Published in final edited form as:

J Stat Comput Simul. 2014 ; 85(15): 3108–3128. doi:10.1080/00949655.2014.951855.

Calibrating the prior distribution for a normal model with conjugate prior

Susan A. Alber and J. Jack Lee

Department of Biostatistics, University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, USA

Abstract

For a normal model with a conjugate prior, we provide an in depth examination of the effects of the hyperparameters on the long-run frequentist properties of posterior point and interval estimates. Under an assumed sampling model for the data generating mechanism, we examine how hyperparameter values affect the mean squared error (MSE) of posterior means and the true coverage of credible intervals. We develop two types of hyperparameter optimality. MSE optimal hyperparameters minimize the MSE of posterior point estimates. Credible interval optimal hyperparameters result in credible intervals that have minimum length while still retaining nominal coverage. A poor choice of hyperparameters has a worse consequence on the credible interval coverage than on the MSE of posterior point estimates. We give an example to demonstrate how our results can be used to evaluate the potential consequences of hyperparameter choices.

Keywords

Bayesian Model; Hyperparameter; Mean Squared Error; Point and Interval Estimates; Prior Specification

1 Introduction

Bayesian methods have drawn much attention in both theoretical development and applications to many applied fields in recent years. An important part of the model specification for the Bayesian method is the choice of the prior distribution, which represents prior knowledge about the unknown parameters in the likelihood function. There are many different perspectives on how the prior distribution should be chosen. For example, see Bernardo, Irony, and Singpurwalla (1979), Bernardo (1997), Kass and Wasserman (1996), Delampady et al. (2001), Kadane et al. (1980), Kadane and Wolfson (1998), and Garthwaite et al. (2005) for a wide varieties of principle on choosing priors. The distribution for the observed data, which constitutes the formulation of the likelihood function, is often chosen (or assumed to be) from the exponential family. This family contains many commonly used distributions, including normal, binomial, Poisson, and Gamma, etc. These distributions have many interesting and unique properties, including the fact that for every member of the exponential family there is a conjugate prior (Brown (1986)). In many applications, the probability distribution of the data can be considered as one of these distributions. In this case the data, the data likelihood function can be easily constructed, and the corresponding conjugate prior is a natural choice for the prior for the parameters. After

these easy choices have been made, one then needs to choose the parameters in the prior to fully specify the model. The main goal of this paper is to calibrate the prior distribution to evaluate the effect of the hyperparameters on the long-run frequentist properties of the posterior point and interval estimates.

When the likelihood function is a member of the exponential family and the prior is conjugate, the posterior mean of $\theta = E(X)$ can be written as a weighted average $E(\theta|\mathbf{x}) = wE(\theta) + (1 - w)\bar{x}$, where $E(\theta)$ is the prior mean, $\bar{x}$ is the sample mean, and $w = n_0/(n_0 + n)$, where n is the sample size and n_0 is a function of the prior variance of θ (Ericson (1970), Brown (1986), Robert (2001), page 122). This shows that the posterior mean is a weighted average of the prior mean and the sample mean, with weights proportional to n_0 and n , respectively. The prior can be viewed as adding n_0 observations from a hypothetical study. As the form of $E(\theta|\mathbf{x})$ shows, the value of n_0 is most influential when n is small. For fixed n_0 , as n approaches infinity, w approaches zero. Therefore, asymptotically the value of n_0 has no effect. Given θ_0 is the true mean, then since $\bar{x}$ is an unbiased estimate of θ_0 (i.e., $E(\bar{x}) = \theta_0$, where expectation is taken with respect to $\mathbf{x}$) and assuming $E(\theta) = \theta_0$, for a finite sample size $E(\theta|\mathbf{x})$ is a biased estimate of θ_0 . However, the variance of $E(\theta|\mathbf{x})$ (taken with respect to $\mathbf{x}$) is smaller than the variance of $\bar{x}$. As w increases, the absolute value of the bias of $E(\theta|\mathbf{x})$ increases, while the variance of $E(\theta|\mathbf{x})$ decreases, where both the bias and variance are with respect to $\mathbf{x}$.

This type of variance bias tradeoff is also found in many frequentist methods, where there is a choice between two different estimators: one of which has a smaller bias but a larger variance while the other one has a larger bias but a smaller variance. In practical applications, one may prefer the estimator with the larger bias (and smaller variance) if it also has a smaller mean squared error (MSE). The logic is that since we only have a single sample, we may want to choose the estimator such that our one estimate is likely to be closer to the true value considering the sampling variability, even though it is biased over repeated sampling. In the Bayesian approach, as the posterior mean given above shows, instead of choosing one estimator over the other, the biased estimate (from the prior) and the unbiased estimate (from the sample) are combined into a single estimate.

Various approaches have been proposed for measuring the sensitivity of the posterior to the prior. For example, see Kass (1989), McCulloch (1989), Weiss (1996), and Millar and Stewart (2007). One approach is to examine the effects of the prior on the frequentist properties of posterior point and interval estimates (Ren et al. (2004)). This approach can also be used to show the relationships between both posterior point estimates and frequentist estimates and between credible intervals and confidence intervals. Using frequentist properties to guide the selection of the prior has been called the calibrated Bayes approach. Arguments in favor of this approach are discussed by Dawid (1982) and Rubin (1984). The calibrated Bayes approach can be viewed as a bridge between Bayesian and frequentist methods (Little, 2006, and Bayarri and Berger, 2004).

For a normal model with a conjugate prior, we provide an in depth examination of the effects of the hyperparameters on the long-run frequentist properties of posterior point and interval estimates. We demonstrate the effects of hyperparameters on the mean squared error (MSE)

of posterior point estimates and on the coverage of credible intervals. We compare the MSE of the posterior means and coverage of credible intervals to the MSE and coverage of confidence intervals from maximum likelihood estimation. We define two types of hyperparameter optimality. An MSE optimal hyperparameter gives the minimum MSE of the posterior expected value. A credible interval optimal hyperparameter gives a credible interval with the minimum length such that the true coverage equals the nominal coverage. We discuss the known variance model in Section 2 and the unknown variance model in Section 3. In Section 4 we demonstrate how our results can be used to evaluate the potential consequences of hyperparameter choices in a real data analysis. We close with a discussion in Section 5 where we highlight some general principles from our results that are likely to extend to other models. We also provide an appendix that generalizes our key results to any estimate that is a weighted average, and give an example of the application to estimation of the difference between two normal means.

2 Known Variance Model

Let $\mathbf{y} = (y_1, \dots, y_n)$ be the data, where n is the sample size and for $i = 1, \dots, n$, y_i is normally distributed with mean μ and known variance τ_o . Let the prior for μ be normal with mean μ_m and variance τ_o/n_μ . The model is

$$y_i | \mu \sim N(\mu, \tau_o) \quad i=1, \dots, n \quad (1)$$

$$\mu \sim N\left(\mu_m, \frac{\tau_o}{n_\mu}\right),$$

which is a standard model. (For example, see Gelman et al. (2004), pages 42–45.) The likelihood function (first line of the model) assumes the y_i s are independent with identical marginal normal distributions.

All parameters that do not appear on the left of a tilde are fixed. Subscript o is used for known parameters in the sampling density, such as τ_o . We call all remaining parameters hyperparameters. In model (1) the hyperparameters are μ_m and n_μ . We assume that μ_m has been selected and focus on how to select n_μ .

Assume the true distribution of the data is

$$y_i \sim N(\mu_o, \tau_o) \quad i=1, \dots, n \quad (2)$$

for the entire paper, including Section 3 for the unknown variance model. All expectations, variances, and mean squared errors taken with respect to $\mathbf{y}$ assume model (2). Define the bias of the hyperparameter μ_m with respect to the true mean μ_o as

$$\Delta_\mu = \mu_m - \mu_o.$$

Given model (1), the posterior distribution of μ is

$$\mu|\mathbf{y} \sim N\left(w_{\mu}\mu_m + (1 - w_{\mu})\bar{y}, (1 - w_{\mu})\frac{\tau_o}{n}\right), \quad (3)$$

where

$$w_{\mu} = \frac{n_{\mu}}{n_{\mu} + n}$$

and $\bar{y}$ is the mean of $\mathbf{y}$. The posterior expected value of μ , given in equation (3), is a weighted average of the prior mean μ_m and the sample mean $\bar{y}$, with weights proportional to n_{μ} and n , respectively (Gelman et al. (2004), page 43). Therefore, n_{μ} can be viewed as a prior sample size, $n_{\mu} + n$ as the total effective sample size, and w_{μ} as the weight of the prior as a fraction of the total effective sample size.

2.1 Effect of w_{μ} on the MSE of $E(\mu|\mathbf{y})$

With respect to the distribution of $\mathbf{y}$ given by model (2), the posterior expected value $E(\mu|\mathbf{y})$ is normally distributed with mean, variance, and mean squared error (MSE)

$$E_{\mathbf{y}}[E(\mu|\mathbf{y})] = \mu_o + w_{\mu}\Delta_{\mu},$$

$$\text{var}_{\mathbf{y}}[E(\mu|\mathbf{y})] = (1 - w_{\mu})^2 \frac{\tau_o}{n},$$

and

$$\text{MSE}[E(\mu|\mathbf{y})] = \text{var}_{\mathbf{y}}[E(\mu|\mathbf{y})] + \text{bias}^2[E(\mu|\mathbf{y})] = (1 - w_{\mu})^2 \frac{\tau_o}{n} + w_{\mu}^2 \Delta_{\mu}^2.$$

We use subscript $\mathbf{y}$ for expectation and variance when they are taken with respect to the distribution of $\mathbf{y}$ to differentiate them from posterior expectation and variance. We do not use a subscript for the MSE or the bias since these are always taken with respect to $\mathbf{y}$.

For the MSE of $E(\mu|\mathbf{y})$, we have the usual tradeoff between bias and variance; as w_{μ} increases, weighting $E(\mu|\mathbf{y})$ towards μ_m , the bias increases and the variance decreases. Figure 1(a) shows the effect of w_{μ} on the MSE of $E(\mu|\mathbf{y})$. The y-axis is the ratio of the MSE of $E(\mu|\mathbf{y})$ to the MSE of $\bar{y}$

$$\frac{\text{MSE}[E(\mu|\mathbf{y})]}{\text{MSE}(\bar{y})} = (1 - w_{\mu})^2 + w_{\mu}^2 \frac{\Delta_{\mu}^2}{\frac{\tau_o}{n}}, \quad (4)$$

where $\text{MSE}(\bar{y}) = \text{var}(\bar{y}) = \tau_o/n$. We can think of $|\Delta_\mu|/\sqrt{\tau_o/n}$ as the standardized bias, which is the distance, measured in standard deviations of $\bar{y}$, between the prior mean μ_m and the true mean μ_o . The smaller the value of $|\Delta_\mu|/\sqrt{\tau_o/n}$ the greater the possible reduction in the MSE of $E(\mu|y)$.

The value of $w_\mu \in (0, 1)$ that minimizes the MSE of $E(\mu|y)$, which we call the MSE optimal choice and denote as w_μ^* , is

$$w_\mu^* = \frac{\frac{\tau_o}{n}}{\Delta_\mu^2 + \frac{\tau_o}{n}} = \frac{\frac{1}{\text{MSE}(\mu_m)}}{\frac{1}{\text{MSE}(\mu_m)} + \frac{1}{\text{MSE}(\bar{y})}}, \quad (5)$$

where $\text{MSE}(\mu_m) = \text{bias}^2(\mu_m) = \Delta_\mu^2$. Since $\bar{y}$ is unbiased, $1/\text{MSE}(\bar{y}) = 1/\text{var}(\bar{y})$, which is the precision of $\bar{y}$. More generally, we can interpret the inverse of the MSE of a (possibly biased) estimate as a measure of the amount of information provided by the estimate. Then, the amount of information provided by μ_m is $1/\text{MSE}(\mu_m)$, the amount of information provided by $\bar{y}$ is $1/\text{MSE}(\bar{y})$, and w_μ^* is the fraction of information provided by μ_m relative to the sum of the information provided by $\bar{y}$ and μ_m .

When $w_\mu = w_\mu^*$, the posterior expected value is

$$E(\mu|y, w_\mu^*) = \frac{\frac{1}{\text{MSE}(\mu_m)}\mu_m + \frac{1}{\text{MSE}(\bar{y})}\bar{y}}{\frac{1}{\text{MSE}(\mu_m)} + \frac{1}{\text{MSE}(\bar{y})}}, \quad (6)$$

which is a weighted average of μ_m and $\bar{y}$, with weights proportional to $1/\text{MSE}(\mu_m)$, the amount of information provided by μ_m , and $1/\text{MSE}(\bar{y})$, the amount of information provided by $\bar{y}$, respectively, and the posterior variance is

$$\text{var}(\mu|y, w_\mu^*) = (1 - w_\mu^*)\text{MSE}(\bar{y}) = w_\mu^*\text{MSE}(\mu_m) = \left(\frac{1}{\text{MSE}(\mu_m)} + \frac{1}{\text{MSE}(\bar{y})} \right)^{-1},$$

which is the inverse of the sum of the information provided by μ_m and $\bar{y}$.

Returning to the original model parameterization, the MSE optimal n_μ is

$$n_\mu^* = n \left(\frac{w_\mu^*}{1 - w_\mu^*} \right) = \frac{\tau_o}{\Delta_\mu^2} = \frac{1}{\left(\frac{1}{n} \right) \frac{1}{\text{MSE}(\bar{y})}}. \quad (7)$$

We can interpret $(1/n)[1/\text{MSE}(\bar{y})]$ as a measure of the amount of information provided by a single observation y_i . Then, n_μ^* is the fraction of information provided by the prior mean μ_m

relative to the information provided by a single observation y_i . When $n_\mu = n_\mu^*$, equivalently $w_\mu = w_\mu^*$, we have $\text{var}(\mu) = \tau_o / n_\mu^* = \Delta_\mu^2$. Therefore, when $n_\mu = n_\mu^*$ the standard deviation of μ is $|\mu|$, and thus the prior mean μ_m is one prior standard deviation away from the true mean μ_o .

As seen in Figure 1(a) and below in equation (9), the MSE of $E(\mu|y)$ is symmetric around w_μ^* and

$$\text{MSE}[E(\mu|y)] \begin{cases} < \text{MSE}(\bar{y}) & \text{if } 0 < w_\mu < \min\{2w_\mu^*, 1\}; \\ = \text{MSE}(\bar{y}) & \text{if } w_\mu = 0 \text{ or } w_\mu = \min\{2w_\mu^*, 1\}; \\ > \text{MSE}(\bar{y}) & \text{if } 2w_\mu^* < w_\mu \leq 1. \end{cases} \quad (8)$$

If $\text{MSE}(\bar{y}) > \text{MSE}(\mu_m)$, then $w_\mu^* > 1/2$ and $\text{MSE}[E(\mu|y)] < \text{MSE}(\bar{y})$ for any $w_\mu \in (0, 1)$.

The value of w_μ affects both the MSE of $E(\mu|y)$ and the posterior variance $\text{var}(\mu|y)$. These effects have important implications for credible intervals. Written as functions of w_μ^* and $w_\mu - w_\mu^*$, we have

$$\text{MSE}[E(\mu|y)] = (1 - w_\mu^*) \frac{\tau_o}{n} + (w_\mu - w_\mu^*)^2 \left(\Delta_\mu^2 + \frac{\tau_o}{n} \right) \quad (9)$$

and

$$\text{var}(\mu|y) = (1 - w_\mu^*) \frac{\tau_o}{n} - (w_\mu - w_\mu^*) \frac{\tau_o}{n}. \quad (10)$$

Subtracting equation (10) from equation (9), substituting $w_\mu^* = (\tau_o/n)/(\Delta_\mu^2 + \tau_o/n)$, and rearranging gives

$$\text{MSE}[E(\mu|y)] = \text{var}(\mu|y) + w_\mu (w_\mu - w_\mu^*) \left(\Delta_\mu^2 + \frac{\tau_o}{n} \right), \quad (11)$$

which shows that the MSE of $E(\mu|y)$ is less than, equal to, or greater than $\text{var}(\mu|y)$ if w_μ is less than, equal to, or greater than w_μ^* , respectively. Therefore, when $w_\mu < w_\mu^*$ the posterior variance provides a conservative estimate of the true uncertainty (as quantified by the MSE) of $E(\mu|y)$ and when $w_\mu > w_\mu^*$ the posterior variance provides an underestimate of the true uncertainty.

2.2 Effect of w_μ on Credible Intervals

A $(1 - \alpha)100\%$ credible interval for μ is

$$E(\mu|\mathbf{y}) \pm Z_{1-\alpha/2} \sqrt{(1-w_\mu) \frac{\tau_o}{n}}.$$

The probability that this interval contains μ_o (i.e, the true coverage) is

$$C_\alpha(w_\mu) = \Phi \left[\left(\frac{w_\mu}{1-w_\mu} \right) \frac{|\Delta_\mu|}{\sqrt{\frac{\tau_o}{n}}} + \frac{Z_{1-\alpha/2}}{\sqrt{1-w_\mu}} \right] - \Phi \left[\left(\frac{w_\mu}{1-w_\mu} \right) \frac{|\Delta_\mu|}{\sqrt{\frac{\tau_o}{n}}} - \frac{Z_{1-\alpha/2}}{\sqrt{1-w_\mu}} \right], \quad (12)$$

where $\Phi(\cdot)$ is the distribution function and Z_p is the p quantile of a standard normal variable. The length of a $(1-\alpha)100\%$ credible interval is

$$L_\alpha(w_\mu) = 2Z_{1-\alpha/2} \sqrt{(1-w_\mu) \frac{\tau_o}{n}} = \sqrt{(1-w_\mu)} [\text{length of a } (1-\alpha)100\% \text{ confidence interval}].$$

True coverage, given in equation (12), can be written as

$$C_\alpha(w_\mu) = \Phi \left[\frac{|\text{bias}[E(\mu|\mathbf{y})]|}{\text{SD}_y[E(\mu|\mathbf{y})]} + Z_{1-\alpha/2} \frac{\text{SD}(\mu|\mathbf{y})}{\text{SD}_y[E(\mu|\mathbf{y})]} \right] - \Phi \left[\frac{|\text{bias}[E(\mu|\mathbf{y})]|}{\text{SD}_y[E(\mu|\mathbf{y})]} - Z_{1-\alpha/2} \frac{\text{SD}(\mu|\mathbf{y})}{\text{SD}_y[E(\mu|\mathbf{y})]} \right],$$

where $\text{SD}(\mu|\mathbf{y}) = \sqrt{\text{var}(\mu|\mathbf{y})} = \sqrt{(1-w_\mu)(\tau_o/n)}$ and

$\text{SD}_y[E(\mu|\mathbf{y})] = \sqrt{\text{var}_y[E(\mu|\mathbf{y})]} = (1-w_\mu) \sqrt{\tau_o/n}$. A sufficient condition for $C_\alpha(w_\mu) > (1-\alpha)$ is

$$w_\mu |\Delta_\mu| = |\text{bias}[E(\mu|\mathbf{y})]| < Z_{1-\alpha/2} \{ \text{SD}(\mu|\mathbf{y}) - \text{SD}_y[E(\mu|\mathbf{y})] \}. \quad (13)$$

(See appendix, section A.1 for a derivation of equation (13).) This shows that $C_\alpha(w_\mu) > (1-\alpha)$ if $\text{SD}(\mu|\mathbf{y}) - \text{SD}_y[E(\mu|\mathbf{y})]$ is sufficiently large relative to $|\text{bias}[E(\mu|\mathbf{y})]|$, which means the posterior standard deviation (which determines the length of the credible interval) is sufficiently larger than the standard deviation of $E(\mu|\mathbf{y})$. As α decreases, $Z_{1-\alpha/2}$ increases, and consequently the value of $\text{SD}(\mu|\mathbf{y}) - \text{SD}_y[E(\mu|\mathbf{y})]$ needed to guarantee $C_\alpha(w_\mu) > (1-\alpha)$ decreases.

For a 95% credible interval, the effect of w_μ on $C_{0.05}(w_\mu)$ is shown in Figure 1(b) and the effect of w_μ on the ratio of $L_{0.05}(w_\mu)$ to the length of a 95% confidence interval is shown in Figure 1(c). For $w_\mu = 0$, $L_\alpha(w_\mu)$ is the same as the length of a $(1-\alpha)100\%$ confidence interval. As w_μ begins to increase above 0, although $L_\alpha(w_\mu)$ decreases making the credible interval smaller than the confidence interval, $C_\alpha(w_\mu)$ increases because the MSE of $E(\mu|\mathbf{y})$ is decreasing at a sufficiently fast rate. However, as w_μ continues to increase, at some point $C_\alpha(w_\mu)$ reaches a maximum, and then decreases rapidly. At the maximum $C_\alpha(w_\mu)$, we have

$w_\mu < w_\mu^*$ because as w_μ increases above w_μ^* , $C_\alpha(w_\mu)$ decreases since the MSE of $E(\mu|y)$ is increasing while $L_\alpha(w_\mu)$ is decreasing. As $|\Delta_\mu|/\sqrt{\frac{\tau_o}{n}}$ decreases, there is an increase in both the maximum $C_\alpha(w_\mu)$ and the value of w_μ at this maximum.

Figure 2 shows the effect of w_μ on $C_\alpha(w_\mu)$ for $\alpha \in \{0.01, 0.05, 0.1, 0.15\}$ and $w_\mu^* = 0.1$. The general pattern is the same for the different α . When $w_\mu < w_\mu^*$, $C_\alpha(w_\mu)$ is above or approximately equal to $1 - \alpha$. As w_μ increases above w_μ^* , $C_\alpha(w_\mu)$ decreases rapidly with the fastest declines for larger α .

We define the $(1 - \alpha)100\%$ credible interval optimal w_μ as the value that gives a $(1 - \alpha)100\%$ credible interval with the smallest $L_\alpha(w_\mu)$ such that $C_\alpha(w_\mu) = 1 - \alpha$. We denote this value as $w_\mu^\dagger(\alpha)$. Since $L_\alpha(w_\mu)$ decreases as w_μ increases, $w_\mu^\dagger(\alpha)$ is the largest $w_\mu \in (0, 1)$ such that $C_\alpha(w_\mu) = 1 - \alpha$. There is no closed form solution for $w_\mu^\dagger(\alpha)$, but the value is easily approximated using a numerical search algorithm.

Figure 3 shows $w_\mu^\dagger(\alpha)$ as a function of $|\Delta_\mu|/\sqrt{\tau_o/n}$ for $\alpha \in \{0.01, 0.05, 0.1, 0.2\}$. For any $|\Delta_\mu|/\sqrt{\tau_o/n}$, the value of $w_\mu^\dagger(\alpha)$ decreases as α increases. Figure 3 also shows the relationship between $w_\mu^\dagger(\alpha)$ and w_μ^* (shown as a solid line). This relationship is important because if $w_\mu^\dagger(\alpha) > w_\mu^*$, then $C_\alpha(w_\mu^*) > 1 - \alpha$, since $C_\alpha(w_\mu)$ decreases as w_μ increases above w_μ^* . The differences between $w_\mu^\dagger(\alpha)$ (the credible optimal value) and w_μ^* (the MSE optimal value) are generally not large and are quite small when $|\Delta_\mu|/\sqrt{\tau_o/n}$ is very small or very large.

3 Unknown Variance Model

In this section we discuss the effects of hyperparameters in a normal model with unknown variance. For $i = 1, \dots, n$, let y_i be normally distributed with mean μ and variance τ . Let the prior for μ be normal with mean μ_m and variance τ/n_μ and the prior for τ be inverse gamma with shape parameter a , scale parameter b , and density proportional to $\tau^{-(a+1)} \exp(-b/\tau)$, where $a > 0$ and $b > 0$. The model is

$$y_i | \mu, \tau \sim N(\mu, \tau) \quad i=1, \dots, n$$

$$\mu | \tau \sim N\left(\mu_m, \frac{\tau}{n_\mu}\right)$$

$$\tau \sim \text{Inverse Gamma}(a, b),$$

which is a standard model. (For example, see Gelman et al. (2004), pages 71–73.) The model makes the same assumptions for the likelihood function (first line of the model) as the known variance model which assumes the y_i s are independent with identical marginal normal distributions. The hyperparameters (values that need to be selected) are μ_m , n_μ , a , and b .

As in Section 2, we assume that equation (2) is the true distribution of $\mathbf{y} = (y_1, \dots, y_n)$. We discuss the effects of hyperparameters on the MSE of $E(\mu|\mathbf{y})$ in Section 3.1 and on the MSE of $E(\tau|\mathbf{y})$ in Section 3.2. We discuss the effects of hyperparameters on the credible interval for μ in Section 3.3.

3.1 Effect of w_μ on the MSE of $E(\mu|\mathbf{y})$

Unconditional on τ , the posterior distribution of μ is

$$\mu|\mathbf{y} \sim t \left\{ 2a+n, w_\mu \mu_m + (1-w_\mu) \bar{y}, \frac{1-w_\mu}{2a+n} \left[\frac{2b}{n} + w_\mu \frac{R(\mu_m)}{n} + (1-w_\mu) \frac{R(\bar{y})}{n} \right] \right\},$$

where

$$R(\mu_m) = \sum_{i=1}^n (y_i - \mu_m)^2, \quad (14)$$

$$R(\bar{y}) = \sum_{i=1}^n (y_i - \bar{y})^2, \quad (15)$$

and $t(d, e, f)$ is a t distribution with degrees of freedom d , location parameter e , and scale parameter f (Gelman et al. (2004), page 73). The posterior expected value $E(\mu|\mathbf{y})$ is the same as for the known variance models. Therefore all results in Section 2.1 on minimization of the MSE of $E(\mu|\mathbf{y})$ apply to both models.

3.2 Effect of w_μ and w_τ on the MSE of $E(\mu|\mathbf{y})$

The posterior distribution of τ is

$$\tau|\mathbf{y} \sim \text{Inverse Gamma} \left(\frac{n_\tau + n + 2}{2}, \frac{1}{2} [n_\tau \tau_m + w_\mu R(\mu_m) + (1-w_\mu) R(\bar{y})] \right)$$

and the posterior mean of τ is

$$E(\tau|\mathbf{y}) = \frac{2b + w_\mu R(\mu_m) + (1-w_\mu) R(\bar{y})}{n_\tau + n}, \quad (16)$$

where

$$n_{\tau}=2(a-1) \quad (17)$$

and $R(\mu_m)$ and $R(\bar{y})$ are given by equations (14) and (15) (Gelman et al. (2004), page 72). Given the form of $E(\tau|\mathbf{y})$, we interpret $2b$ as the prior sum of squares and n_{τ} as the prior sample size. The numerator of $E(\tau|\mathbf{y})$ is the sum of the prior sum of squares $2b$ and a weighted average of two sum of squares from the data: one around the prior mean μ_m and one around the sample mean $\bar{y}$. The denominator of $E(\tau|\mathbf{y})$ is the prior sample size n_{τ} plus the data sample size n . Let

$$\tau_m = \frac{b}{a-1}, \quad (18)$$

which is the prior mean of τ if $a > 1$. If $a > 1$, then the prior mean is equal to the prior sum of squares $2b$ divided by the prior sample size $n_{\tau} = 2(a-1)$. Let

$$w_{\tau} = \frac{n_{\tau}}{n_{\tau} + n}.$$

We assume μ_m and τ_m have been chosen, and focus on how w_{μ} and w_{τ} affect the posterior. Given τ_m and w_{τ} , values of a and b follow from the inverse functions $a = n_{\tau}/2 + 1$ and $b = \tau_m(n_{\tau}/2)$, where $n_{\tau} = nw_{\tau}/(1 - w_{\tau})$.

We interpret $w_{\tau} = 0$ ($n_{\tau} = 0$, $a = 1$) as a prior sample size of zero. If $a = 1$, then as both b and w_{μ} approach zero, $E(\tau|\mathbf{y})$ approaches the maximum likelihood estimate of τ_0 . Assuming fixed τ_m and using the inverse function given above, as a approaches 1, b approaches 0. If $a = 1$, then as b approaches 0, the limiting prior distribution for the variance is not flat ($\tau \propto \tau^{-2}$). However the prior for the precision ($1/\tau$) is a gamma distribution with shape parameter a , scale parameter b , and density proportional to $\tau^{(a-1)} \exp(-\tau b)$, which does have a flat limiting distribution ($1/\tau \propto 1$) when $a = 1$ and b approaches 0.

We assume $a \neq 1$, which means $w_{\tau} \neq 0$. If $a = 1$, then τ_m is indeterminate and the prior mean does not exist. However, this does not cause problems with any of the equations given in the paper because τ_m is always multiplied by n_{τ} , and $n_{\tau}\tau_m = 2(a-1)[b/(a-1)] = 2b$.

The posterior expected value of τ given by equation (16) can also be written as

$$E(\tau|\mathbf{y}) = w_{\tau}\tau_m + (1 - w_{\tau})\overline{\Sigma}(w_{\mu}), \quad (19)$$

where

$$\overline{\Sigma}(w_\mu) = w_\mu \frac{R(\mu_m)}{n} + (1 - w_\mu) \frac{R(\bar{y})}{n}, \quad (20)$$

and $R(\mu_m)$ and $R(\bar{y})$ are given by equations (14) and (15). The quantity $\Sigma(w_\mu)$ is a weighted average of $R(\mu_m)/n$, the maximum likelihood estimate of τ_o if μ_m is the known true mean, and $R(\bar{y})/n$, the maximum likelihood estimate of τ_o if the mean is unknown, with weights proportional to n_μ and n , respectively. The expectations, biases, variances, and MSEs of τ_m , $R(\mu_m)/n$ and $R(\bar{y})/n$ are given in Table 1.

The posterior expected value, given by equation (19), is a weighted average of the prior mean τ_m and $\Sigma(w_\mu)$ with weights proportional to n_τ and n , respectively. Therefore, n_τ can be viewed as the prior sample size, $n_\tau + n$ as the total effective sample size, and w_τ as the weight of the prior as a fraction of the total effective sample size.

The bias of $E(\tau|\mathbf{y})$ is

$$\text{bias } [E(\tau|\mathbf{y})] = E(\tau|\mathbf{y}) - \tau_o = w_\tau \Delta_\tau + (1 - w_\tau) \text{bias } [\overline{\Sigma}(w_\mu)],$$

where

$$\Delta_\tau = \tau_m - \tau_o$$

is the bias of the hyperparameter τ_m and

$$\text{bias } [\overline{\Sigma}(w_\mu)] = w_\mu \Delta_\mu^2 + (1 - w_\mu) \left(-\frac{\tau_o}{n} \right),$$

is the bias of $\Sigma(w_\mu)$, which can also be written as

$$\text{bias } [\overline{\Sigma}(w_\mu)] = (w_\mu - w_\mu^*) \left(\Delta_\mu^2 + \frac{\tau_o}{n} \right). \quad (21)$$

Equation (21) shows that the bias of $\Sigma(w_\mu)$ is less than, equal to, or greater than 0 if w_μ is less than, equal to, or greater than w_μ^* , respectively. Table 2 shows the effect of w_μ on the magnitude of the bias of $\Sigma(w_\mu)$. If $w_\mu < w_\mu^*$, then $-(1 - w_\mu)\tau_o/n < \text{bias } [\Sigma(w_\mu)] < 0$. Therefore, if $w_\mu < w_\mu^*$, then as n increases, the bias of $\Sigma(w_\mu)$ approaches zero. If $w_\mu > w_\mu^*$, then $0 < \text{bias } [\overline{\Sigma}(w_\mu)] < w_\mu \Delta_\mu^2$.

The MSE of $E(\tau|\mathbf{y})$ is

$$\text{MSE } [E(\tau|\mathbf{y})] = w_\tau^2 \Delta_\tau^2 + (1 - w_\tau)^2 \text{MSE}[\overline{\Sigma}(w_\mu)] + 2w_\tau(1 - w_\tau) \Delta_\tau \text{bias } [\overline{\Sigma}(w_\mu)]. \quad (22)$$

The effect of the direction of the bias of τ_m (i.e., the sign of τ) on the MSE of $E(\tau|y)$ depends on the direction of the bias of $Z(w_\mu)$ which, as explained above, depends on w_μ . If $w_\mu = w_\mu^*$, then $\text{bias}[Z(w_\mu)] = 0$, the last term in equation (22) is zero, and the MSE does not depend on the sign of τ . If $w_\mu \neq w_\mu^*$, then the MSE of $E(\tau|y)$ is smaller when the biases of τ_m and $Z(w_\mu)$ are in opposite directions: if $w_\mu < w_\mu^*$, then $\text{bias}[Z(w_\mu)] < 0$ and the MSE of $E(\tau|y)$ is smaller when $\tau > 0$; if $w_\mu > w_\mu^*$, then $\text{bias}[Z(w_\mu)] > 0$ and the MSE of $E(\tau|y)$ is smaller when $\tau < 0$.

The ratio of the MSE of $E(\tau|y)$ to the MSE of $R(\bar{y})/(n-1)$ (an unbiased frequentist estimate of τ_o) is

$$\begin{aligned} \frac{\text{MSE}[E(\tau|y)]}{\text{MSE}\left[\frac{R(\bar{y})}{n-1}\right]} &= w_\tau^2 \frac{\Delta_\tau^2}{\frac{2\tau_o^2}{n}} + (1-w_\tau)^2 \left(1 - \frac{1}{n}\right) + \sqrt{\frac{2}{n}} w_\tau (1-w_\tau) \frac{\Delta_\tau}{\sqrt{\frac{2\tau_o^2}{n}}} \left(\frac{w_\mu}{w_\mu^*} - 1\right) \\ &\quad + \frac{1}{2n} (1-w_\tau)^2 \left(\frac{w_\mu}{w_\mu^*} - 1\right)^2 + \frac{1}{n} (1-w_\tau)^2 \left(\frac{w_\mu}{w_\mu^*}\right)^2 [1 - (1-w_\mu^*)^2], \end{aligned} \quad (23)$$

where $\text{MSE}[R(\bar{y})/(n-1)] = 2\tau_o^2/n$. We can think of

$\Delta_\tau / \sqrt{\text{MSE}[R(\bar{y})/(n-1)]} = \Delta_\tau / \sqrt{2\tau_o^2/n}$ as the standardized bias of τ_m . As n increases the ratio of MSEs becomes less dependent on w_μ . When n is large the last three terms in equation (23) are small and therefore w_μ has minimal effect the ratio of MEs.

Figure 4 shows $\text{MSE}[E(\tau|y)]/\text{MSE}[R(\bar{y})/(n-1)]$ as a function of w_τ , $\Delta_\tau / \sqrt{2\tau_o^2/n}$, and w_μ/w_μ^* with $n = 20$ and $w_\mu^* = 0.2$. (The shape of the function is minimally dependent on the value of w_μ^* .) For any fixed w_μ , the ratio of MSEs decreases as w_τ increases above zero, increases as w_τ increases above w_μ^* (the MSE optimal choice, see Section 3.2.2), and is symmetric around w_μ^* . The effect of w_μ on the ratio of MSEs is generally small and the direction of the effect depends on the sign of τ . If $\tau > 0$, then as w_μ increases the ratio of MSEs increases. If $\tau < 0$, then as w_μ increases the ratio of MSEs increases if w_τ is small and decreases if w_τ is large.

3.2.1 Minimization of the MSE of $E(\tau|y)$ with respect to w_μ —We show the value of w_μ that minimizes the MSE of $E(\tau|y)$, which we denote as $w_\mu^{**}(n_\tau)$, as a function of n_τ instead of w_τ , because it simplifies the function. We also give $w_\mu^{**}(n_\tau)$ as a function of w_μ^* because we would like to be able to set w_μ with the primary focus on estimation of μ . Therefore the relationship between $w_\mu^{**}(n_\tau)$ and w_μ^* is important.

For a given n_τ , the value of $w_\mu \in [0, 1]$ that minimizes the MSE of $E(\tau|y)$ is

$$w_{\mu}^{**}(n_{\tau}) = \begin{cases} 0 & \text{if } \frac{n_{\tau}\Delta_{\tau}}{\tau_0} \geq 1; \\ w_{\mu}^* \left[\frac{1 - \frac{n_{\tau}\Delta_{\tau}}{\tau_0}}{3 - 2(1 - w_{\mu}^*)^2} \right] & \text{if } -A \leq \frac{n_{\tau}\Delta_{\tau}}{\tau_0} \leq 1 \\ 1 & \text{if } \frac{n_{\tau}\Delta_{\tau}}{\tau_0} \leq -A, \end{cases} \quad (24)$$

where

$$A = \frac{\frac{\Delta_{\mu}}{\left(\frac{\tau_0}{n}\right)^2} + 5\frac{\Delta_{\mu}^2}{\frac{\tau_0}{n}} + 2}{\frac{\Delta_{\mu}^2}{\frac{\tau_0}{n}} + 1}.$$

The first condition is equivalent to the second case equation being less than or equal to zero. The third condition is equivalent to the second case equation being greater than or equal to 1. Equation (24) shows that $w_{\mu}^{**}(n_{\tau})$ is less than, equal to, or greater than w_{μ}^* if n_{τ}/τ_0 is greater than, equal to, or less than $2[(1 - w_{\mu}^*)^2 - 1]$, respectively. If $n_{\tau}/\tau_0 > 0$ then $w_{\mu}^{**}(n_{\tau}) < w_{\mu}^*$ because $2[(1 - w_{\mu}^*)^2 - 1] < 0$. If the second condition holds, then $w_{\mu}^{**}(n_{\tau})$ decreases linearly as n_{τ}/τ_0 increases.

3.2.2 Minimization of the MSE of $E(\tau|y)$ with respect to w_{μ} —For a given w_{μ} , the value of w_{τ} that minimizes the MSE of $E(\tau|y)$, which we call the MSE optimal w_{τ} and denote as w_{τ}^* , is

$$w_{\tau}^* = \begin{cases} 0 & \text{if } \text{MSE}(\bar{\tau}) \leq \Delta_{\tau} \text{bias}(\bar{\tau}) < \text{MSE}(\tau_m); \\ \frac{\text{MSE}(\bar{\tau}) - \Delta_{\tau} \text{bias}(\bar{\tau})}{\text{MSE}(\tau_m) + \text{MSE}(\bar{\tau}) - 2\Delta_{\tau} \text{bias}(\bar{\tau})} & \text{if } \Delta_{\tau} \text{bias}(\bar{\tau}) \leq \min \{ \text{MSE}(\bar{\tau}), \text{MSE}(\tau_m) \}; \\ 1 & \text{if } \text{MSE}(\tau_m) \leq \Delta_{\tau} \text{bias}(\bar{\tau}) < \text{MSE}(\bar{\tau}), \end{cases} \quad (25)$$

where $\mathbf{Z} = \mathbf{Z}(w_{\mu})$ and $\text{MSE}(\tau_m) = \Delta_{\tau}^2$. The three conditions are mutually exclusive because $\tau \text{bias}[\mathbf{Z}(w_{\mu})] = \max \{ \text{MSE}(\mathbf{Z}), \text{MSE}(\tau_m) \}$. (See Section A.2.) The first condition is equivalent to the second case equation being less than or equal to 0. The third condition is equivalent to the second case equation being greater than or equal to 1.

The second condition holds if $\tau \text{bias}[\mathbf{Z}(w_{\mu})] = 0$. As explained in Section 3.2, the bias of $\mathbf{Z}(w_{\mu})$ is less than, equal to, or greater than zero if w_{μ} is less than, equal to, or greater than w_{μ}^* , respectively. Therefore, if either $w_{\mu} = w_{\mu}^*$ ($w_{\mu} < w_{\mu}^*$ and $\tau > 0$), or ($w_{\mu} > w_{\mu}^*$ and $\tau < 0$) then $\tau \text{bias}[\mathbf{Z}(w_{\mu})] = 0$ and the second condition holds. When the second condition doesn't hold, $w_{\tau}^* = 0$ if $\text{MSE}[\mathbf{Z}(w_{\mu})] < \text{MSE}(\tau_m)$ and $w_{\tau}^* = 1$ if $\text{MSE}(\tau_m) < \text{MSE}[\mathbf{Z}(w_{\mu})]$.

3.3 Effect of w_{μ} and w_{τ} on Credible Intervals for μ

A $(1 - \alpha)100\%$ credible interval for μ is

$$E(\mu|\mathbf{y}) \pm t_{n_\tau+n+2, 1-\alpha/2} \sqrt{\left(\frac{n_\tau+n}{n_\tau+n+2}\right) (1-w_\mu) \frac{E(\tau|\mathbf{y})}{n}},$$

where $t_{d,p}$ is the p quantile of a standard t distribution with d degrees of freedom. The probability that this interval contains μ_o (i.e., the true coverage) is

$$\begin{aligned} & E_{\mathbf{y}^\Phi} \left[\left(\frac{w_\mu}{1-w_\mu} \right) \frac{|\Delta_\mu|}{\sqrt{\frac{\tau_o}{n}}} + \sqrt{\frac{E(\tau|\mathbf{y})}{\tau_o}} \left(\frac{1}{\sqrt{1-w_\mu}} \right) \sqrt{\frac{n_\tau+n}{n_\tau+n+2}} t_{n_\tau+n+2, 1-\alpha/2} \right] \\ & - E_{\mathbf{y}^\Phi} \left[\left(\frac{w_\mu}{1-w_\mu} \right) \frac{|\Delta_\mu|}{\sqrt{\frac{\tau_o}{n}}} - \sqrt{\frac{E(\tau|\mathbf{y})}{\tau_o}} \left(\frac{1}{\sqrt{1-w_\mu}} \right) \sqrt{\frac{n_\tau+n}{n_\tau+n+2}} t_{n_\tau+n+2, 1-\alpha/2} \right]. \end{aligned} \quad (26)$$

If $w_\mu = w_\tau = 0$ and n is small, then true coverage is slightly below nominal coverage because, although $E(\mu|\mathbf{y}, w_\mu = w_\tau = 0) = \bar{y}$, we also have $E(\tau|\mathbf{y}, w_\mu = w_\tau = 0) = R(\bar{y})/n$, which has a small negative bias as an estimate of τ_o , and the credible interval is

$E(\mu|\mathbf{y}) \pm t_{n+2, 1-\alpha/2} \sqrt{E(\tau|\mathbf{y})/(n+2)}$, which is slightly shorter than a confidence interval.

As w_μ increases between 0 and w_μ^* , the absolute value of the bias of $E(\tau|\mathbf{y})$ decreases. When $w_\mu = w_\mu^*$ the bias of $E(\tau|\mathbf{y})$ is zero. As w_μ increases above w_μ^* , the (positive) bias of $E(\tau|\mathbf{y})$ increases. Therefore, when $w_\tau = 0$, coverage in the unknown variance model is slightly below $1 - \alpha$ for small n and small w_μ and approaches coverage in the known variance model as n or w_μ increases. Figure 5 shows the coverage of a 95% credible interval when $n = 20$ and $w_\tau = 0$. (Compare Figure 5 with Figure 1(b) for the known variance model.)

If n is large and $w_\mu \leq w_\mu^*$, then the true coverage is approximately equal to

$$\Phi \left[\left(\frac{w_\mu}{1-w_\mu} \right) \frac{|\Delta_\mu|}{\sqrt{\frac{\tau_o}{n}}} + \sqrt{1+w_\tau} \left(\frac{\Delta_\tau}{\tau_o} \right) \left(\frac{Z_{1-\alpha/2}}{\sqrt{1-w_\mu}} \right) \right] - \Phi \left[\left(\frac{w_\mu}{1-w_\mu} \right) \frac{|\Delta_\mu|}{\sqrt{\frac{\tau_o}{n}}} - \sqrt{1+w_\tau} \left(\frac{\Delta_\tau}{\tau_o} \right) \left(\frac{Z_{1-\alpha/2}}{\sqrt{1-w_\mu}} \right) \right], \quad (27)$$

which approaches the true coverage in the known variance model when $w_\tau \rightarrow \tau$ approaches 0. Equation (27) shows that as w_τ increases, the true coverage increases if $\tau > 0$ and decreases if $\tau < 0$.

4 Effects of hyperparameter values: An Example

Using estimation of mean body temperature in healthy adults as an example, we demonstrate how our results can be used to evaluate the potential consequences of hyperparameter choices in the unknown variance model. In a study published in 1992 in the Journal of the American Medical Association, the average temperature of a sample of 148 healthy adults

aged 18 to 40 was 98.2°F (36.8°C) (Mackowiak et al., 1992). The authors argued that the standard value for normal body temperature should be changed from the commonly accepted 98.6°F (36.8°C) to 98.2°F (37°C). Although these results and conclusion were controversial and the study received considerable attention, we will not address questions about the validity of the study or the authors' conclusions. The published study had 148 subjects. However, we will assume the sample size n is 20 to better illustrate the effect of the prior when the sample size is small. The study had multiple measurements for each subject and the paper does not contain any Bayesian methods. However, we will assume only one measurement per subject, that the goal is to estimate body temperature using the unknown variance model, and that we have a subject expert who has chosen prior mean $\mu_m = 98.6$ and prior expected variance $\tau_m = 1$, which has been intentionally selected such that it is likely to be larger than the true variance τ_o . Please note that we give this example only to illustrate the statistical methods of calibrating the prior distribution and not to perform the full scope of data analysis. Hence, we chose to ignore many factors such as gender, time of day, etc., that one would consider if this were a real analysis for analyzing body temperature data.

Figure 6 shows the effect of w_μ and w_τ on the posterior point estimate of μ and the 95% credible interval assuming $\mu_m = 98.6$, $\tau_m = 1$, $\bar{y} = 98.2$ and $R(\bar{y})/n = 0.3$. Our goal is to help our subject expert understand the consequences of their choices for w_μ and w_τ by examining their effects on the posterior over a range of values for w_μ and w_τ . Setting hyperparameters n_μ , a , and b as functions of choices for w_μ and w_τ means that the hyperparameters will depend on the sample size n . This would be inappropriate if n were random. However, in the models given in sections 2 and 3, n is known ahead of time and therefore is not random.

4.1 Effects of w_μ

The MSE of $E(\mu|y)$ will be less than the MSE of $\bar{y}$ if $w_\mu < 2w_\mu^*$. However, true coverage of credible intervals decreases rapidly as w_μ increases above w_μ^* . Therefore, assuming we want both a point estimate and an interval estimate, we want $w_\mu \leq w_\mu^*$. For the original model parameterization, this means that we should try to choose an n_μ that is preferably close to, but definitely not larger than the MSE optimal value n_μ^* , which is $(|\Delta_\mu|/\sqrt{\tau_o})^{-2}$. The quantity $|\Delta_\mu|/\sqrt{\tau_o}$ is the number of standard deviations between the prior mean μ_m and the true mean μ_o . We could set $n_\mu = \delta^{-2}$, where δ is the smallest value such that our subject expert is reasonably confident is larger than $|\Delta_\mu|/\sqrt{\tau_o}$, which would give $n_\mu < n_\mu^*$ as desired.

One may want different degrees of confidence for different situations. For example, suppose our subject expert wants a conservative prior and is very confident that $|\Delta_\mu|/\sqrt{\tau_o} < 2$, then we can set $n_\mu = 2^{-2} = 0.25$ ($w_\mu \approx 0.01$). For a less conservative prior, if our subject expert is comfortable assuming that $|\Delta_\mu|/\sqrt{\tau_o} < 0.5$, then we can set $n_\mu = 0.5^{-2} = 4$ ($w_\mu \approx 0.17$).

Although setting n_μ^* by choosing a conservative value for $|\Delta_\mu|/\sqrt{\tau_o}$ is conceptually appealing, it may be easier to select a conservative value for $|\mu|$. With this approach we need a value for τ_o to plug in to get δ . We want δ to be less than the true value $|\Delta_\mu|/\sqrt{\tau_o}$. Therefore, a conservative value for τ_o is one that is smaller than the true value. Since $\tau_m = 1$

was intentionally chosen to be likely to be larger than τ_o , we should plug in a smaller value, say $\tau_o = 0.4$.

We can also help our subject expert evaluate an n_μ they have chosen using their own preferred method. For example, suppose they are considering setting $n_\mu = 0.5$ ($w_\mu \approx 0.02$), which they may view as conservative. We can quantify the actual conservativeness in terms of the maximum value of $|\Delta_\mu|$ such that there will be a reduction in the MSE of the posterior point estimate. Assuming $\tau_o = 0.4$, if $n_\mu = 0.5$, then $w_\mu < w_\mu^*$ if $|\Delta_\mu| < \sqrt{\tau_o/n_\mu} \approx 0.89$ and $w_\mu < 2w_\mu^*$ if $|\Delta_\mu| < \sqrt{2(\tau_o/n_\mu) + \tau_o/n} \approx 1.3$. Telling our subject expert the implicit assumption they are making will help them decide if $n_\mu = 0.5$ is a good choice.

The above approach is straightforward, but provides no information about the potential effects of w_μ on the magnitude of the MSE of $E(\mu|y)$ and the true coverage of credible intervals. Without this information one may choose a prior that is too conservative. We recommend selecting a set of candidate values for w_μ and examining their effects over a range of plausible $|\Delta_\mu|/\sqrt{\tau_o/n}$ values. Figure 7 illustrates how to do this for $w_\mu \in \{0.1, 0.2, 0.3, 0.4\}$ and $|\Delta_\mu|/\sqrt{\tau_o/n} \in [0, 3]$. (Values of w_μ and $|\Delta_\mu|/\sqrt{\tau_o/n}$ were chosen purely for illustration. In practice one should select a range of $|\Delta_\mu|/\sqrt{\tau_o/n}$ values highly likely to contain the true value.) Figure 7(a) shows the ratio of the MSE of $E(\mu|y)$ to the MSE of $\bar{y}$. Figure 7(b) shows the true coverage of a 95% credible interval when $w_\tau = 0$ ($a = 1, b = 0$). Figure 7(c) shows the ratio of the MSE of $E(\mu|y)$ to the MSE of $\bar{y}$ and the coverage of 95% credible intervals when $w_\mu = w_\mu^*$. Using Figure 7 one can make a fully informed choice for w_μ .

Given the potential effects of w_μ on the posterior shown in Figure 7, a reasonably conservative choice for estimation of body temperature might be $w_\mu = 0.1$, because with this choice, if $|\Delta_\mu|/\sqrt{\tau_o/n} < 3$ or (assuming $\tau_o = 0.4$) if $|\Delta_\mu| < 0.42$, then $\text{MSE}[E(\mu|y)] < \text{MSE}[\bar{y}]$. If $w_\mu = 0.1$, then $E(\mu|y) = 98.24$.

4.2 Effects of w_τ

In this section we discuss the effect of w_τ on the posterior of μ . The value of w_τ does not affect $E(\mu|y)$, but does affect the length and therefore coverage of credible intervals for μ . As can be seen in Figure 6, as w_τ increases, the length of the 95% credible interval increases. This occurs because $\tau_m > R(\bar{y})/n$ and $\text{var}(\mu|y) = (1 - w_\mu)E(\tau|y)/n$, where $E(\tau|y)$ is a weighted average of τ_m and $Z(w_\mu) = w_\mu R(\bar{y})/n + (1 - w_\mu)R(\mu_m)/n$, with weights proportional to w_τ and $1 - w_\tau$, respectively. We need to be cautious not to set w_τ too large, because if τ_m had been set too small, we would have $\tau_m < R(\bar{y})/n$ and then the length of credible intervals would decrease as w_τ increases.

Assuming $w_\mu < w_\mu^*$, then if $w_\tau = 0$ and n were large, coverage would be approximately equal to or above the nominal value, and we would not want to change interval length by setting $w_\tau > 0$. Therefore, when n is large, unless there is very good reason to believe that τ_m is very

close to the true variance τ_o , in general we recommend setting $w_\tau = 0$ or very small. However, for the body temperature example, w_τ should be set larger than 0 because the sample size is small ($n = 20$) which results in a small reduction of coverage below that of the known variance model when $w_\tau = 0$. (See figure 5.) Assuming we are confident that $\tau_m > \tau_o$, we could set $w_\tau = 0.1$, which with $w_\mu = 0.1$, gives a 95% credible interval of (98.0, 98.5).

5 Discussion

For any Bayesian model, selection of the prior should be informed by a full understanding of the prior's effects on posterior point and interval estimates. In most cases, as a prior becomes more informative there is an increase in bias and a decrease in variance of the posterior point estimate. Therefore an informative prior has the potential to increase or decrease the quality of estimation. However, if one fully understands the effects of the prior on the bias, variance, MSE of the posterior point estimate, and the coverage and length of the credible interval, one can choose a conservative prior that has a high likelihood to improve the quality of estimation.

For the normal conjugate models in sections 2 and 3, when $w_\mu \leq w_\mu^*$ there are multiple benefits in both the known and unknown variance models. In both models, if $w_\mu = w_\mu^*$ then $E(\mu|y)$ has minimum MSE and if $w_\mu < w_\mu^*$ then the MSE of $E(\mu|y)$ is smaller than the MSE of $\bar{y}$. In the known variance model, if $w_\mu \leq w_\mu^*$, then the posterior variance $\text{var}(\mu|y)$ is equal to or larger than the MSE of $E(\mu|y)$ (see equation (11)) and the true coverage of credible intervals is approximately equal to or above nominal coverage. In the unknown variance model, if $w_\mu \leq w_\mu^*$, n is moderately large, $a = 1$, and $b = 0$, then true coverage of credible intervals is approximately equal to or above nominal coverage. An additional benefit to having $w_\mu < w_\mu^*$ in the unknown variance model is that if $\tau > 0$ then the value of w_μ that minimizes the MSE of $E(\tau|y)$ is less than w_μ^* .

Having $w_\mu > w_\mu^*$ is especially problematic for the credible interval. As w_μ increases above w_μ^* , credible interval coverage decreases rapidly, because the MSE of $E(\mu|y)$ increases and $\text{var}(\mu|y)$ decreases, which decreases the width of the credible interval. This result would also apply to any other model where weighting too strongly towards the prior increases the MSE and decreases posterior variance.

As demonstrated in the example in Section 4, our results can be used to inform the selection of hyperparameters by evaluating their potential effects on the posterior. We recommend that w_μ be selected such that there is strong belief that $w_\mu < w_\mu^*$. When setting hyperparameters a and b in the unknown variance model, one should primarily consider their effects on the coverage of credible intervals for μ . These hyperparameters have no effect on $E(\mu|y)$, but they affect $\text{var}(\mu|y)$ and consequently credible interval length and coverage. If $w_\mu < w_\mu^*$, $a = 1$, and $b = 0$, then for moderately large n , the credible interval length and coverage will be approximately the same as in the known variance model. Therefore, unless there is good reason to believe that τ_m is very close to τ_o , we recommend setting $w_\tau = 0$ or very small. If $w_\tau = 0$, then $a = 1$, which using the inverse function given in section 3.2 gives $b = 0$. One can

set $a = 1$ and then either set b very small or set $b = 0$ giving the improper prior $\tau \propto \tau^{-2}$, which still results in a proper posterior distribution.

Our results demonstrate some general long-run frequentist properties of posterior point estimates and credible intervals that are likely to extend to many Bayesian models. Specifically they are most directly applicable to cases where the posterior mean can be written as a weighted average of the prior mean and the sample mean, with weights that depend on the sample size and a prior variance hyperparameter. As discussed in Section 1, this is the case when the likelihood is a member of the exponential family and the prior is conjugate (Ericson (1970), Brown (1986)). The exact functional forms of the MSE of posterior point estimates and the coverage of credible intervals will depend on the specific distributions. However, we would expect that the general shapes of their dependencies on both the weight and the bias of the prior mean are likely to be similar to the graphs shown in figures 1–5 and 7.

A well recognized problem in eliciting priors is that people are often overconfident about their beliefs and weight too strongly towards the prior. Our results suggest that, in general, some degree of overconfidence is tolerable for the posterior point estimate, but even a small amount of overconfidence can result in the true coverage of posterior credible intervals falling below the nominal value. For any model, regardless of the method used to select hyperparameters, one should at least be aware that they are making an implicit assumption about the size of the bias of the prior point estimate and understand the consequences for the posterior if their assumption is incorrect.

Acknowledgments

This work was supported in part by grants CA 16672 and CA 97007 from the National Cancer Institute and W81XWH-06-1-303 from the Department of Defense. The authors also thank the reviewers for constructive critiques and the editorial assistance from Ms. LeeAnn Chastain.

References

- Bayarri MJ, Berger JO. The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*. 2004; 19:58–80.
- Bernardo JM. Non-informative priors do not exist A dialogue with Jos M. Bernardo. *Journal of Statistical Planning and Inference*. 1997; 65:159177.
- Bernardo JM, Irony TZ, Singpurwalla ND. Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1979; 41:113–147.
- Brown, LD. Monograph Series. Vol. 6. Hayward, California: IMS lecture notes; 1986. Foundations of Exponential Families.
- Dawid AP. The Well-Calibrated Bayesian. *Journal of the American Statistical Association*. 1982; 77:605–610.
- Delampady M, DasGupta A, Casella G, Rubin H, Strawderman WE. A New Approach to Default Priors and Robust Bayes Methodology. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*. 2001; 29:437–450.
- Ericson WA. On the Posterior Mean and Variance of a Population Mean. *Journal of the American Statistical Association*. 1970; 65:649–652.
- Garthwaite PH, Kadane JB, O’Hagen A. Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association*. 2005; 100:680–700.
- Gelman, A., Carlin, JB., Stern, HS., Rubin, D. second. Chapman & Hall; 2004.

- Kadane JB, Dickey JM, Winkler RL, Smith WS, Peters SC. Interactive Elicitation of Opinion for a Normal Linear Model. *Journal of the American Statistical Association*. 1980; 75:845–854.
- Kadane JB, Wolfson LJ. Experiences in Elicitation. *The Statistician*. 1998; 47:3–19.
- Kass RE. Approximate Methods for Assessing Influence and Sensitivity in Bayesian Analysis. *Biometrika*. 1989; 76:663–674.
- Kass RE, Wasserman L. The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*. 1996; 91:1343–1370.
- Little R. Calibrated Bayes: A Bayes/Frequentist Roadmap. *The American Statistician*. 2006; 60:1–11.
- Mackowiak PA, Wasserman SS, Levine MM. A Critical Appraisal of 98.6°F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich. *Journal of the American Medical Association*. 1992; 268:1578–1580. [PubMed: 1302471]
- McCulloch RE. Local Model Influence. *Journal of the American Statistical Association*. 1989; 84:473–478.
- Millar RB, Stewart WS. Assessment of Locally Influential Observations in Bayesian Models. *Bayesian Analysis*. 2007; 2:365–384.
- Ren C, Sun D, Dey DK. Comparison of Bayesian and Frequentist Estimation and Prediction for a Normal Population. *Sankhyā : The Indian Journal of Statistics*. 2004; 66:678–706.
- Robert, CP. *The Bayesian Choice*. second. Springer; 2001.
- Rubin DB. Bayesian Justifiability and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*. 1984; 12:1151–1172.
- Weiss RE. An Approach to Bayesian Sensitivity Analysis. *Journal of the Royal Statistical Society B*. 1996; 58:739–750.

A Appendix

A.1 Derivation of equation (13) in section 2.2

True coverage in the known variance model, given by equation (12) in section 2.2, can be written as

$$C_{\alpha}(w_{\mu}) = \Phi(Z_g) - \Phi(Z_h)$$

where

$$Z_g = \frac{|\text{bias}[E(\mu|\mathbf{y})]|}{\text{SD}_{\mathbf{y}}[E(\mu|\mathbf{y})]} + Z_{1-\alpha/2} \frac{\text{SD}(\mu|\mathbf{y})}{\text{SD}_{\mathbf{y}}[E(\mu|\mathbf{y})]}$$

and

$$Z_h = \frac{|\text{bias}[E(\mu|\mathbf{y})]|}{\text{SD}_{\mathbf{y}}[E(\mu|\mathbf{y})]} - Z_{1-\alpha/2} \frac{\text{SD}(\mu|\mathbf{y})}{\text{SD}_{\mathbf{y}}[E(\mu|\mathbf{y})]},$$

where $\Phi(\cdot)$ is the distribution function and Z_p is the p quantile of a standard normal variable,

$\text{SD}(\mu|\mathbf{y}) = \sqrt{\text{var}(\mu|\mathbf{y})} = \sqrt{(1 - w_{\mu})(\tau_o/n)}$ and

$\text{SD}_{\mathbf{y}}[E(\mu|\mathbf{y})] = \sqrt{\text{var}_{\mathbf{y}}[E(\mu|\mathbf{y})]} = (1 - w_{\mu}) \sqrt{\tau_o/n}$. We have $C_{\alpha}(w_{\mu}) = 1 - \alpha$ if and only if $h = g - (1 - \alpha)$. We know $Z_{g-(1-\alpha)} < Z_{\alpha}$ because $g-(1-\alpha) < \alpha$ since $g < 1$. We also know $Z_{g-(1-\alpha)} > Z_{\alpha/2}$ because $Z_g > Z_{1-\alpha/2}$, which follows from the fact that $|\text{bias}[E(\mu|\mathbf{y})]| \neq 0$ and

$SD(\mu|y)/SD_y[E(\mu|y)] > 1$. Therefore, when $C_\alpha(w_\mu) = 1 - \alpha$, we have $Z_{\alpha/2} < Z_h < Z_\alpha$. Since coverage increases as Z_h decreases, $Z_h < Z_{\alpha/2}$ is a sufficient condition for $C_\alpha(w_\mu) > 1 - \alpha$ and $Z_h > Z_\alpha$ is a sufficient condition for $C_\alpha(w_\mu) < 1 - \alpha$. These conditions can be written as

$$w_\mu|\Delta_\mu| = |\text{bias}[E(\mu|y)]| < Z_{1-\alpha/2} \{SD(\mu|y) - SD_y[E(\mu|y)]\}$$

is sufficient for $C_\alpha(w_\mu) > (1 - \alpha)$ and

$$w_\mu|\Delta_\mu| = |\text{bias}[E(\mu|y)]| > Z_{1-\alpha/2}SD(\mu|y) - Z_{1-\alpha}SD_y[E(\mu|y)]$$

is sufficient for $C_\alpha(w_\mu) < (1 - \alpha)$.

A.2 General results for weighted averages

The effects of w_μ and w_τ on the MSEs of the posterior means of μ and τ , respectively, can be generalized to any estimator (Bayesian or Frequentist) that is a weighted average. Since many posterior means are weighted averages, these general results can be used for many different Bayesian models. Consider estimation of a parameter β , where the estimator $\hat{\beta}(w)$ is

$$\hat{\beta}(w) = w\hat{\beta}_1 + (1 - w)\hat{\beta}_2,$$

where $w \in [0, 1]$ and $\hat{\beta}_1$ and $\hat{\beta}_2$ are functions of either fixed parameters and/or the observed data. The MSE of $\hat{\beta}(w)$ is

$$\text{MSE}[\hat{\beta}(w)] = w^2\text{MSE}(\hat{\beta}_1) + (1 - w)^2\text{MSE}(\hat{\beta}_2) + 2w(1 - w)e_{12}, \quad (28)$$

where

$$e_{12} = \text{cov}(\hat{\beta}_1, \hat{\beta}_2) + \Delta_1\Delta_2$$

and

$$\Delta_i = \text{bias}(\hat{\beta}_i) = E(\hat{\beta}_i) - \beta \text{ for } i=1, 2.$$

All MSEs, covariances, and expectations are taken with respect to the distribution of the data. Equation (28) shows that the MSE of $\hat{\beta}(w)$ is smaller when the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ is negative and/or the biases Δ_1 and Δ_2 are in opposite directions.

The values of w that minimizes the MSE of $\hat{\beta}(w)$ is

$$w^* = \begin{cases} 0 & \text{if } \text{MSE}(\hat{\beta}_2) \leq e_{12} < \text{MSE}(\hat{\beta}_1); \\ \frac{\text{MSE}(\hat{\beta}_2) - e_{12}}{\text{MSE}(\hat{\beta}_1) + \text{MSE}(\hat{\beta}_2) - 2e_{12}} & \text{if } e_{12} < \min \{ \text{MSE}(\hat{\beta}_1), \text{MSE}(\hat{\beta}_2) \}; \\ 1 & \text{if } \text{MSE}(\hat{\beta}_1) \leq e_{12} < \text{MSE}(\hat{\beta}_2) \end{cases}$$

The three conditions are mutually exclusive because

$$\text{MSE}(\hat{\beta}_1) + \text{MSE}(\hat{\beta}_2) - 2e_{12} = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) + (\Delta_1 - \Delta_2)^2 > 0 \text{ and therefore } e_{12} < \max \{ \text{MSE}(\hat{\beta}_1), \text{MSE}(\hat{\beta}_2) \}.$$

If $w^* \notin \{0, 1\}$ then the MSE of $\hat{\beta}(w)$ is symmetric around w^* and can be written as either

$$\text{MSE}[\hat{\beta}(w)] = (1 - w^*)\text{MSE}(\hat{\beta}_2) + (w - w^*)^2 [\text{MSE}(\hat{\beta}_1) + \text{MSE}(\hat{\beta}_2) - 2e_{12}] + w^*e_{12}$$

or

$$\text{MSE}[\hat{\beta}(w)] = w^*\text{MSE}(\hat{\beta}_1) + (w - w^*)^2 [\text{MSE}(\hat{\beta}_1) + \text{MSE}(\hat{\beta}_2) - 2e_{12}] + (1 - w^*)e_{12}.$$

Therefore

$$\text{MSE}[\hat{\beta}(w)] \begin{cases} < \text{MSE}(\hat{\beta}_2) & \text{if } 0 < w < \min\{2w^*, 1\}; \\ = \text{MSE}(\hat{\beta}_2) & \text{if } w = 0 \text{ or } w = \min\{2w^*, 1\}; \\ > \text{MSE}(\hat{\beta}_2) & \text{if } 2w^* < w \leq 1. \end{cases}$$

In a Bayesian model where $\hat{\beta}(w)$ is the posterior expected value and $\hat{\beta}_1$ is a prior mean, $\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = 0$ and e_{12} is reduced to $\frac{1}{2}$. If in addition $\hat{\beta}_2$ is an unbiased estimate then $\frac{1}{2} = 0$ and e_{12} is reduced to zero. This is the case for the posterior expected value of the mean in both the known and unknown variance models in sections 2 and 3.

A.2.1 Difference between two normal means

The above results can easily be applied to any posterior estimate that is a weighted average. As an example, consider the posterior expected value of the difference between two means μ_1 and μ_2 in the normal conjugate model where for $i = 1, 2$ and $j = 1, 2, \dots, n$, y_{ij} is normally distributed with mean μ_j and variance τ and the prior for μ_j is normal with mean μ_m and variance τ/n_j . Results for the MSE of the posterior mean of the difference do not depend on the prior for τ , which is therefore left unspecified as distribution F . The model is

$$y_{ij} | \mu_i, \tau \sim N(\mu_i, \tau) \quad i = 1, 2, j = 1, \dots, n$$

$$\mu_i|\tau \sim N\left(\mu_m, \frac{\tau}{n_\mu}\right) i=1, 2$$

$$\tau \sim F.$$

The posterior expected value of $\mu_1 - \mu_2$ is

$$E(\mu_1 - \mu_2|\mathbf{y}) = (1 - w)(\bar{y}_1 - \bar{y}_2) \quad (29)$$

where

$$w = \frac{n_\mu}{n_\mu + n}.$$

Equation (29) is a weighted average of the prior expected difference $E(\mu_1 - \mu_2) = E(\mu_m - \mu_m) = 0$ and $\bar{y}_1 - \bar{y}_2$. Since $\bar{y}_1 - \bar{y}_2$ is an unbiased estimate of $\mu_1 - \mu_2$ and is uncorrelated with the prior expected difference, $e_{12} = 0$. Assuming that for $i = 1, 2$ and $j = 1, \dots, n$, y_{ij} is distributed normally with mean μ_{0i} and variance τ_o , the MSE of $E(\mu_1 - \mu_2)$ is $(\mu_{01} - \mu_{02})^2$, the MSE of $(\bar{y}_1 - \bar{y}_2)$ is $2\tau_o/n$, the MSE of $E(\mu_1 - \mu_2|\mathbf{y})$ is

$$\text{MSE}[E(\mu_1 - \mu_2|\mathbf{y})] = w^2(\mu_{01} - \mu_{02})^2 + (1 - w)^2 \frac{2\tau_o}{n},$$

and the MSE optimal w is

$$w^* = \frac{\text{MSE}(\bar{y}_1 - \bar{y}_2)}{\text{MSE}[E(\mu_1 - \mu_2)] + \text{MSE}(\bar{y}_1 - \bar{y}_2)} = \frac{\frac{2\tau_o}{n}}{(\mu_{01} - \mu_{02})^2 + \frac{2\tau_o}{n}}.$$

The MSE of $E(\mu_1 - \mu_2|\mathbf{y})$ is symmetric around w^* . If $w < w^*$ then $\text{MSE}[E(\mu_1 - \mu_2|\mathbf{y})] < \text{MSE}(\bar{y}_1 - \bar{y}_2)$.

Figure 1. Effect of w_μ in both the known and unknown variance models on (a) the ratio of the MSE of $E(\mu|y)$ to the MSE of $\bar{y}$. Effect of w_μ in the known variance models on (b) the coverage of a 95% credible interval and (c) the ratio of the length of a 95% credible interval to the length of a 95% confidence interval. The short vertical bar is used to denote w_μ^* , the MSE optimal w_μ .

Figure 2.

Credible interval coverage for the known variance model with $|\Delta_\mu|/\sqrt{\tau_0/n}=3(w_\mu^*=0.1)$.

Figure 3. Effect of the standardized bias on the MSE optimal (w_μ^*) and credible interval optimal ($w_\mu^\dagger(\alpha)$) values of w_μ in the known variance model.

Figure 4. Effect of w_τ and w_μ on the ratio of the MSE of $E(\tau|\mathbf{y})$ to the MSE of $R(\bar{y})/(n-1)$ in the unknown variance model with $n=20$ and $w_\mu^*=0.2$. The MSE of $R(\bar{y})/(n-1)$ is $(2\tau_o^2)/n$. (See also equation (23).)

Figure 5. Ninety-five percent credible interval coverage in the unknown variance model when $n = 20$ and $w_\tau = 0$ ($a = 1$, $b = 0$). Coverage values were estimated using 40,000 simulated values of $R(\mu)$ and $R(\bar{y})$, which are sufficient statistics for $E(\tau|\mathbf{y})$. The triangles are values of w_μ that are MSE optimal for estimation of μ_0 . The circles are values of w_μ that are MSE optimal for estimation of τ_0 .

Figure 6. Posterior mean and 95% credible interval for the body temperature example in section 4 with $n = 20$, $\mu_m = 98.6$, $\tau_m = 1$, $\bar{y} = 98.2$ and $R(\bar{y})/n = 0.3$. The solid line is $E(\mu|y)$ and the dashed lines are 95% credible intervals.

Figure 7.

For the unknown variance model with $n = 20$ and $w_\tau = 0$ ($a = 1$, $b = 0$), the effect of the standardized bias ($|\Delta_\mu|/\sqrt{\tau_0/n}$) on the ratio of the MSE of $E(\mu|y)$ to the MSE of $\bar{y}$ and on the true coverage of a 95% credible interval, where (a) and (b) are for $w_\mu \in \{0.1, 0.2, 0.3, 0.4\}$ and (c) is for $w_\mu = w_\mu^*$, where the solid line corresponds to the left axis (MSE ratio) and the dashed line corresponds to the right axis (credible interval coverage). In (b) credible

interval coverage was estimated using 40,000 simulated values of $R(\mu)$ and $R(\bar{y})$, which are the sufficient statistics for $E(\tau|\mathbf{y})$

Table 1

Expectation, bias, variance, and mean squared error (MSE) for each term in $E(\tau|y)$ (See equation (19)). The covariance between $R(\mu_m)/n$ and $R(\bar{y})/n$ is equal to the variance of $R(\bar{y})/n$.

	expectation	bias	variance	MSE
τ_m	τ_m	τ	0	Δ_τ^2
$\frac{R(\mu_m)}{n} = \frac{1}{n} \sum_{i=1}^n (y_i - \mu_m)^2$	$\tau_\circ + \Delta_\mu^2$	Δ_μ^2	$\frac{2\tau_\circ}{n} (\tau_\circ + 2\Delta_\mu^2)$	$\Delta_\mu^4 + \frac{2\tau_\circ}{n} (\tau_\circ + 2\Delta_\mu^2)$
$\frac{R(\bar{y})}{n} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$	$\tau_\circ \left(\frac{n-1}{n} \right)$	$\frac{-\tau_\circ}{n}$	$2\tau_\circ^2 \left(\frac{n-1}{n^2} \right)$	$\tau_\circ^2 \left(\frac{2n-1}{n^2} \right)$

Table 2

Effect of w_μ on the bias of $\mathbf{Z}(w_\mu) = w_\mu R(\mu_m)/n + (1 - w_\mu)R(\bar{y})/n$ and the bias of $E(\tau|\mathbf{y})$.

$w_\mu < w_\mu^*$	$w_\mu = w_\mu^*$	$w_\mu > w_\mu^*$
$-(1 - w_\mu) \frac{\tau_o}{n} < \text{bias} \left[\bar{\Sigma}(w_\mu) \right] < 0$ $w_\tau \Delta_\tau - (1 - w_\tau)(1 - w_\mu) \left(\frac{\tau_o}{n} \right) < \text{bias} [E(\tau \mathbf{y})] < w_\tau \Delta_\tau$	$\text{bias} [\mathbf{Z}(w_\mu)] = 0$ $\text{bias} [E(\tau \mathbf{y})] = \frac{w_\tau \tau}{w_\tau}$	$0 < \text{bias} \left[\bar{\Sigma}(w_\mu) \right] < w_\mu \Delta_\mu^2$ $w_\tau \Delta_\tau < \text{bias} [E(\tau \mathbf{y})] < w_\tau \Delta_\tau + (1 - w_\tau) w_\mu \Delta_\mu^2$