\sum_j(y_{ij}-\mu_i)^2}{2\sigma^2}\right]\right) \\
&= (2\pi)^{\frac{-m-\sum n_i}{2}}(\tau^2)^{-\frac{m}{2}}(\sigma^2)^{-\frac{\sum n_i}{2}}\exp\left[-\frac{\sum_i(\mu_i-\mu)^2}{2\tau^2}\right]\exp\left[-\frac{\sum_i\sum_j(y_{ij}-\mu_i)^2}{2\sigma^2}\right] \\
&= (2\pi)^{\frac{-m-\sum n_i}{2}}(\tau^2)^{-\frac{m}{2}}(\sigma^2)^{-\frac{\sum n_i}{2}}\exp\left[-\frac{\sum_i\mu_i^2-2\mu\sum_i\mu_i+m\mu^2}{2\tau^2}\right] \\
&\quad \exp\left[-\frac{\sum_i\sum_j y_{ij}^2-2\sum_i\mu_i n_i\bar{y}_i+\sum_i n_i\mu_i^2}{2\sigma^2}\right] \qquad (1.22)
\end{aligned}$$

Below we derive the conditional distributions of each element of θ given the other coordinates and the observations, the so-called full conditionals, used in a Gibbs sampling scheme. For brevity we denote the full set of conditioning variables by $\cdots$, even though this set is different in different instances. It is helpful to remember, that a quadratic function Ax^2+Bx+c can be re-written as $A(x+\frac{B}{2A})^2-\frac{B^2}{2A}+C$ if $A\neq 0$. This allows to recognize the values of the mean and the variance if the conditional posterior is Gaussian.

Posterior for μ_i, $\forall i\in\{1,2,\cdots,m\}$

$$\begin{aligned}
p(\mu_i \mid \cdots) &\propto \exp\left[-\frac{\mu_i^2-2\mu\mu_i}{2\tau^2}-\frac{-2\mu_i n_i\bar{y}_i+n_i\mu_i^2}{2\sigma^2}\right] \\
&\propto \exp\left[-\frac{(\sigma^2+n_i\tau^2)\mu_i^2-2(n_i\bar{y}_i\tau^2+\mu\sigma^2)\mu_i}{2\tau^2\sigma^2}\right] \qquad (1.23)
\end{aligned}$$

The above formula implies that $\mu_i \mid \cdots \sim N(\frac{n_i\bar{y}_i\tau^2+\mu\sigma^2}{\sigma^2+n_i\tau^2},\frac{\tau^2\sigma^2}{\sigma^2+n_i\tau^2})$.

Posterior for μ

$$\begin{aligned}
p(\mu \mid \cdots) &\propto \exp\left[-\frac{\mu^2-2\mu_0\mu+\mu_0^2}{2\,\sigma_0^2}-\frac{m\mu^2-2\mu\sum_i\mu_i}{2\tau^2}\right] \\
&\propto \exp\left[-\frac{(\tau^2+m\,\sigma_0^2)\mu^2-2(\mu_0\tau^2+\sum_i\mu_i\,\sigma_0^2)\mu+\mu_0^2\tau^2}{2\,\sigma_0^2\tau^2}\right] \qquad (1.24)
\end{aligned}$$

The above formula implies that $\mu \mid \cdots \sim N(\frac{\mu_0\tau^2+\sum_i \mu_i \sigma_0^2}{\tau^2+m\sigma_0^2}, \frac{\sigma_0^2\tau^2}{\tau^2+m\sigma_0^2})$.

Posterior for τ^2

$$\begin{aligned} p(\tau^2 \mid \cdots) &\propto (\tau^2)^{-a_1-1}\exp\left[-\frac{b_1}{\tau^2}\right](\tau^2)^{-\frac{m}{2}}\exp\left[-\frac{\sum_i(\mu_i-\mu)^2}{2\tau^2}\right] \\ &\propto (\tau^2)^{-(a_1+\frac{m}{2})-1}\exp\left[-\frac{2b_1+\sum_i(\mu_i-\mu)^2}{2\tau^2}\right] \end{aligned} \tag{1.25}$$

The above formula implies that $\tau^2 \mid \cdots \sim IG(a_1+\frac{m}{2}, b_1+\frac{\sum_i(\mu_i-\mu)^2}{2})$.

Posterior for σ^2

$$\begin{aligned} p(\sigma^2 \mid \cdots) &\propto (\sigma^2)^{-a_2-1}\exp\left[-\frac{b_2}{\sigma^2}\right](\sigma^2)^{-\frac{\sum_i n_i}{2}}\exp\left[-\frac{\sum_i\sum_j(y_{ij}-\mu_i)^2}{2\sigma^2}\right] \\ &\propto (\sigma^2)^{-(a_2+\frac{\sum_i n_i}{2})-1}\exp\left[-\frac{2b_2+\sum_i\sum_j(y_{ij}-\mu_i)^2}{2\sigma^2}\right] \end{aligned} \tag{1.26}$$

The above formula implies that $\sigma^2 \mid \cdots \sim IG(a_2+\frac{\sum_i n_i}{2}, b_2+\frac{\sum_i\sum_j(y_{ij}-\mu_i)^2}{2})$.

1.7.2 Scenario 2: Formulation (a) and Prior (b)

Keep the setup in Scenario 1 except its prior $p(\theta)$, which will now be replaced by the following

$$p(\theta) = p(\mu)p(\tau^2)p(\sigma^2), \quad \textit{where}$$
$$\mu \sim N(\mu_0, \sigma_0^2), \quad \sigma_0^2 \sim IG(a_0,b_0), \quad \tau^2 \sim IG(a_1,b_1), \quad \sigma^2 \sim IG(a_2,b_2)$$
$$\mu_0, a_0, b_0, a_1, b_1, a_2, b_2 \textit{ are all given.}$$

It can be seen that $\mathscr{Q}$ will stay the same as in (1.22) under this construction of the prior for μ, and hence the full conditionals of $\mu_i, \mu, \tau^2, \sigma^2$ remain unchanged as in (1.23) - (1.26). The full conditional of σ_0^2 is also needed to build the Gibbs sampler, and can be derived as follows.

$$\begin{aligned} p(\sigma_0^2 \mid \cdots) &\propto (\sigma_0^2)^{-a_0-1}\exp\left[-\frac{b_0}{\sigma_0^2}\right](\sigma_0^2)^{-\frac{1}{2}}\exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right] \\ &\propto (\sigma_0^2)^{-(a_0+\frac{1}{2})-1}\exp\left[-\frac{2b_0+(\mu-\mu_0)^2}{2\sigma_0^2}\right] \end{aligned} \tag{1.27}$$

The above formula implies that $\sigma_0^2 \mid \cdots \sim IG(a_0 + \frac{1}{2}, b_0 + \frac{(\mu-\mu_0)^2}{2})$.

1.7.3 Scenario 3: Formulation (b) and Prior (a)

We now assume a different formulation for the distribution of y_{ij}, suggested in [17] among others, which is believed to lead to a better performance of the resulting Gibbs sampler. Let

$$y_{ij} = \mu + a_i + \varepsilon_{ij}, \quad a_i \sim N(0, \tau^2), \quad \varepsilon \sim N(0, \sigma^2)$$

where $i \in \{1, 2, \cdots, m\}$, $j \in \{1, 2, \cdots, n_i\}$, and the variables in the set $\{a_i\}_i \cup \{\varepsilon_{ij}\}_{i,j}$ are assumed mutually independent. The rest of the setting is the same as in Scenario 1. With these updated notations, the posterior is given by

$$\begin{aligned} p(\theta \mid Data) &\propto \frac{p(\theta) p(y_1, \cdots, y_m, a_1, \cdots, a_m \mid \mu, \tau^2, \sigma^2)}{p(a_1, \cdots, a_m \mid \tau^2)} \\ &= \frac{p(\theta)\mathscr{Q}'}{p(a_1, \cdots, a_m \mid \tau^2)} \propto p(\theta)\mathscr{Q}' \end{aligned}$$

where $\mathscr{Q}'$ can be elaborated as below.

$$\begin{aligned} \mathscr{Q}' &= p(y_1, \cdots, y_m \mid a_1, \cdots, a_m, \mu, \tau^2, \sigma^2) p(a_1, \cdots, a_m \mid \mu, \tau^2) \\ &= \prod_{i=1}^{m} \left((2\pi)^{-\frac{1}{2}} (\tau^2)^{-\frac{1}{2}} \exp\left[-\frac{a_i^2}{2\tau^2} \right] \prod_{j=1}^{n_i} (2\pi)^{-\frac{1}{2}} (\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{(y_{ij} - \mu - a_i)^2}{2\sigma^2} \right] \right) \\ &= (2\pi)^{\frac{-m-\sum n_i}{2}} (\tau^2)^{-\frac{m}{2}} (\sigma^2)^{-\frac{\sum n_i}{2}} \exp\left[-\frac{a_i^2}{2\tau^2} \right] \exp\left[-\frac{\sum_i \sum_j (y_{ij} - \mu - a_i)^2}{2\sigma^2} \right] \end{aligned} \tag{1.28}$$

The last factor in (1.28) can be further detailed as

$$\exp\left[-\frac{\sum_i \sum_j y_{ij}^2 - 2\mu \sum_i (n_i \bar{y}_i) - 2\sum_i (n_i a_i \bar{y}_i) + \mu^2 \sum_i n_i + 2\mu \sum_i (n_i a_i) + \sum_i n_i a_i^2}{2\sigma^2} \right] \tag{1.29}$$

Posterior for a_i, $\forall i \in \{1, 2, \cdots, m\}$

$$\begin{aligned} p(a_i \mid \cdots) &\propto \exp\left[-\frac{a_i^2}{2\tau^2} - \frac{-2(n_i a_i \bar{y}_i) + 2\mu n_i a_i + n_i a_i^2}{2\sigma^2} \right] \\ &\propto \exp\left[-\frac{(\sigma^2 + n_i\tau^2) a_i^2 - 2\tau^2 n_i (\bar{y}_i - \mu) a_i^2}{2\tau^2\sigma^2} \right] \end{aligned} \tag{1.30}$$

The above formula implies that $a_i \mid \cdots \sim N(\frac{\tau^2 n_i (\bar{y}_i - \mu)}{\sigma^2 + n_i \tau^2}, \frac{\tau^2 \sigma^2}{\sigma^2 + n_i \tau^2})$.

Posterior for μ

$$\begin{aligned} p(\mu \mid \cdots) &\propto \exp\left[-\frac{\mu^2-2\mu_0\mu+\mu_0^2}{2\,\sigma_0^2}-\frac{-2\mu\sum_i(n_i\bar{y}_i)+\mu^2\sum_i n_i+2\mu\sum_i(n_i a_i)}{2\sigma^2}\right] \\ &\propto \exp\left[-\frac{(\sigma^2+\sum_i n_i\,\sigma_0^2)\mu^2-2[\mu_0\sigma^2+(\sum_i n_i\bar{y}_i-\sum_i n_i a_i)\,\sigma_0^2]\mu+\mu_0^2\sigma^2}{2\,\sigma_0^2\sigma^2}\right] \end{aligned} \tag{1.31}$$

The above formula implies that $\mu \mid \cdots \sim N(\frac{\mu_0\sigma^2+(\sum_i n_i\bar{y}_i-\sum_i n_i a_i)\,\sigma_0^2}{\sigma^2+\sum_i n_i\,\sigma_0^2}, \frac{\sigma_0^2\sigma^2}{\sigma^2+\sum_i n_i\,\sigma_0^2})$.

Posterior for τ^2

$$\begin{aligned} p(\tau^2 \mid \cdots) &\propto (\tau^2)^{-a_1-1}\exp\left[-\frac{b_1}{\tau^2}\right](\tau^2)^{-\frac{m}{2}}\exp\left[-\frac{\sum_i a_i^2}{2\tau^2}\right] \\ &\propto (\tau^2)^{-(a_1+\frac{m}{2})-1}\exp\left[-\frac{2b_1+\sum_i a_i^2}{2\tau^2}\right] \end{aligned} \tag{1.32}$$

The above formula implies that $\tau^2 \mid \cdots \sim IG(a_1+\frac{m}{2}, b_1+\frac{\sum_i a_i^2}{2})$.

Posterior for σ^2

$$\begin{aligned} p(\sigma^2 \mid \cdots) &\propto (\sigma^2)^{-a_2-1}\exp\left[-\frac{b_2}{\sigma^2}\right](\sigma^2)^{-\frac{\sum_i n_i}{2}}\exp\left[-\frac{\sum_i\sum_j(y_{ij}-\mu-a_i)^2}{2\sigma^2}\right] \\ &\propto (\sigma^2)^{-(a_2+\frac{\sum_i n_i}{2})-1}\exp\left[-\frac{2b_2+\sum_i\sum_j(y_{ij}-\mu-a_i)^2}{2\sigma^2}\right] \end{aligned} \tag{1.33}$$

The above formula implies that $\sigma^2 \mid \cdots \sim IG(a_2+\frac{\sum_i n_i}{2}, b_2+\frac{\sum_i\sum_j(y_{ij}-\mu-a_i)^2}{2})$.

1.7.4 Scenario 4: Formulation (b) and Prior (b)

Keep the setup in Scenario 3 except its prior $p(\theta)$, which will now be replaced by the prior setting of Scenario 2.

It can be seen that $\mathscr{Q}'$ will stay the same as formula (1.28), and hence the colorred full conditionals for $a_i, \mu, \tau^2, \sigma^2$ remain unchanged and are given in (1.30)-(1.33). The full conditional for σ_0^2 is also needed to build the Gibbs sampler, and it is exactly the same as (1.27) in Scenario 2.

1.7.5 Additional Note

Giving the prior setting (b), we assume that the priors of μ, τ^2 and σ^2 are mutually independent, and hence their corresponding posteriors are not conditionally conjugate. The posterior of μ and τ^2 are mutually dependent under formulation (a), while the posterior of μ and σ^2 are mutually dependent under formulation (b). Since we usually have much more information to estimate σ^2 than τ^2 in the data, formulation (b) may perform better.

A regime is to set $\sigma_0^2 = \tau^2\omega$ under formulation (a) or $\sigma_0^2 = \sigma^2\omega$ under formulation (b), for some large ω reflecting the non-informativeness. Thus, $p(\mu \mid \cdots)$ becomes

$$N(\frac{\mu_0 + \omega\sum_i \mu_i}{1+m\omega}, \frac{\omega\tau^2}{1+m\omega})$$

in (1.24), and

$$N(\frac{\mu_0 + (\sum_i n_i\bar{y}_i - \sum_i n_i a_i)\omega}{1+\omega\sum_i n_i}, \frac{\omega\sigma^2}{1+\omega\sum_i n_i})$$

in (1.31).