APPROXIMATION OF THE META-ANALYTIC-PREDICTIVE PRIOR DISTRIBUTION IN THE ONE-WAY RANDOM EFFECTS MODEL WITH UNKNOWN VARIANCE

Harunori Mori*

In order to use historical data in the design of sample surveys with a Bayesian approach, the information from the historical data must be expressed as a prior distribution. Then, the best prior distribution for the parameter of interest is a predictive distribution. The density function of the predictive distribution generally is not available in an analytical form. From the perspective of practical use, Schmidli et al. (2014) proposed an approximation for the predictive distribution using a mixture of conjugate prior distributions. Their method relies on random numbers drawn from the predictive distribution. However, if the population distribution includes a nuisance parameter, their method becomes impractical. We propose a new approximation method that does not rely on these simulated numbers. Our approximation instead minimizes the mean squared error between the exact Bayes estimator and the one corresponding to the approximated predictive distribution.

Key words and phrases: Bayes estimation, meta-analytic-predictive prior, one-way random effects model.

1. Introduction

The past two decades have seen an increasing number of studies using Bayesian methods in clinical trials. Many of these studies are concerned with the inference of unknown treatment effects by using the results from related studies. If historical data are used appropriately in the design of clinical trials, we will be able to reduce the costs, trial durations, and sample sizes of studies. Especially, a reduction of sample size is ethically desirable. From a statistical perspective, the inference of unknown effects by utilizing historical data is more precise than inference using only the current data. Thus the use of historical data is attractive for clinical trials.

We assume that the results from similar studies are available. These studies are indexed by \( c = 1, \ldots, C \). We denote observations from study \( c \) by \( y_{cn} \), where \( n = 1, \ldots, N_c \). We assume that the observations \( y_{c1}, \ldots, y_{cN_c} \) are drawn from some population distribution \( F \) which is defined by a single parameter \( \alpha_c \) for each \( c \). Then, \( C \) independent samples are defined by

\[
Y_{cn} \sim F(\alpha_c), \quad n = 1, \ldots, N_c, \quad c = 1, \ldots, C.
\]

Received February 6, 2017. Revised September 17, 2017. Accepted September 26, 2017.

*Faculty of Urban Liberal Arts, Tokyo Metropolitan University, Minami-osawa 1-1, Hachioji-shi, Tokyo 192-0397, Japan.
An inference of $\alpha_1, \ldots, \alpha_C$ is not of interest here. When we refer to multiple variables $x_i, i = 1, \ldots, n$, such as for the parameters $\alpha_1, \ldots, \alpha_C$, we will use the abbreviation $\{x_i\}$.

Now we denote an independent sample from the current study by

$$X_m \sim F(\alpha), \quad m = 1, \ldots, M.$$ 

We are interested in estimating the parameter $\alpha$.

We assume that $\alpha$ and $\{\alpha_c\}$ are exchangeable (Spiegelhalter et al. (2004)) following Schmidli et al. (2014).

Then, they are distributed according to a common prior distribution, which is denoted by

$$\alpha \sim G(\beta) \quad \text{and} \quad \alpha_c \sim G(\beta), \quad c = 1, \ldots, C,$$

where $\beta$ is a hyperparameter of $G$. We can apply the empirical Bayesian method or the shrinkage estimation method to estimate $\alpha$ by using all of the available data, namely $\{y_{cn}\}$ and $\{x_m\}$. Here, we consider this estimation problem with a hierarchical Bayesian model by assuming that the hyperparameter $\beta$ also has a prior distribution. Then, $\alpha$ can be estimated with one of two approaches (Neuenschwander (2011)). In the first approach, $\alpha$ is estimated simultaneously with $\{\alpha_c\}$ by using the Bayesian method, which is referred to as the meta-analytic-combined (MAC) approach. The second approach consists of two steps. First, we derive a prior distribution of $\alpha$ from historical data, i.e., data from previous studies. Then $\alpha$ is estimated from the current study’s data and the prior distribution using the Bayesian method. This approach is referred to as the meta-analytic-predictive (MAP) approach.

In this estimation problem represented by the hierarchical Bayesian model, the MAC approach is valid from a statistical perspective. In the MAP approach, we cannot correctly estimate $\alpha$ if an incorrect prior distribution is specified in the first step. However, the MAC approach cannot be used for planning studies since the analysis can only be carried out after obtaining observations from the current study. Thus, the MAP approach is more practical in clinical trials. As demonstrated in Subsection 2.2, the MAC approach is equivalent to the MAP approach when the predictive distribution of $\alpha$ given $\{y_{cn}\}$ is set as the prior distribution of $\alpha$ (Neuenschwander (2011)). The density of the predictive distribution is defined by

$$\pi(\alpha | \{y_{cn}\}) = \int g(\alpha | \beta)\pi(\beta | \{y_{cn}\})d\beta,$$

where $g(\alpha | \beta)$ is the prior density of $\alpha$ and $\pi(\beta | \{y_{cn}\})$ is the posterior density of $\beta$ given $\{y_{cn}\}$. The predictive distribution of $\alpha$ given $\{y_{cn}\}$ is referred to as the MAP prior distribution. Although the predictive density is not available in analytical form in many cases, $\pi(\alpha | \{y_{cn}\})$ can be easily approximated using Monte Carlo integration. For sufficiently large $S$, let $\beta_s, s = 1, \ldots, S$ be a sequence of random numbers drawn from the posterior distribution, then a
Monte Carlo estimate of \( \pi(\alpha \mid \{y_{cn}\}) \) is given by \( S^{-1} \sum_{s=1}^{S} g(\alpha \mid \beta_s) \). Simulating random numbers from the posterior distribution using the Markov chain Monte Carlo (MCMC) method is straightforward. However, since the number of simulated random numbers \( S \) generally ranges from tens of thousands to hundreds of thousands, Bayesian inference of \( \alpha \) using the Monte Carlo estimate as the prior distribution of \( \alpha \) is computationally expensive. From the point of view of practical application, Schmidli et al. (2014) proposed approximating the MAP prior distribution by a mixture of conjugate prior distributions. Since the number of components is much smaller than \( S \), the computational cost of Bayesian inference is greatly decreased. They evaluated the quality of the approximation in terms of the Kullback-Leibler divergence. The weights and hyperparameters of the mixture distributions can be obtained as maximum likelihood estimates by treating the random numbers drawn from the Monte Carlo estimate \( S^{-1} \sum_{s=1}^{S} g(\alpha \mid \beta_s) \) as observations.

As discussed in Subsection 2.2, it is not practical to obtain the approximated MAP prior distribution by applying the method of Schmidli et al. (2014) when, as with a normal distribution with unknown variance, the population distribution includes a nuisance parameter. As they themselves pointed out, they do not adequately consider the problem. In this article, we approximate the MAP prior distribution directly with a mixture of conjugate prior distributions without relying on random numbers drawn from the MAP prior distribution. The approximation is achieved by minimizing the mean squared error (MSE) between the exact Bayes estimator and the one corresponding to an approximated MAP prior distribution. In this article, we consider the one-way random effects model, and assume the error terms follow a normal distribution with unknown variance.

The remainder of this article is organized as follows. In the following section, we derive the best prior distribution with the MAP approach. In Section 3, we propose a new method to approximate the MAP prior distribution with a mixture distribution. In Section 4, we present the results from a simulation study, and in Section 5 we discuss our findings.

2. The best prior distribution in the MAP approach

2.1. The model

In this article, we consider the one-way random effect model given by

\[
Y_{cn} \sim N(\alpha_c, \sigma^2), \quad n = 1, \ldots, N_c, \quad c = 1, \ldots, C,
\]

where \( \{Y_{cn}\} \) are samples corresponding to historical data. In the current study, we denote a sample by

\[
X_m \sim N(\alpha, \sigma^2), \quad m = 1, \ldots, M.
\]

Thus, we suppose that the variance is the same in all studies. When the variance is different for each study, the predictive distribution of \((\alpha, \sigma^2)\) can be easily approximated by the method of Schmidli et al. (2014), which is outlined in Appendix A.1.
We suppose that random effects $\alpha$ and $\{\alpha_c\}$ are independently distributed according to
\begin{equation}
(2.3) \quad \alpha \sim N(\beta, \sigma^2/\rho) \quad \text{and} \quad \alpha_c \sim N(\beta, \sigma^2/\rho), \quad c = 1, \ldots, C,
\end{equation}
where hyperparameters $\beta$ and $\rho$ are unknown. The sufficient statistics for $\alpha$ and $\sigma^2$ with respect to samples (2.2) are the sample mean $\bar{X}$ and the sum of the squared deviation $V$, which are defined by
\begin{align*}
\bar{X} &= \frac{1}{M} \sum_{m=1}^{M} X_m \quad \text{and} \quad V = \sum_{m=1}^{M} (X_m - \bar{X})^2.
\end{align*}
Then $\bar{X}$ and $V$ are independently distributed according to
\begin{equation}
(2.4) \quad \bar{X} | \alpha, \sigma^2 \sim N(\alpha, \sigma^2/M) \quad \text{and} \quad V | \sigma^2 \sim Ga((M - 1)/2, 2\sigma^2),
\end{equation}
respectively, where $Ga((M - 1)/2, 2\sigma^2)$ denotes a gamma distribution with shape parameter $(M - 1)/2$ and scale parameter $2\sigma^2$. This means that $V/\sigma^2$ is distributed according to a chi-squared distribution with $M - 1$ degrees of freedom.

For $\sigma^2$, we assume that $\sigma^{-2} \sim Ga(\nu_\sigma/2, 2/\tau_\sigma^2)$, where $\nu_\sigma$ and $\tau_\sigma^2$ are known. This means that
\begin{equation}
(2.5) \quad \sigma^2 \sim Ga^{-1}(\nu_\sigma/2, 2/\tau_\sigma^2),
\end{equation}
which is the inverse-gamma distribution. We note that $E[\sigma^2] = \tau_\sigma^2/(\nu_\sigma - 2)$ and $V[\sigma^2] = 2\tau_\sigma^4/\{(\nu_\sigma - 2)^2(\nu_\sigma - 4)\}$.

2.2. The best prior distribution in the MAP approach

First we consider the case where $\sigma^2$ is known. Since $\beta$ and $\rho$ are unknown, we must assume suitable prior distributions for $\beta$ and $\rho$. The joint density function is denoted by $\pi(\beta, \rho)$. Throughout this article, we will denote the density function of the sample by $p$ and that for the parameters by $\pi$. The predictive density of $\alpha$ is defined by
\begin{equation}
(2.6) \quad \pi(\alpha | \{y_{cn}\}) = \int \pi(\alpha | \beta, \rho)\pi(\beta, \rho | \{y_{cn}\})d(\beta, \rho).
\end{equation}
In the MAC approach, $\alpha$ is estimated using a posterior distribution of $\alpha$ given samples (2.1) and (2.2). The posterior density is defined by
\begin{equation*}
\pi(\alpha | \bar{x}, \{y_{cn}\}) = \int \pi(\alpha, \{\alpha_c\}, \beta, \rho | \bar{x}, \{y_{cn}\})d(\{\alpha_c\}, \beta, \rho).
\end{equation*}
Using Bayes theorem, we have
\begin{align*}
\pi(\alpha | \bar{x}, \{y_{cn}\}) &\propto \int p(\bar{x} | \alpha)\pi(\alpha | \beta, \rho)p(\{y_{cn}\} | \{\alpha_c\})\pi(\{\alpha_c\} | \beta, \rho) \\
&\quad \times \pi(\beta, \rho)d(\{\alpha_c\}, \beta, \rho) \\
&\propto p(\bar{x} | \alpha) \int \pi(\alpha | \beta, \rho) \int \pi(\{\alpha_c\} | \beta, \rho, \{y_{cn}\})d(\{\alpha_c\})d(\beta, \rho) \\
&= p(\bar{x} | \alpha) \int \pi(\alpha | \beta, \rho)\pi(\beta, \rho | \{y_{cn}\})d(\beta, \rho) \\
&= p(\bar{x} | \alpha)\pi(\alpha | \{y_{cn}\}),
\end{align*}
which shows that the MAC approach is equivalent to the MAP approach using (2.6) as the prior distribution of \( \alpha \). Since the Bayes estimator of \( \alpha \) in the MAC approach is defined as the estimator that minimizes the Bayes risk in this hierarchical Bayesian model, the Bayes risk of the Bayes estimator of \( \alpha \) in the MAC approach is always less than or equal to the Bayes risk in the MAP approach. Thus, we see that (2.6) is the best prior distribution for \( \alpha \).

For the case where \( \sigma^2 \) is unknown, a similar result can be obtained. Similar to the case where \( \sigma^2 \) is known, we have

\[
\pi(\alpha, \sigma^2 | \bar{x}, v, \{y_{cn}\}) \propto p(\bar{x}, v | \alpha, \sigma^2) \int \pi(\alpha | \sigma^2, \beta, \rho) \int \pi(\sigma^2, \beta, \rho | \{y_{cn}\}) \int \pi(\alpha | \sigma^2, \beta, \rho) \int \pi(\sigma^2, \beta, \rho | \{y_{cn}\}) d(\beta, \rho).
\]

Let

(2.7) \[ \pi(\alpha, \sigma^2 | \{y_{cn}\}) = \int \pi(\alpha | \sigma^2, \beta, \rho) \pi(\sigma^2, \beta, \rho | \{y_{cn}\}) d(\beta, \rho). \]

Then we have

\[
\pi(\alpha, \sigma^2 | \bar{x}, v, \{y_{cn}\}) \propto p(\bar{x}, v | \alpha, \sigma^2) \pi(\alpha, \sigma^2 | \{y_{cn}\}),
\]

which shows that the MAP approach is equivalent to the MAC approach. Thus, (2.7) is the MAP prior density of \((\alpha, \sigma^2)\). The predictive density of \( \alpha \) is defined by \( \pi(\alpha | \{y_{cn}\}) = \int \pi(\alpha | \sigma^2, \beta, \rho) \pi(\sigma^2, \beta, \rho | \{y_{cn}\}) d(\beta, \rho). \) We note that, however, \( \pi(\sigma^2 | \{y_{cn}\}) = \int \pi(\alpha, \sigma^2 | \{y_{cn}\}) d\alpha \) is not the predictive density of \( \sigma^2 \).

The natural conjugate prior distribution of \( \alpha \) given \( \sigma^2 \) in (2.2) is the normal distribution and that for \( \sigma^2 \) is the inverse-gamma distribution. Therefore, we approximate (2.7) with the following mixture of these densities with positive weights \( \omega_k \), \( k = 1, \ldots, K \) summing to one:

(2.8) \[ \hat{\pi}_K(\alpha, \sigma^2 | \{y_{cn}\}) = \sum_{k=1}^{K} \omega_k n(\alpha | b_k, \sigma^2 / r_k) \gamma^{-1}(\sigma^2 | \nu_k / 2, 2 / \tau_k^2), \]

where \( n(\alpha) \) and \( \gamma^{-1}(\sigma^2) \) are the density functions of the normal distribution and the inverse-gamma distribution, respectively. This approximation is justified by Theorem 3.1 of Dalal and Hall (1983).

The joint density (7) can be expressed as

(2.9) \[ \pi(\alpha, \sigma^2 | \{y_{cn}\}) = \pi(\alpha | \sigma^2, \{y_{cn}\}) \pi(\sigma^2 | \{y_{cn}\}). \]

Although it is possible to draw random numbers from (2.7) using equation (2.9), this is not practical for the following reasons. Let \( s^2 \) denote a random number.
drawn from the posterior distribution of $\sigma^2$ given $\{y_{cn}\}$. Then we can draw random numbers from $\pi(\beta, \rho \mid s^2, \{y_{cn}\})$, which are denoted by $(\beta_s, \rho_s), s = 1, \ldots, S$. The density $\pi(\alpha \mid s^2, \{y_{cn}\})$ can be approximated using Monte Carlo integration, which is given by $S^{-1} \sum_{s=1}^{S} n(\alpha \mid \beta_s, s^2/\rho_s)$. Therefore, in order to draw random numbers from (2.7), we must draw a random number from the above Monte Carlo estimate for each $s^2$. Since it is necessary to generate so many random numbers corresponding to $(\alpha, \sigma^2)$, this method is extremely inefficient.

3. Proposed method

3.1. Standard results for Bayes estimation

In this subsection, we summarize standard results about Bayes estimation, where the sample is as in (2.2) and the prior distributions are (2.3) and (2.5). These results are necessary in the following subsection. Here, we refer to Anderson (2003) for the derivation of these results.

From (2.3) and (2.5), the prior density of $\alpha$ is

$$
\pi(\alpha) = \int_0^\infty \pi(\alpha \mid \sigma^2)\pi(\sigma^2)d\sigma^2
$$

$$
= \sqrt{\frac{\rho}{\pi \tau^2}} \frac{\Gamma \left( \frac{\nu + 1}{2} \right)}{\Gamma \left( \frac{\nu}{2} \right)} \left\{ 1 + \frac{\rho}{\tau^2} (\alpha - \beta)^2 \right\}^{-(\nu + 1)/2}.
$$

(3.1)

Let $T^2 = \frac{\nu \sigma}{\tau^2} (\alpha - \beta)^2$; then, $T$ is distributed according to a $t$ distribution with $\nu$ degrees of freedom.

The posterior distribution of $\alpha$ given $\sigma^2$ and $\bar{x}$ is

$$
\alpha \mid \sigma^2, \bar{x} \sim N(\bar{\alpha}, \sigma^2/(M + \rho)),
$$

(3.2)

where

$$
\bar{\alpha} = \bar{x} - \eta(\bar{x} - \beta) \quad \text{and} \quad \eta = \frac{\rho}{M + \rho}.
$$

Since the posterior mean $\bar{\alpha}$ does not include $\sigma^2$, $\bar{\alpha}$ is the Bayes estimator of $\alpha$. The posterior distribution of $\sigma^2$ given $\bar{x}$ and $v$ is

$$
\sigma^2 \mid \bar{x}, v \sim Ga^{-1}((M + \nu)/2, 2/\tilde{\tau}^2),
$$

(3.3)

where

$$
\tilde{\tau}^2 = M\eta(\bar{x} - \beta)^2 + v + \tau^2.
$$

From (3.2) and (3.3), the posterior density of $\alpha$ given $\bar{x}$ and $v$ is

$$
\pi(\alpha \mid \bar{x}, v) = \int_0^\infty \pi(\alpha \mid \sigma^2, \bar{x}, v)\pi(\sigma^2 \mid \bar{x}, v)d\sigma^2
$$

$$
\propto \left\{ 1 + \frac{M + \rho}{\tilde{\tau}^2} (\alpha - \bar{\alpha})^2 \right\}^{-(M + \nu + 1)/2}.
$$

(3.4)
Let \( T^2 = \frac{(M+\rho)(M+\nu_\sigma)}{\tilde{\tau}^2} (\alpha - \tilde{\alpha})^2 \); then, \( T \) is distributed according to a \( t \) distribution with \( M + \nu_\sigma \) degrees of freedom. From the properties of the \( t \) distribution, we can see that

\[
E[\alpha \mid \bar{x}, v] = \tilde{\alpha} \quad \text{and} \quad V[\alpha \mid \bar{x}, v] = \frac{\tilde{\tau}^2}{(M + \rho)(M + \nu_\sigma - 2)}
\]

From a simple calculation, we have

\[
\bar{X} \mid \sigma^2 \sim N(\beta, (M^{-1} + \rho^{-1})\sigma^2).
\]

Therefore, the marginal density of \((\bar{X}, V)\) is

\[
p(\bar{x}, v) = \int_0^\infty p(\bar{x} \mid \sigma^2)p(v \mid \sigma^2)\pi(\sigma^2) d\sigma^2
\]

\[
= \frac{\sqrt{M\eta}(\tilde{\tau}_v^2)^\nu_\sigma/2\Gamma\left(\frac{M + \nu_\sigma}{2}\right)}{\sqrt{\pi\Gamma\left(\frac{M - 1}{2}\right)\Gamma\left(\frac{\nu_\sigma}{2}\right)}} v^{(M-3)/2}(\tilde{\tau}_v^2)^{-(M+\nu_\sigma)/2}.
\]

### 3.2. Proposed method

Denote the Bayes estimator of \( \alpha \) in the MAC approach by \( \hat{\alpha}_0 \), which is given by

\[
\hat{\alpha}_0 = E[\alpha \mid \bar{X}, V, \{Y_{cn}\}] = \int \alpha \pi(\alpha, \sigma^2 \mid \bar{X}, V, \{Y_{cn}\}) d(\alpha, \sigma^2).
\]

As mentioned in Subsection 2.2, \( \hat{\alpha}_0 \) is also the Bayes estimator with respect to the MAP prior density (2.7) in the MAP approach.

From (2.8), the prior distribution of \((\alpha, \sigma^2)\) used in the second step of the MAP approach are

\[
(\alpha, \sigma^2) \sim \sum_{k=1}^K \omega_k N(\alpha \mid b_k, \sigma^2/r_k)Ga^{-1}(\sigma^2 \mid \nu_k/2, 2/\tau_k^2).
\]

The Bayes estimator of \( \alpha \) with respect to the prior distributions (3.7) is denoted by \( \hat{\alpha}_K \), which is derived in Subsection 3.3. Since the hyperparameter \( r_k \) is used in the definition of \( h_k \), namely \( h_k = r_k/(M+r_k) \), in \( \hat{\alpha}_K \), we use \( h_k \) instead of \( r_k \) in the explanation below. The Bayes estimator \( \hat{\alpha}_K \) includes the unknown mixture weights \( \omega_k \) and the unknown hyperparameters \( \{b_k, h_k, \nu_k, \tau_k^2\} \). To simplify the notation, let \( \theta_k = (\omega_k, b_k, h_k, \nu_k, \tau_k^2) \).

We define these parameters as the values that minimize the MSE between \( \hat{\alpha}_0 \) and \( \hat{\alpha}_K \). The MSE is given by

\[
MSE(\hat{\alpha}_0, \hat{\alpha}_K \mid \{\theta_k\}) = \int (\hat{\alpha}_0 - \hat{\alpha}_K)^2 \pi(\bar{x}, v, \{y_{cn}\}) d(\bar{x}, v, \{y_{cn}\}),
\]

where \( \pi(\bar{x}, v, \{y_{cn}\}) \) is the marginal density of \((\bar{X}, V, \{Y_{cn}\})\).
From Theorem 3.1 of Dalal and Hall (1983), there exists a sequence of densities \( \{ \hat{\pi}_K(\alpha, \sigma^2 | \{ y_{cn} \}) \} \) converging to the MAP prior density (2.7), namely \( \pi(\alpha, \sigma^2 | \{ y_{cn} \}) \), as \( K \to \infty \) for all \( \alpha \) and \( \sigma^2 \). In addition, the posterior density of \( (\alpha, \sigma^2) \) with respect to \( \hat{\pi}_K \) also converges to the posterior density with respect to (2.7). If the prior distribution of \( (\beta, \rho) \) is non-informative, the posterior distribution of \( (\alpha, \sigma^2) \) with respect to (2.7) may also be non-informative when \( C, \{ N_c \} \), and \( M \) are small. In this case, the Bayes estimator \( \hat{\alpha}_0 \) does not exist and the MSE does not converge to 0 as \( K \to \infty \). However, we can assume that the second-order moment of \( \hat{\alpha}_0 \) with respect to the marginal distribution of \((\bar{X}, \bar{V}, \{ \bar{Y}_{cn} \})\) exists in most practical situations. Then, the MSE converges to 0 as \( K \to \infty \) under this assumption. We can also show that, from Theorem 25.12 and its Corollary in Billingsley (1995), the convergence of the MSE holds under some mild assumptions about \( \hat{\pi}_K \).

The MAP prior density (2.7) does not properly include the sample size \( M \) of (2.2). However, since both \( \hat{\alpha}_0 \) and \( \hat{\alpha}_K \) depend on \( M \), the values of \( \{ \theta_k \} \) that minimize the MSE may also depend on \( M \) for small \( K \).

Since \( \hat{\alpha}_0 \) is not available in analytical form, it is not practical to minimize directly the MSE with respect to \( \{ \theta_k \} \). As shown below, minimizing the MSE is equivalent to minimizing the Bayes risk of \( \hat{\alpha}_K \) under the squared loss. Values of \( \{ \theta_k \} \) that minimize the Bayes risk can be more easily obtained.

The Bayes risk of \( \hat{\alpha}_K \) under the squared loss is defined by

\[
BR(\hat{\alpha}_K | \{ \theta_k \}) = \int (\alpha - \hat{\alpha}_K)^2 p(\bar{x}, \bar{v} | \alpha, \sigma^2)p(\{ y_{cn} \} | \{ \alpha_c \}, \sigma^2) \times \pi(\alpha | \sigma^2, \beta, \rho)\pi(\{ \alpha_c \} | \sigma^2, \beta, \rho)\pi(\sigma^2, \beta, \rho) \times d(\bar{x}, \bar{v}, \{ y_{cn} \}, \alpha, \{ \alpha_c \}, \sigma^2, \beta, \rho)
\]

\[
= \int \int (\alpha - \hat{\alpha}_K)^2 p(\bar{x}, \bar{v} | \alpha, \sigma^2)\pi(\alpha | \sigma^2, \beta, \rho) \times \pi(\sigma^2, \beta, \rho | \{ y_{cn} \})d(\bar{x}, \bar{v}, \alpha, \sigma^2, \beta, \rho)\pi(\{ y_{cn} \})d(\{ y_{cn} \})
\]

\[
= \int \int (\alpha - \hat{\alpha}_K)^2 p(\bar{x}, \bar{v} | \alpha, \sigma^2)\pi(\alpha, \sigma^2 | \{ y_{cn} \})d(\bar{x}, \bar{v}, \alpha, \sigma^2)
\]

\[
\times \pi(\{ y_{cn} \})d(\{ y_{cn} \}).
\]

In the third equality, we use (2.7), namely the definition of \( \pi(\alpha, \sigma^2 | \{ y_{cn} \}) \). Substituting \( \alpha - \hat{\alpha}_K = (\alpha - \hat{\alpha}_0) + (\hat{\alpha}_0 - \hat{\alpha}_K) \), from the definition of \( \hat{\alpha}_0 \), we have

\[
BR(\hat{\alpha}_K | \{ \theta_k \}) = \int \int (\alpha - \hat{\alpha})^2 p(\bar{x}, \bar{v} | \alpha, \sigma^2)\pi(\alpha, \sigma^2 | \{ y_{cn} \})d(\bar{x}, \bar{v}, \alpha, \sigma^2)
\]

\[
\times \pi(\{ y_{cn} \})d(\{ y_{cn} \}) + MSE(\hat{\alpha}_0, \hat{\alpha}_K | \{ \theta_k \}).
\]
Since the first term is constant with respect to \( \{ \theta_k \} \), we see that minimizing the MSE is equivalent to minimizing the Bayes risk.

The Bayes risk can be written as

\[
BR(\hat{\alpha}_K \mid \{ \theta_k \}) = \int L_X(\hat{\alpha}_K \mid \{ \theta_k \}) \pi(\alpha \mid \sigma^2, \beta, \rho) \, d(\bar{x}, v, \alpha)
\]

(3.8)

where

\[
L_X(\hat{\alpha}_K \mid \{ \theta_k \}) = \int (\alpha - \hat{\alpha}_K)^2 p(\bar{x}, v \mid \alpha, \sigma^2) \pi(\alpha \mid \sigma^2, \beta, \rho) \, d(\bar{x}, v, \alpha).
\]

(3.9)

### 3.3. The Bayes estimator and the Bayes risk

Since (3.7) is a mixture of conjugate prior distributions, the posterior distribution is also a mixture of posterior distributions with updated mixture weights. The Bayes estimator \( \hat{\alpha}_K \) is defined by

\[
\hat{\alpha}_K = E[\alpha \mid \bar{X}, V] = \sum_{k=1}^{K} p_k \hat{\alpha}_k,
\]

(3.10)

where, from (3.2), the posterior mean \( \hat{\alpha}_k \) corresponding to the \( k \)th component is given by

\[
\hat{\alpha}_k = \bar{X} - h_k(\bar{X} - b_k) \quad \text{and} \quad h_k = \frac{r_k}{M + r_k}, \quad k = 1, \ldots, K.
\]

From the marginal density (3.6), the updated mixture weight \( p_k \) is given by

\[
p_k = \frac{\omega_k q_k}{\sum_{k'=1}^{K} \omega_{k'} q_{k'}}
\]

where

\[
q_k = \frac{\sqrt{r_k} \tau_k^{\nu_k/2}/2 \Gamma \left( \frac{M + \nu_k}{2} \right)}{\Gamma \left( \frac{\nu_k}{2} \right)} (\bar{x}_k^{(2)} - (M + \nu_k)/2)
\]
and
\[ \hat{\tau}_k^2 = Mh_k(\bar{X} - b_k)^2 + V + \tau_k^2. \]
Thus, the Bayes estimator \( \hat{\alpha}_K \) is written as \( \hat{\alpha}_K = \bar{X} - G(\bar{X}, V) \), where
\[ G(\bar{X}, V) = \sum_{k=1}^{K} h_k p_k(\bar{X} - b_k). \]

Next, we derive the loss function (3.8). From (3.2), we have
\[
E[\alpha | \bar{X}, V, \sigma^2, \beta, \rho] = \bar{X} - \eta(\bar{X} - \beta),
\]
\[
E[\alpha^2 | \bar{X}, V, \sigma^2, \beta, \rho] = \{\bar{X} - \eta(\bar{X} - \beta)\}^2 + \frac{1}{M + \rho} \sigma^2.
\]
Therefore, the expectation of \((\alpha - \hat{\alpha}_K)^2\) with respect to \( \alpha \) is given by
\[
E[(\alpha - \hat{\alpha}_K)^2 | \bar{X}, V, \sigma^2, \beta, \rho] = \frac{1}{M + \rho} \sigma^2 + \eta^2(\bar{X} - \beta)^2 - 2\eta(\bar{X} - \beta)G(\bar{X}, V) + G^2(\bar{X}, V).
\]

By taking the expectation of the above conditional expectation with respect to \( \bar{X} \) and \( V \), (3.8) can be derived from (3.5) and (2.4),
\[
L_X(\hat{\alpha}_K | \{\theta_k\}) = E[(\alpha - \hat{\alpha}_K)^2 | \sigma^2, \beta, \rho]
= \frac{1}{M} \sigma^2 - 2E[L_1 | \sigma^2, \beta, \rho] + E[L_2 | \sigma^2, \beta, \rho],
\]
(3.11)

where
\[
L_1(\{\theta_k\} | \bar{X}, V, \beta, \rho) = \eta(\bar{X} - \beta)G(\bar{X}, V),
\]
\[
L_2(\{\theta_k\} | \bar{X}, V, \beta, \rho) = G^2(\bar{X}, V).
\]

We note that, although the two functions \( L_1 \) and \( L_2 \) do not include \( \sigma^2 \), the expectations include \( \sigma^2 \).

The posterior risk (3.9) can be derived by taking the expectation of (3.11) with respect to \( \sigma^2, \beta, \) and \( \rho \) given \( \{y_{cn}\} \),
\[
E[L_X(\hat{\alpha}_K | \{\theta_k\}) | \{y_{cn}\}] = \frac{1}{M} E[\sigma^2 | \{y_{cn}\}] - 2E[E[L_1 | \sigma^2, \beta, \rho] | \{y_{cn}\}] + E[E[L_2 | \sigma^2, \beta, \rho] | \{y_{cn}\}]\]
(3.12)

However, since the expectations of \( L_1 \) and \( L_2 \) are not available in analytical form, we must use Monte Carlo integration in order to approximate these expectations.

The joint density of \( \bar{x}, v, \sigma^2, \beta, \) and \( \rho \) is expressed as
\[
\pi(\bar{x}, v, \sigma^2, \beta, \rho | \{y_{cn}\}) = p(\bar{x}, v | \sigma^2, \beta, \rho)\pi(\sigma^2, \beta, \rho | \{y_{cn}\})
= p(\bar{x} | \sigma^2, \beta, \rho)p(v | \sigma^2)\pi(\sigma^2, \beta, \rho | \{y_{cn}\}).
\]
For sufficiently large $S$, let $(\bar{x}_s, v_s, \sigma^2_s, \beta_s, \rho_s), \ s = 1, \ldots, S$ be a sequence of random numbers drawn from the joint posterior distribution. Then a Monte Carlo estimate of (3.12) is given by

$$
\frac{1}{MS} \sum_{s=1}^{S} \sigma^2_s - \frac{2}{S} \sum_{s=1}^{S} L_1(\{\theta_k\} | \bar{x}_s, v_s, \beta_s, \rho_s)
$$

$$
+ \frac{1}{S} \sum_{s=1}^{S} L_2(\{\theta_k\} | \bar{x}_s, v_s, \beta_s, \rho_s).
$$

(3.13)

In order to minimize (3.13) with respect to $\{\theta_k\}$, we use a numerical analysis. In the numerical analysis, partial derivatives of (3.13) with respect to these variables are required. We summarize these partial derivatives in Appendix A.2.

### 3.4. The case where $K = 1$

When the number of components $K$ in (2.7) is one, the numerical analysis to minimize (3.13) is not necessary. To simplify the notation, we omit the subscript $k$. When $K = 1$, the prior distribution (3.7) is

$$(\alpha, \sigma^2) \sim N(b, \sigma^2 / r)Ga^{-1}(\nu/2, \tau^2/2)$$

Then the Bayes estimator of $\alpha$ is given by $\hat{\alpha}_1 = \bar{X} - h(\bar{X} - b)$, where $h = \frac{r}{M + r}$. The loss function $L_X(\hat{\alpha}_1 | b, h)$ can be derived analytically. By taking the expectation of $(\alpha - \hat{\alpha}_1)^2$ first with respect to $\bar{X}$ given $\alpha$ and next with respect to $\alpha$, we have

$$L_X(\hat{\alpha}_1 | b, h) = \left\{ \frac{1}{M^2} \sigma^2 + (\beta - b)^2 \right\} h^2 - \frac{2}{M} \sigma^2 h + \frac{1}{M} \sigma^2.$$

The posterior risk is given by

$$E[L_X(\hat{\alpha}_1 | b, h) | \{y_{cn}\}] = \left\{ \frac{1}{M} E \left[ \frac{1}{\eta} \sigma^2 \mid \{y_{cn}\} \right] + E[(\beta - b)^2 \mid \{y_{cn}\}] \right\} h^2
$$

$$- \frac{2}{M} E[\sigma^2 \mid \{y_{cn}\}] h + \frac{1}{M} E[\sigma^2 \mid \{y_{cn}\}].$$

(3.14)

The $b$ and $h$ that minimize the posterior risk, denoted by $b^*$ and $h^*$, are given by

$$b^* = E[\beta \mid \{y_{cn}\}]$$

and

$$h^* = \frac{E[\sigma^2 \mid \{y_{cn}\}]}{E[\eta^{-1} \sigma^2 \mid \{y_{cn}\}] + ME[\beta \mid \{y_{cn}\}]}$$

$$= \frac{E[\sigma^2 \mid \{y_{cn}\}]}{ME[\rho^{-1} \sigma^2 \mid \{y_{cn}\}] + E[\sigma^2 \mid \{y_{cn}\}] + MV[\beta \mid \{y_{cn}\}]}.$$

The $r$ corresponding to $h^*$, which is denoted by $r^*$, is given by

$$r^* = \frac{E[\sigma^2 \mid \{y_{cn}\}]}{E[\rho^{-1} \sigma^2 \mid \{y_{cn}\}] + V[\beta \mid \{y_{cn}\}]}.$$

We note that both $b^*$ and $r^*$ do not depend on the sample size $M$ of (2.2).
4. Simulations

4.1. Data and prior distributions

In the following simulation study, we apply the proposed method to artificially constructed data under two settings. In the first setting, which will be referred to as Setting 1, data are constructed based on the one-way random effect model (2.1) with \( \sigma^2 = 100 \). Random effects \( \{\alpha_c\} \) are generated from the prior distribution (2.3) with \( \beta = 0 \) and \( \rho = 1 \), namely \( N(0,10^2) \).

In the simulation, the prior distributions of \( \beta \) and \( \rho \) must be specified. We assume that \( \beta \sim N(\mu_\beta, \tau_\beta^2) \) and \( \rho \sim Ga(\nu_\rho/2, 2\tau_\rho^2) \). For the estimation, we choose values of hyperparameters in order that the prior distributions may be close to being non-informative. For the prior distribution of \( \beta \), the hyperparameters are set as \( \mu_\beta = 0 \) and \( \tau_\beta^2 = 1000 \). For the prior distribution of \( \rho \), we set \( \nu_\rho = 0.5 \) and \( \tau_\rho^2 = 2 \). The hyperparameters of (2.5) are set as \( \nu_\sigma = 4 \) and \( \tau_\sigma^2 = 200 \), which means that the variance of \( \sigma^2 \) does not exist. We note that their expectations are the same as the values of \( \sigma^2, \beta, \) and \( \rho \) defined above in order to construct the data.

In order to minimize the approximated posterior risk (3.13), we generate random numbers from the posterior distributions of \( \sigma^2, \beta, \) and \( \rho \). Since we are considering a hierarchical model defined by its conjugate prior distribution, the random numbers can be easily simulated with a Gibbs sampler (Chib (2001), Geweke (2005)). The sampling algorithm proceeds as follows.

(i) Draw \( \alpha_c \mid \ast \sim N(w_c\bar{y}_c + (1-w_c)\beta, \sigma^2/(N+\rho)) \) independently for \( c = 1, \ldots, C \). The notation \( \alpha_c \mid \ast \) indicates that all parameters other than \( \alpha_c \) are conditions. Here \( w_c = N/(N+\rho) \) and \( \bar{y}_c = N^{-1}\sum_{n=1}^{N}y_{nc} \).

(ii) Draw \( \sigma^2 \mid \ast \sim Ga^{-1}((CN + \nu_\sigma)/2, 2/(W_y + \tau_\sigma^2)) \).

Here \( W_y = \sum_{c=1}^{C} \sum_{n=1}^{N} (y_{cn} - \alpha_c)^2 \).

(iii) Draw \( \beta \mid \ast \sim N(w\bar{\alpha} + (1-w)\mu_\beta, \sigma^2\tau_\beta^2/(C\rho\tau_\beta^2 + \sigma^2)) \).

Here \( w = C\rho\tau_\beta^2/(C\rho\tau_\beta^2 + \sigma^2) \) and \( \bar{\alpha} = C^{-1}\sum_{c=1}^{C} \alpha_c \).

(iv) Draw \( \rho \mid \ast \sim Ga((C + \nu_\rho)/2, 2\sigma^2\tau_\rho^2/(\tau_\rho^2W_\alpha + \sigma^2)) \).

Here \( W_\alpha = \sum_{c=1}^{C}(\alpha_c - \beta)^2 \).

We generate 1,000,000 random numbers after a burn-in period. Convergence of the Markov chain was assessed by testing the hypothesis that the population means of the first 100,000 iterations and the last 500,000 iterations are equal. The standard error of the sample mean was estimated using the Parzen window with bandwidth of 500 (Geweke (1992)). Finally we select a subset of 100,000 random numbers that are the \((10i - 9)\)th value for \( i = 1 \) to 100,000 in order to reduce the auto-correlation of the sequence.

For simplicity, the sample size \( N_c \) of (2.1) is constant for all \( c \), and thus can be denoted by \( N \). The sample size \( N \) is set to 10 and the number of studies \( C \) is set to 10. We consider sample sizes \( M \) from (2.2), of \( M = 5, 10, \) and \( 20 \).

In the second setting, which will be referred as Setting 2, we set \( \sigma^2 = 188.0, \beta = 40.5, \) and \( \rho = 3.0 \). These values correspond to the 90th percentiles of their prior distributions. Setting 2 considers the effect on the results of using an inappropriate prior distribution for the data.
Table 1. Approximated posterior risks.

| Setting 1 | Setting 2 |
|-----------|-----------|
| $M = 5$   | $M = 5$   |
| $M = 10$  | $M = 10$  |
| $M = 20$  | $M = 20$  |

| $K = 1^*$ |          |
|-----------|----------|
| 16.0423   | 27.7418  |
| 8.5061    | 14.8758  |
| 4.3806    | 7.7042   |

| $K = 1$   |          |
|-----------|----------|
| 15.6999   | 27.0116  |
| 8.3956    | 14.6282  |
| 4.3490    | 7.6312   |

| $K = 2$   |          |
|-----------|----------|
| 15.6154   | 26.7702  |
| 8.3587    | 14.5565  |
| 4.3384    | 7.6077   |

| $K = 3$   |          |
|-----------|----------|
| 15.6144   | 26.7679  |
| 8.3583    | 14.5555  |
| 4.3383    | 7.6077   |

| $K = 4$   |          |
|-----------|----------|
| 15.6138   | 26.7667  |
| 8.3580    | 14.5549  |
| 4.3382    | 7.6075   |

| $K = 5$   |          |
|-----------|----------|
| 15.6134   | 26.7662  |
| 8.3578    | 14.5545  |
| 4.3381    | 7.6074   |

All simulation programs were written using the SAS/IML programming language. Specifically, the NLPQN subroutine that applies the quasi-Newton method was used to minimize the approximated posterior risk (3.13).

4.2. Simulation results

We applied (2.8) varying the number of components $K$ from 1 to 5 to approximate the MAP prior density (2.7). Table 1 shows the approximated posterior risks (3.13) for each sample size $M$.

When $K = 1$, we obtain the posterior risk by the method described in Subsection 3.4. For reference, we also obtain the posterior risk (3.14) with $E[\rho | \{y_{cn}\}] / (M + E[\rho | \{y_{cn}\}])$ that is a plug-in estimate of $\eta$. The posterior risks are presented in the row indexed by $K = 1^*$. For all cases, the posterior risk with the plug-in estimate is greatest. Unfortunately, appropriate criteria for determining the number of components $K$ are currently not known. Since the posterior risk decreases gradually with increasing $K$, we set $K = 5$ for all cases.

Table 2 shows estimates of $\{\omega_k, b_k, r_k, \nu_k, \tau_k^2\}$ for each case. The components are ordered by mixture weights.

As mentioned in Subsection 3.2, $\hat{\pi}_K(\alpha, \sigma^2 | \{y_{cn}\})$ obtained by our method depends on the sample size $M$ of (2.2). Looking at Table 2, $\hat{\pi}_K$ seems to be similar for each $M$. We derived the density function $\hat{\pi}_K(\alpha | \{y_{cn}\})$ for each $M$ by using the values of the hyperparameters presented in Table 2, and compared these densities by their plots. From (3.1), the density of $\alpha$ is given by

$$
\hat{\pi}_K(\alpha | \{y_{cn}\}) = \int_0^\infty \hat{\pi}_K(\alpha, \sigma^2 | \{y_{cn}\})d\sigma^2
$$

$$
= \sum_{k=1}^K \omega_k \int_0^\infty n(\alpha | b_k, \sigma^2/r_k)\gamma^{-1}(\sigma^2 | \nu_k/2, 2/\tau_k^2)d\sigma^2
$$

$$
= \sum_{k=1}^K \omega_k \sqrt{\frac{r_k}{\pi \tau_k^2}} \Gamma \left(\frac{\nu_k + 1}{2}\right) \left\{1 + \frac{r_k}{\tau_k^2}(\alpha - b_k)^2\right\}^{-(\nu_k+1)/2}.
$$

Almost no difference can be seen between these graphs (not shown). Thus, the influence of sample size $M$ on the approximation of (2.7) with $\hat{\pi}_K$ by our method seems to be small.
It is difficult to evaluate directly the accuracy of the approximation of \( \hat{\pi}_K \) obtained by our method. However, if \( \hat{\pi}_K \) approximates the MAP prior density (2.7) well, then the Bayes estimator \( \hat{\alpha}_K \) should be close to \( \hat{\alpha}_0 \), namely the Bayes estimator of \( \alpha \) in the MAC approach. We simulate additional data \( \{x_m\} \) with \( M = 10 \) for several values of \( \alpha \) to obtain \( \hat{\alpha}_0 \) and \( \hat{\alpha}_K \). Table 3(a) and (b) show the values of \( \hat{\alpha}_0 \) and \( \hat{\alpha}_K \) and their standard deviations. From (3.4), the standard deviation of \( \hat{\alpha}_K \) is given by \( \sqrt{\sum_{k=1}^{2} \omega_k V_k[\alpha \mid \bar{x}, v]} \), where \( V_k[\alpha \mid \bar{x}, v] = \hat{\tau}^2_k / \{(M + r_k)(M + \nu_k - 2)\} \).

Table 3 shows that \( \hat{\alpha}_K \) is very close to \( \hat{\alpha}_0 \) even in Setting 2 corresponding to the case where the prior distributions are inappropriate for the data. This result suggests that \( \hat{\pi}_K \) closely approximates (2.7) under our method.

### 5. Discussion

Regarding the use of historical data, we referred mainly to studies on clinical trials, which represent an important application for the statistical analysis of hist-
historical data. In clinical trials, historical data can be used not only for estimating unknown effects but also for planning the study under consideration. Although observations in clinical trials rarely follow a normal distribution, the distribution of their logarithms is often well approximated by a normal distribution (Spiegelhalter et al. (2004)). Then we can approximate the MAP prior distribution with a mixture of conjugate prior distributions by applying our method.

In other fields, historical data are often used for determining the sample size. In psychology or sensory evaluation studies, which commonly involve clinical trials with human subjects, it is desirable to determine the sample size necessary to achieve the desired power for hypothesis tests. If sufficient historical data are not available, a preliminary experiment is often carried out to obtain prior information for the parameter of interest. Then, the MAP approach can be applied for the determination of the sample size.

An example of potential applications of the MAP approach outside of clinical trials is substantive tests in auditing. Each transaction included in a set of accounts has a recorded value called the book value and a true value called the audit value. The audit value is objectively estimated based on evidence. The error in a transaction is defined as the book value minus the audit value. The auditor must investigate whether misstatements exist in the accounts based on these errors. However, since the number of transactions included in a set of accounts is usually enormous, the auditor cannot investigate the error for all transactions. Therefore, in auditing, a sample survey is carried out. If prior information such as the results of auditing from previous years is available, then the auditor can determine the effective sample size by using this information from past audits (Sahu and Smith (2006), Laws and O’Hagan (2000, 2002)). The MAP

### Table 3(a). Setting 1.

| MAC approach | MAP approach |
|--------------|--------------|
| $\hat{\alpha}_0$ | $\hat{\alpha}_K$ |
| s.d. | s.d. |
| $\alpha = -20$ | $-20.32$ | $-20.31$ |
| s.d. | 2.88 | 2.83 |
| $\alpha = -10$ | $-14.06$ | $-14.06$ |
| s.d. | 2.88 | 2.87 |
| $\alpha = 0$ | $-2.67$ | $-2.68$ |
| s.d. | 2.86 | 2.86 |
| $\alpha = 10$ | $5.09$ | $5.08$ |
| s.d. | 2.91 | 2.90 |
| $\alpha = 20$ | $19.84$ | $19.87$ |
| s.d. | 2.90 | 2.84 |

### Table 3(b). Setting 2.

| MAC approach | MAP approach |
|--------------|--------------|
| $\hat{\alpha}_0$ | $\hat{\alpha}_K$ |
| s.d. | s.d. |
| $\alpha = 24$ | $26.61$ | $26.64$ |
| s.d. | 3.84 | 3.81 |
| $\alpha = 32$ | $30.41$ | $30.39$ |
| s.d. | 3.91 | 3.89 |
| $\alpha = 40$ | $37.00$ | $37.01$ |
| s.d. | 3.70 | 3.69 |
| $\alpha = 48$ | $57.00$ | $57.00$ |
| s.d. | 3.86 | 3.82 |
| $\alpha = 56$ | $50.75$ | $50.72$ |
| s.d. | 3.69 | 3.67 |
approach would also be useful in this situation. Since the error in a transaction
is often zero, we must assume a mixture distribution with a positive probability
assigned to zero for the population distribution of errors. Compared to the one-
way random effects model considered in this article, the numerical analysis to
minimize the approximated posterior risk will likely be more complex.

When we use historical data in a meta-analysis, we must consider the prior-
data conflict problem. Despite careful selection of historical trials, past infor-
mation may not be relevant for the new trial due to unanticipated differences in
study design, conduct or patient population (Schmidli et al. (2014)). In some
cases, we can only use the average or the variance described in the literature,
not the complete dataset. Then, we often assume a normal distribution for the
population distribution of observations for convenience. If the historical data or
a statistical model is not relevant for the current study, the MAP prior distribu-
tion may lead to incorrect estimates. Therefore, in the meta-analysis, we must
consider a prior distribution that discounts the information from historical data.
For this purpose, the power prior distribution proposed by Ibrahim and Chen
(2000) and the commensurate prior distribution proposed by Hobbs et al. (2011)
can be used. Schmidli et al. (2014) proposed a simple and intuitive prior dis-
tribution that is different from the above two prior distributions. They referred
to this as the robust MAP prior distribution. It is defined as a mixture of the
approximated MAP prior distribution and a vague conjugate prior distribution.
They showed that it avoids the prior-data conflict. We note that, since we also
approximate the MAP prior distribution with a mixture distribution, we can
easily construct the robust MAP prior distribution.

In the simulation study, we used the maximum value of $K$ among all of
the values examined since an appropriate criterion to determine the optimum
number of components in our problem is not available. From the point of view
of practical applications, the approximation accuracy of $\hat{\pi}_K$ may be sufficiently high
with $K = 3$. It is an area for future work to establish an appropriate criterion to
determine the number of components.

Appendix

A.1. The case where the variance is different for each study

We denote the samples corresponding to historical data by

$$Y_{cn} \sim N(\alpha_c, \sigma_c^2), \quad n = 1, \ldots, N_c, \quad c = 1, \ldots, C.$$ 

In the current study, a sample is defined by (2.2). Schmidli et al. (2014) con-
sidered this problem in their web appendix. Similar to many studies on meta-
analysis, they treated the variances $\{\sigma_c^2\}$ as known. Moreover, they hardly con-
sider how $\sigma^2$ should be estimated. Thus, when regarding this problem as a purely
theoretical one, their treatment of this problem is not sufficiently rigorous. If $\sigma^2$
is the parameter of interest, the problem is straightforward. As mentioned in
Subsection 2.1, the sufficient statistic for the variance follows a gamma distribu-
tion, which does not include the mean, namely a nuisance parameter. Schmidli
et al. (2016) applied the method of Schmidli et al. (2014) to approximate the MAP prior distribution of $\sigma^2$.

We assume that not only random effects $\alpha$ and $\{\alpha_c\}$ but also variances $\sigma^2$ and $\{\sigma_c^2\}$ are exchangeable. We assume that $\sigma^2$ and $\{\sigma_c^2\}$ are independently distributed according to

$$
\sigma^2 \sim \text{Ga}^{-1}(\nu/2, 2/\tau^2) \quad \text{and} \quad \sigma_c^2 \sim \text{Ga}^{-1}(\nu/2, 2/\tau_c^2), \quad c = 1, \ldots, C,
$$

where $\nu$ and $\tau^2$ are unknown. The distributions of $\alpha$ and $\{\alpha_c\}$ are

$$
\alpha \sim N(\beta, \psi^2) \quad \text{and} \quad \alpha_c \sim N(\beta, \psi^2), \quad c = 1, \ldots, C.
$$

We note that the parameter $\rho$ in (2.3) is not used here.

From the discussion in Subsection 2.2, we see that the best prior distribution in the MAP approach is the predictive distribution of $(\alpha, \sigma^2)$. The joint density is given by

$$
\pi(\alpha, \sigma^2 \mid \{y_{cn}\}) = \int \pi(\alpha \mid \beta, \psi^2) \pi(\sigma^2 \mid \nu, \tau^2) \pi(\beta, \tau^2, \nu, \tau^2 \mid \{y_{cn}\}) d(\beta, \tau^2, \nu, \tau^2).
$$

For sufficiently large $S$, a Monte Carlo estimate of $\pi(\alpha, \sigma^2 \mid \{y_{cn}\})$ is given by

$$
S^{-1} \sum_{s=1}^S n(\alpha \mid \beta_s, \psi_s^2) \gamma^{-1}(\sigma^2 \mid \nu_s/2, 2/\tau_s^2), \quad s = 1, \ldots, S
$$

is a sequence of random numbers drawn from the posterior distribution. Finally, we approximate the Monte Carlo estimate with the mixture distribution (2.8). The mixture weights $\{\omega_k\}$ and the hyperparameters $\{b_k, r_k, \nu_k, \tau_k^2\}$ are obtained by the method of Schmidli et al. (2014). In a numerical analysis, the EM algorithm can be applied (McLachlan and Krishnan (1997)). The Bayes estimator of $\alpha$ with the mixture prior distribution is given by (3.10).

A.2. Partial derivatives of $G$

For convenience in the numerical analysis, we represent the mixture weights $\{\omega_k\}$ as

$$
\omega_k = \frac{\exp(u_k)}{\sum_{k'=1}^K \exp(u_{k'})}, \quad k = 1, \ldots, K,
$$

where $u_K = 0$ to identify parameters $\{u_k\}$. In (3.13), since only the function $G(\bar{x}_s, \nu_s)$ for all $s$ includes $\{u_k, b_k, h_k, \nu_k, \tau_k^2\}$, it is sufficient to know the partial derivatives of $G$ with respect to these variables. In order to simplify the notation, we here omit the subscript $s$.

The parameter $u_i$ is included in all $\{\omega_k\}$. From properties of the partial derivative, we have

$$
\frac{\partial G}{\partial u_i} = \sum_{k=1}^K \frac{\partial G}{\partial \omega_k} \frac{\partial \omega_k}{\partial u_i},
$$
where
\[
\frac{\partial G}{\partial \omega_k} = \frac{1}{\sum_{k=1}^{K} \omega_k q_k} \{ -q_k G + h_k q_k (\bar{x} - b_k) \} = \frac{p_k}{\omega_k} \{ h_k (\bar{x} - b_k) - G \},
\]
\[
\frac{\partial \omega_k}{\partial u_i} = \begin{cases} 
-\omega_i \omega_k; & i \neq k \\
\omega_i - \omega_i^2; & i = k.
\end{cases}
\]

Therefore, we have
\[
\frac{\partial G}{\partial u_i} = -\omega_i \sum_{k=1}^{K} p_k \{ h_k (\bar{x} - b_k) - G \} + p_i \{ h_i (\bar{x} - b_i) - G \} = p_i \{ h_i (\bar{x} - b_i) - G \}.
\]

Since only \( b_i \) and \( q_i \) include \( b_i \) in \( G \), the partial derivative of \( G \) with respect to \( b_i \) is
\[
\frac{\partial G}{\partial b_i} = -h_i p_i + \frac{\omega_i}{\sum_{k=1}^{K} \omega_k q_k} \{ h_i (\bar{x} - b_i) - G \} \frac{\partial q_i}{\partial b_i},
\]
\[
\frac{\partial q_i}{\partial b_i} = M h_i (M + \nu_i) (\bar{x} - b_i) q_i (\tilde{\tau}_i^2)^{-1},
\]
where \( \tilde{\tau}_i^2 = M h_i (\bar{x} - b_i)^2 + V + \tau_i^2 \).

Similarly, we have
\[
\frac{\partial G}{\partial h_i} = p_i (\bar{x} - b_i) + \frac{\omega_i}{\sum_{k=1}^{K} \omega_k q_k} \{ h_i (\bar{x} - b_i) - G \} \frac{\partial q_i}{\partial h_i},
\]
\[
\frac{\partial q_i}{\partial h_i} = \frac{1}{2} q_i h_i^{-1} - M (M + \nu_i) (\bar{x} - b_i)^2 (\tilde{\tau}_i^2)^{-1}.
\]

For \( \nu_i \), the partial derivative of \( G \) with respect to \( \nu_i \) is
\[
\frac{\partial G}{\partial \nu_i} = \frac{\omega_i}{\sum_{k=1}^{K} \omega_k q_k} \{ h_i (\bar{x} - b_i) - G \} \frac{\partial q_i}{\partial \nu_i},
\]
\[
\frac{\partial q_i}{\partial \nu_i} = \frac{1}{2} q_i \left\{ \log \tilde{\tau}_i^2 + \Psi \left( \frac{M + \nu_i}{2} \right) - \Psi \left( \frac{\nu_i}{2} \right) - \log \tilde{\tau}_i^2 \right\},
\]
where \( \Psi \) denotes the digamma function defined as \( \Psi(x) = \frac{d}{dx} \log \Gamma(x) \) for \( x > 0 \).

Finally, we have
\[
\frac{\partial G}{\partial \tilde{\tau}_i^2} = \frac{\omega_i}{\sum_{k=1}^{K} \omega_k q_k} \{ h_i (\bar{x} - b_i) - G \} \frac{\partial q_i}{\partial \tilde{\tau}_i^2},
\]
\[
\frac{\partial q_i}{\partial \tilde{\tau}_i^2} = \frac{1}{2} q_i \left\{ \frac{\nu_i}{\tilde{\tau}_i^2} - \frac{(M + \nu_i) (\tilde{\tau}_i^2)^{-1}}{2} \right\}.
\]

**Acknowledgements**

The author is grateful to two anonymous referees for helpful comments that improved this paper considerably.
REFERENCES

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed., Wiley.
Billingsley, P. (1995). *Probability and Measure*, 3rd ed., Wiley.
Chib, S. (2001). Markov chain Monte Carlo methods: computation and inference, *Handbook of Econometrics* 5 (eds. J. J. Heckman and E. Leamer), pp. 3569–3649, Amsterdam, North-Holland.
Dallal, S. and Hall, W. (1983). Approximating priors by mixtures of natural conjugate priors, *J. R. Stat. Soc. Ser. B*, 45, 278–286.
Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, *Bayesian Statistics*, 4 (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 169–193, Oxford University Press, Oxford.
Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*, Wiley.
Hobbs, B. P., Carlin, B. P., Mandekar, S. J. and Sargent, D. J. (2011). Hierarchical commensurate and power prior models or adaptive incorporation of historical information in clinical trials, *Biometrics*, 67, 1047–1056.
Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models, *Stat. Sci.*, 15, 46–60.
Laws, D. J. and O’Hagan, A. (2000). Bayesian inference for rare errors in populations with unequal unit sizes, *J. R. Stat. Soc. Ser. C*, 49, 577–590.
Laws, D. J. and O’Hagan, A. (2002). A hierarchical Bayesian model for multilocation auditing, *J. R. Stat. Soc. Ser. D*, 51, 431–450.
McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, Wiley.
Neuenschwander, B. (2011). From historical data to priors, *Proceedings of Biopharmaceutical Section, Joint Statistical Meetings, American Statistical Association*, 3466–3474.
Sahu, S. K. and Smith, T. M. F. (2006). A Bayesian method of sample size determination with practical applications, *J. R. Stat. Soc. Ser. A*, 169, 235–253.
Schmidli, H., Gsteiger, S., Roychoudhury, S., O’Hagan, H., Spiegelhalter, D. and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information, *Biometrics*, 70, 1023–1032.
Schmidli, H., Neuenschwander, B. and Friede, T. (2016). Meta-analytic-predictive use of historical variance data for the design and analysis of clinical trials, *Comput. Stat. Data Anal*, Spiegelhalter, D. J., Abrams, K. R. and Myles J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Wiley.