1 Bayesian Bias Correction Model

Assuming that n iid samples $\{X_1, \ldots, X_n\}$, were collected from a normal population with mean μ and variance σ^2. The model likelihood has the form,

$$P(\vec{X}|\mu, \sigma^2, T_n > c) = \prod_{i=1}^{n} \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(X_i-\mu)^2}{2\sigma^2}\right]}{1 - \Phi(c - \frac{\mu}{\sigma/\sqrt{n}})} \tag{1.1}$$

where $\vec{X} = (X_1, \ldots, X_n)$ and Φ is the cumulative distribution function (cdf) of the standard normal distribution.

The prior distribution of the model is defined by the following distributions:

$$\begin{aligned} p(\mu|\xi) &= \xi\delta_{\{0\}}(\mu) + (1-\xi)f(\mu), \\ f(\mu) &= \text{Uniform}(0, A), \\ p(\xi) &= \text{Beta}(a, b), \\ p(\sigma^2) &= \text{Inv-Gamma}(\alpha_1, \alpha_2), \end{aligned}$$

where A is the upper bound of log OR. We use $A = 2$ throughout the paper. We choose the shape parameter, α_1, and the scale, α_2, for the inverse gamma distribution such that the prior mean of σ^2 is equal to the sample variance, S^2, and the prior variance of σ^2 is equal to 200. Since the mean of the Inv-Gamma(α_1, α_2) is $\frac{\alpha_2}{\alpha_1-1}$ for $\alpha_1 > 1$, and the variance is $\frac{\alpha_2^2}{(\alpha_1-1)^2(\alpha_1-2)}$, a simple calculation leads to $\alpha_1 = S^4/200+2$, and $\alpha_2 = S^6/200 + S^2$.

We reparameterize the model using $\theta = \mu/2$ and therefore, the proposed Bayesian model has the following hierarchical structure

$$p(\theta|\xi) = \xi g_0(\theta) + (1-\xi)g_1(\theta), \tag{1.2}$$

$$\begin{aligned} p(\xi) &= \text{Beta}(a, b) \\ p(\sigma^2) &= \text{Inv-Gamma}(S^4/200 + 2, S^6/200 + S^2) \end{aligned}$$

where $g_0(\theta) = \delta_{\{0\}}(\theta)$ and $g_1(\theta)$ is the density of Uniform(0,1) and S is the sample standard deviation.

The joint prior distribution for (θ, ξ) is

$$p(\theta, \xi) = p(\theta|\xi)p(\xi) = \xi g_0(\theta)\xi^{a-1}(1-\xi)^{b-1} + (1-\xi)g_1(\theta)\xi^{a-1}(1-\xi)^{b-1}. \quad (1.3)$$

Let Z be the latent mixture indicator so that $Z = 0$ if the significant SNP is a false positive ($\theta = 0$) and $Z = 1$ for a true positive ($\theta > 0$). Then conditional on Z, the sampling distribution is:

$$p(\vec{X}|\theta, \sigma^2, Z, T_n > c) \propto (1/\sigma)^n \left(\frac{\exp\{-\sum_{i=1}^n \frac{X_i^2}{2\sigma^2}\}}{1-\Phi(c)} \right)^{1-Z} \left(\frac{\exp\{-\sum_{i=1}^n \frac{(X_i-2\theta)^2}{2\sigma^2}\}}{1-\Phi(c-\frac{2\theta}{\sigma/\sqrt{n}})} \right)^{Z} \quad (1.4)$$

If Z were observed, the posterior distribution for the vector (θ, ξ, σ^2) can be expressed as:

$$\begin{aligned}
p(\theta, \xi, \sigma^2|\vec{X}, Z, T_n > c) &\propto p(\vec{X}, Z|\theta, \sigma^2, T_n > c)p(\theta|\xi)p(\xi)p(\sigma^2) \\
&\propto (1/\sigma)^n \left(\frac{\exp\{-\sum_{i=1}^n \frac{X_i^2}{2\sigma^2}\}}{1-\Phi(c)} \right)^{1-Z} \left(\frac{\exp\{-\sum_{i=1}^n \frac{(X_i-2\theta)^2}{2\sigma^2}\}}{1-\Phi(c-\frac{2\theta}{\sigma/\sqrt{n}})} \right)^{Z} \\
&\times (\xi g_0(\theta) + (1-\xi)g_1(\theta)) \times \xi^{a-1}(1-\xi)^{b-1} \times p(\sigma^2) \\
&\propto (1/\sigma)^n \left(\frac{\exp\{-\sum_{i=1}^n \frac{X_i^2}{2\sigma^2}\}}{1-\Phi(c)} \right)^{1-Z} \left(\frac{\exp\{-\sum_{i=1}^n \frac{(X_i-2\theta)^2}{2\sigma^2}\}}{1-\Phi(c-\frac{2\theta}{\sigma/\sqrt{n}})} \right)^{Z} \\
&\times \xi^{1-Z}(1-\xi)^Z \xi^{a-1}(1-\xi)^{b-1} \left(\frac{1}{\sigma^2}\right)^{\alpha_1+1} \exp\left\{-\frac{\alpha_2}{\sigma^2}\right\} \\
&= (1/\sigma)^n \left(\frac{\exp\{-\sum_{i=1}^n \frac{X_i^2}{2\sigma^2}\}\xi}{1-\Phi(c)} \right)^{1-Z} \\
&\times \left(\frac{\exp\{-\sum_{i=1}^n \frac{(X_i-2\theta)^2}{2\sigma^2}\}(1-\xi)}{1-\Phi(c-\frac{2\theta}{\sigma/\sqrt{n}})} \right)^{Z} \\
&\times \xi^{a-1}(1-\xi)^{b-1} \left(\frac{1}{\sigma^2}\right)^{\alpha_1+1} \exp\left\{-\frac{\alpha_2}{\sigma^2}\right\}.
\end{aligned}$$

with $\alpha_1 = S^4/200 + 2$, and $\alpha_2 = S^6/200 + S^2$.

2 Supplementary Plots

We present simulation results under a number of additional scenarios:

- Figure 1 illustrates the performance of the estimators under the null hypothesis $(\mu = 0)$.
- Figure 2 shows results when the type one error rate is 10^{-4}.
- The robustness of the model with respect to prior choice is reflected in Figure 3 where summaries for **B.L**, **B.H** and **B.BMA** are presented for different choices of the parameters a and b.
- Simulation results under an additive genetic model with different values of $\mu \in \{0, \log(1.02), \log(1.1), \log(1.5)\}$ and when the significance level is $\alpha = 0.05$ are shown in Figure 4.
- Similar scenario to the one described above but with $\alpha = 0.001$ (Figure 5).
- Comparison of two different burn-in periods shows that discarding the first 5,000 or 15,000 samples produces very similar results (Figures 6 and 7).
- The robustness of the prior to the choice of the upper bound A for μ and prior variance for σ^2 is illustrated in Figures 8 and 9.

Figure 1: ***Performance of the nine estimators under the normal model with a type I error rate of*** 0.05 ***when the true value of*** μ ***is log(1)=0.*** *Each circle represents an estimate,the horizontal bar is the averaged estimate over 200 simulated datasets.The Bias, sample Standard Deviation(SD) and Root Mean Squared Error (RMSE) are also provided for each estimator. One can see that B.L performs best in this case.*

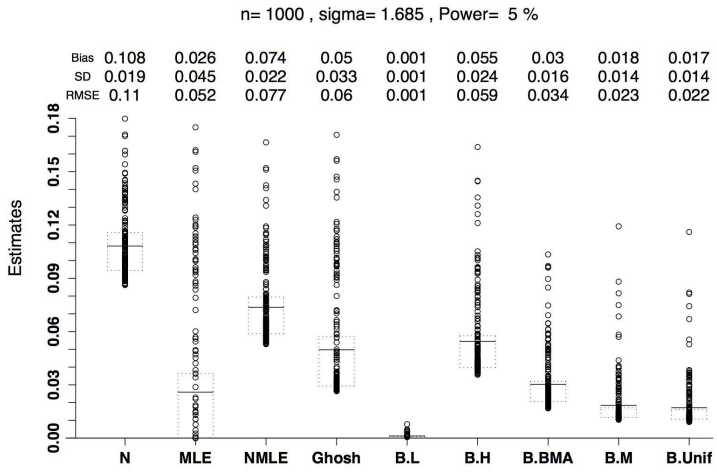

Figure 2: ***Performance of the nine estimators under the normal model with a type I error rate of*** 10^{-4}***.*** *The population mean* $\mu = log(1.1) = 0.0953$ *and power ranging from 5%,20%,50% to 99%. Details of the simulating parameters are given in row 2 of table 1. Each circle represents an estimate, the horizontal bar is the averaged estimate over 200 simulated datasets, and the long horizontal line represents the true value of* μ*. The Bias, sample Standard Deviation(SD) and Root Mean Squared Error (RMSE) are also provided for each estimator.*

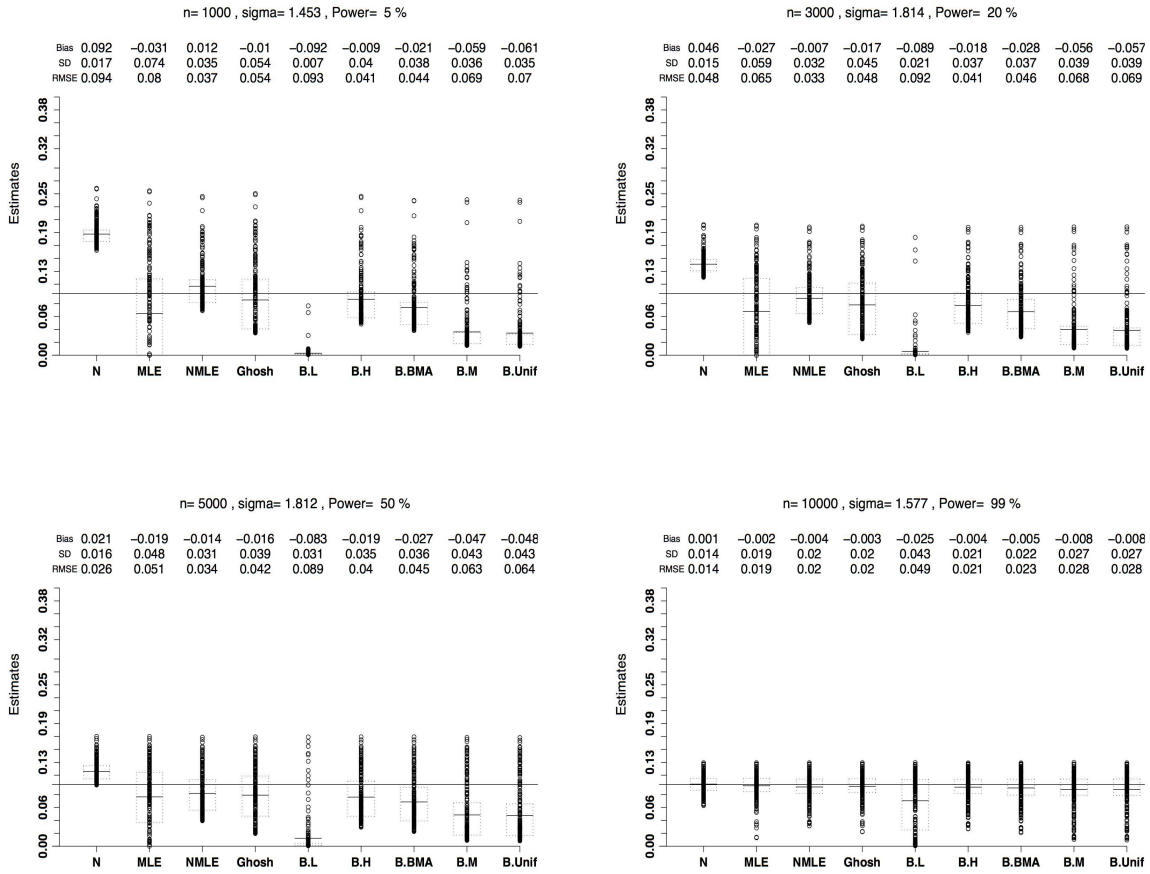

Figure 3: ***Performance of the Bayesian estimators B.L and B.H and the B.BMA averaging over B.L and B.H. for different values settings of a and b in the prior distribution Beta(a,b) for the hyperparameter ξ under the normal model.*** *All estimators with* $(a, b) \in \{(4, 0.5), (8, 0.5), (16, 0.5)\}$ *are of the* **B.L** *type because the density of Beta(a,b) in this case preserves the "inverse" L-shape. Similarly, when* $(a, b) \in \{(0.5, 4), (0.5, 8), (0.5, 16)\}$ *we obtain* **B.H***-type densities that preserve the L-shape. Left: power=10%, type I error* $\alpha = 0.05$*. Right: power=5%, type I error* $\alpha = 10^{-6}$*. Each circle represents an estimate, the horizontal bar is the averaged estimate over 200 simulated datasets, and the long horizontal line represents the true value of* μ*. The Bias, sample Standard Deviation(SD) and Root Mean Squared Error (RMSE) are also provided for each estimator.*

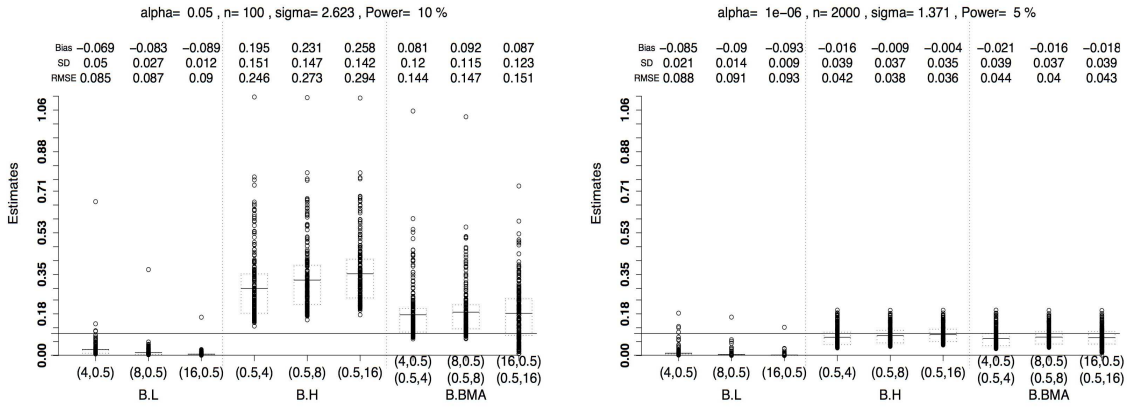

Figure 4: ***Performance of the nine estimators under an additive genetic model with a type I error rate of*** $\alpha = 0.05$***.*** *The sample size is 1,000 (500 cases and 500 controls), the minor allele frequency of the causal SNP is 0.25. The true effects of the SNP on the log OR scale are* $\mu = \beta = log(1) = 0$ *corresponding to the null case, log(1.02) corresponding to power* $\approx 10\%$*, log(1.1) corresponding to power* $\approx 30\%$ *to log(1.5) corresponding power* $> 95\%$*. Each circle represents an estimate, the horizontal bar is the averaged estimate over 200 simulated datasets, and the long horizontal line represents the true value of* μ*. The Bias, sample Standard Deviation(SD) and Root Mean Squared Error (RMSE) are also provided for each estimator.*

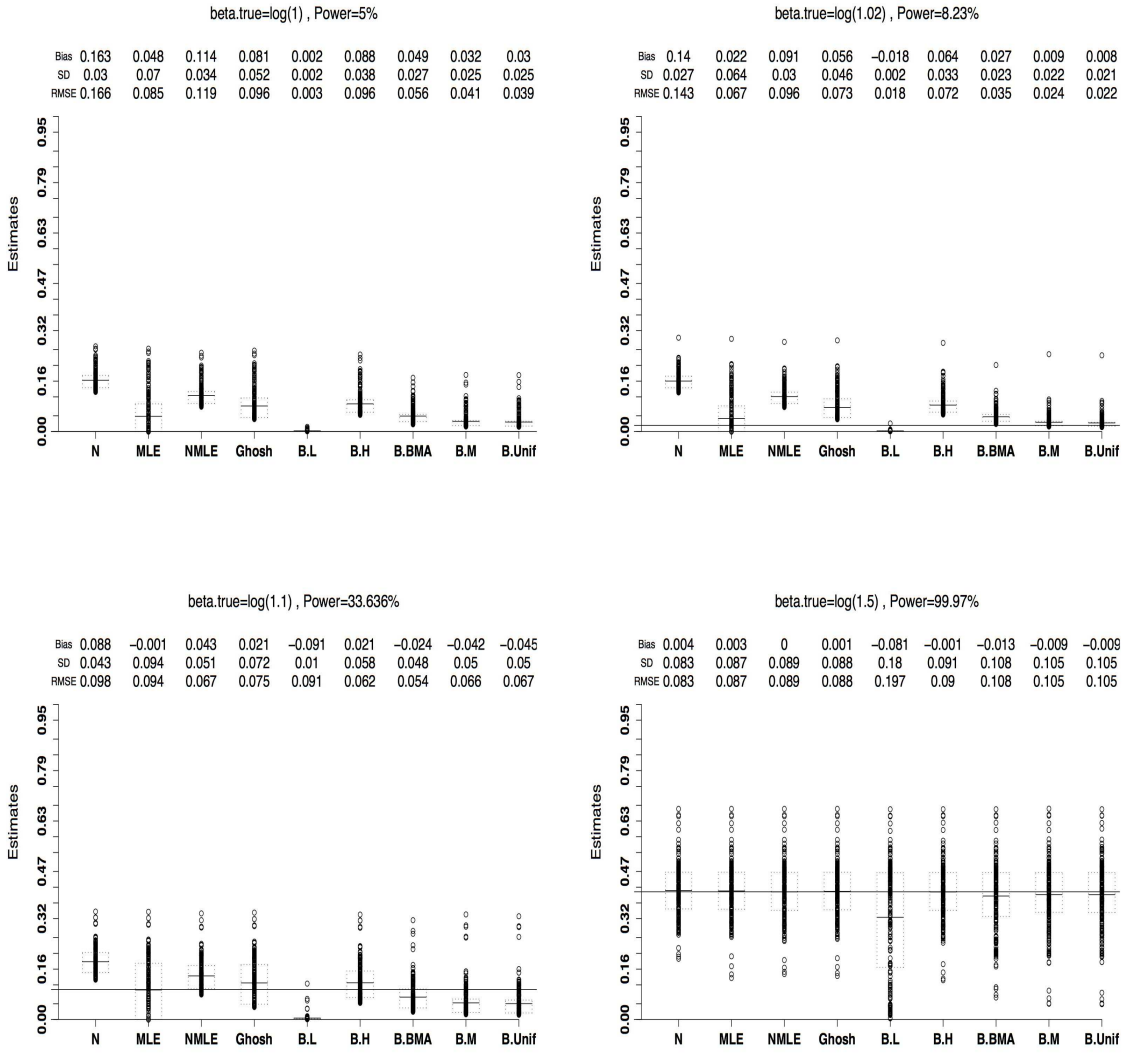

Figure 5: ***Performance of the nine estimators under an additive genetic model with a type I error rate of*** $\alpha = 0.001$***.*** *The sample size is 1,000 (500 cases and 500 controls), the minor allele frequency of the causal SNP is 0.25.The true effect of the SNP on the log OR scale ranging from* $\mu = \beta = log(1.05)$*,log(1.1),log(1.2) to log(1.6) corresponding to power* $< 1\%$, $\approx 5\%$, $\approx 20\%$, $> 95\%$*. Each circle represents an estimate, the horizontal bar is the averaged estimate over 200 simulated datasets, and the long horizontal line represents the true value of* μ*. The Bias, sample Standard Deviation(SD) and Root Mean Squared Error (RMSE) are also provided for each estimator.*

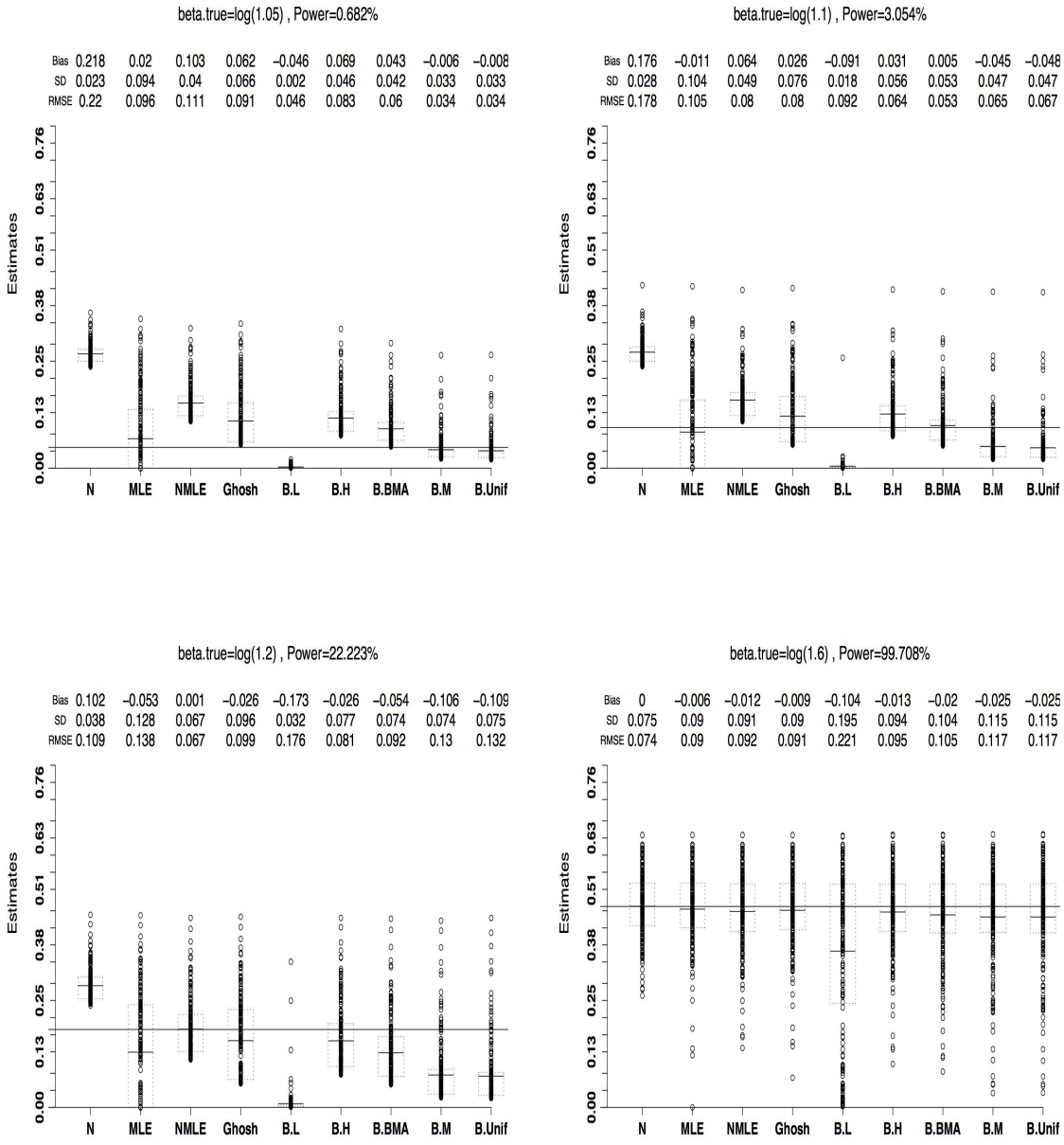

Figure 6: ***Scatter plot of the estimates of μ for the estimator B.Unif under two computation schemes for the MCMC in which the first one is having 15,000 posterior samples after discarding the first 5000 burn-in samples and the second one is having 15,000 posterior samples after discarding the first 15,000 burn-in samples under the normal model for 200 replications.*** *The true value of μ is equal to log(1.1)=0.0953. This plot shows that estimation results are approximately the same for these two schemes. The plots (which are not shown here due to space limitation) for other Bayesian estimators with different value of a and b suggest the same conclusion. Top left:$\alpha = 0.05$,power=10%, Top right: $\alpha = 0.05$,power=20%, bottom left: $\alpha = 10^{-6}$,power=5%, bottom right: $\alpha = 10^{-6}$,power=20%.*

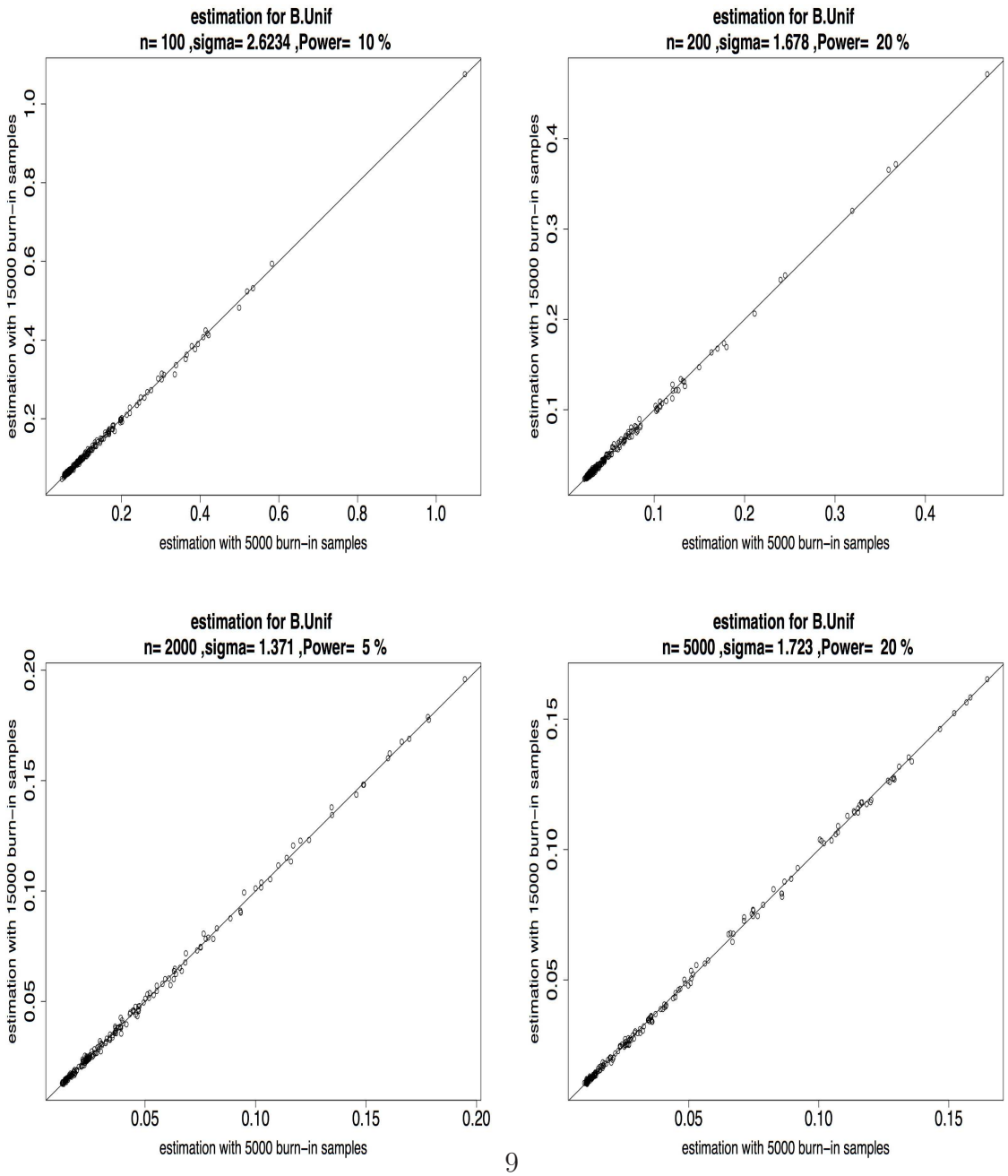

Figure 7: ***Scatter plot of the estimates of μ for the estimator B.Unif under two computation schemes for the MCMC in which the first one is having 15,000 posterior samples after discarding the first 5000 burn-in samples and the second one is having 25,000 posterior samples after discarding the first 5,000 burn-in samples under the normal model for 200 replications.*** *The true value of μ is equal to log(1.1)=0.0953. This plot shows that estimation results are approximately the same for these two schemes. The plots (which are not shown here due to space limitation) for other Bayesian estimators with different value of a and b suggest the same conclusion. Top left: type I error rate of $\alpha = 0.05$,power=10%, Top right: type I error rate of $\alpha = 0.05$,power=20%, bottom left: type I error rate of $\alpha = 10^{-6}$,power=5%, bottom right: type I error rate of $\alpha = 10^{-6}$,power=20%.*

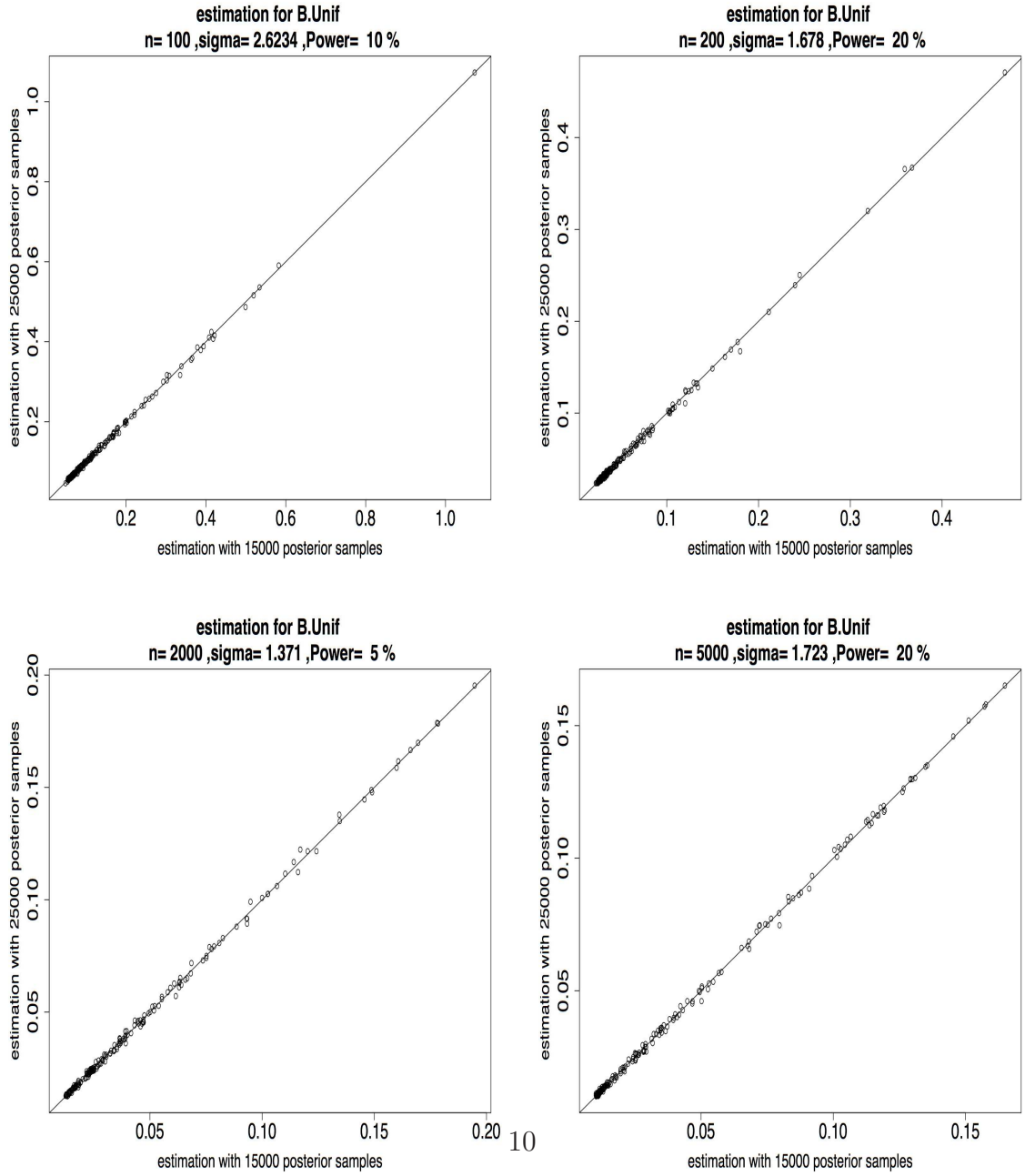

Figure 8: ***Scatter plot of the estimates of μ for the estimator B.Unif when the upper bound for the support of μ A=2 vs. A=6 under the normal model for 200 replications*** *The true value of μ is equal to log(1.1)=0.0953. This plot shows that estimation results are approximately the same for these two schemes. The plots (which are not shown here due to space limitation) for other Bayesian estimators with different value of a and b suggest the same conclusion. Left: type I error rate of $\alpha = 10^{-6}$,power=5%, Right: type I error rate of $\alpha = 10^{-6}$,power=20%.*

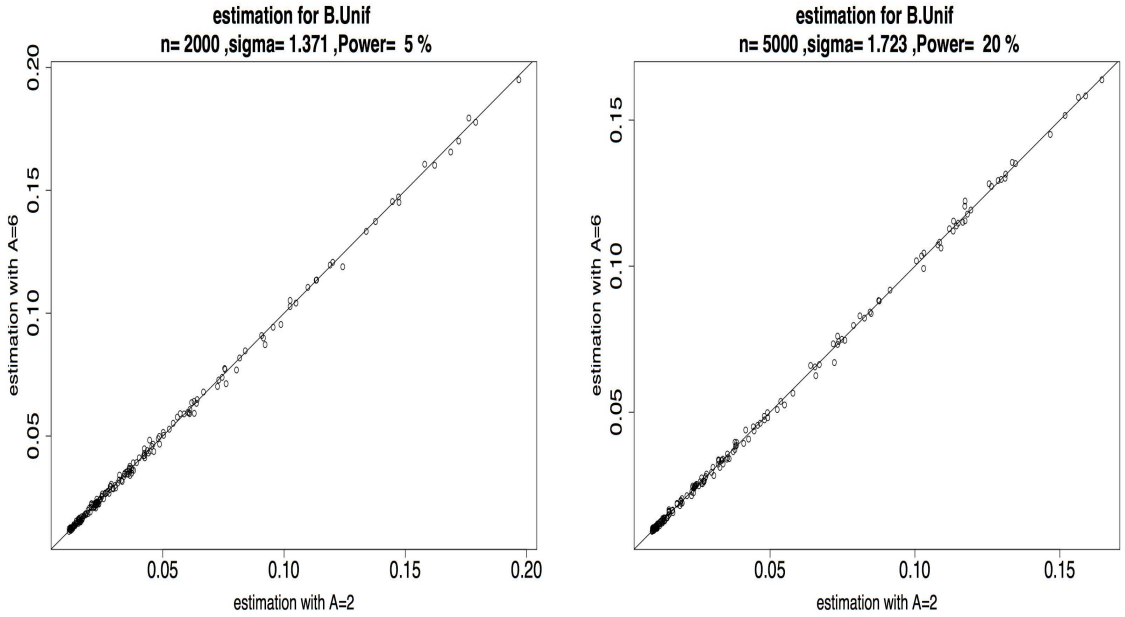

Figure 9: ***Scatter plot of the estimates of μ for the estimator B.Unif when the prior variance of σ^2 is 10 vs. 200 (Left) or 200 vs. 1000(Right) under the normal model for 200 replications when the type I error rate is 0.05 and power=0.1.*** *The true value of μ is equal to log(1.1)=0.0953. This plot shows that estimation results are pretty robust to different settings of the values of prior variance for σ^2. The plots (which are not shown here due to space limitation) for other Bayesian estimators with different value of a and b suggest the same conclusion.*