Supplementary Materials for "Estimation and Testing of Gene Expression Heterosis"

Tieming JI, Peng LIU, and Dan NETTLETON

EVALUATION OF THE APPROXIMATION OF THE JOINT POSTERIOR DENSITY

The Metropolis-Hastings algorithm to draw samples for gene j from the joint posterior density was developed and implemented as follows.

Step 1: Set the initial values for $(\alpha_j^0, \delta_j^0) = (\widehat{\alpha}_j, \widehat{\delta}_j)$, and set $n = 1$.

Step 2: Use the prior mixture model (1)–(2) in the main paper as the proposal distribution, and use

$J(\alpha_j^{new}, \delta_j^{new} | \alpha_j^{n-1}, \delta_j^{n-1})$ to denote the proposal distribution. Thus, we have

$$\begin{aligned} & J(\alpha_j^{new}, \delta_j^{new} | \alpha_j^{n-1}, \delta_j^{n-1}) \\ = \; & [\pi_\alpha \mathbf{1}_{[\alpha_j^{new}=0]} + (1-\pi_\alpha)\mathbf{1}_{[\alpha_j^{new}\neq 0]}\phi(\alpha_j^{new}|\mu_\alpha, \sigma_\alpha^2)] \\ & \times [\pi_\delta \mathbf{1}_{[\delta_j^{new}=0]} + (1-\pi_\delta)\mathbf{1}_{[\delta_j^{new}\neq 0]}\phi(\delta_j^{new}|\mu_\delta, \sigma_\delta^2)]. \end{aligned}$$

Let

$$r_n = \frac{p(\alpha_j^{new}, \delta_j^{new} | \widehat{\alpha}_j, \widehat{\delta}_j, S_j^2)/J(\alpha_j^{new}, \delta_j^{new} | \alpha_j^{n-1}, \delta_j^{n-1})}{p(\alpha_j^{n-1}, \delta_j^{n-1} | \widehat{\alpha}_j, \widehat{\delta}_j, S_j^2)/J(\alpha_j^{n-1}, \delta_j^{n-1} | \alpha_j^{new}, \delta_j^{new})}.$$

If $r_n \geq 1$, let $(\alpha_j^n, \delta_j^n) = (\alpha_j^{new}, \delta_j^{new})$. Otherwise, assign $(\alpha_j^n, \delta_j^n) = (\alpha_j^{new}, \delta_j^{new})$ with probability r_n, and assign $(\alpha_j^n, \delta_j^n) = (\alpha_j^{n-1}, \delta_j^{n-1})$ with probability $(1 - r_n)$. In computing r_n, we use formula (12) in the Appendix B of the main paper to evaluate $p(\alpha_j^{new}, \delta_j^{new} | \widehat{\alpha}_j, \widehat{\delta}_j, S_j^2)$ and $p(\alpha_j^{n-1}, \delta_j^{n-1} | \widehat{\alpha}_j, \widehat{\delta}_j, S_j^2)$.

Tieming Ji (✉) is Assistant Professor, Department of Statistics, University of Missouri at Columbia, Columbia, MO 65211, USA (E-mail: *jit@missouri.edu*). Peng Liu is Associate Professor (E-mail: *pliu@iastate.edu*) and Dan Nettleton is Laurence H. Baker Endowed Chair and Professor (E-mail: *dnett@iastate.edu*), Department of Statistics, Iowa State University, Ames, IA 50011, USA.

Step 3: Let $n = n + 1$, repeat step 2 until $n = N$.

In our implementation, we let $N = 20,000$, and discarded the draws of (α_j, δ_j)'s from the first 10,000 iterations for burn-in, and collected the last 10,000 draws of (α_j, δ_j)'s for statistical inferences.

In order to compare our approximation described in Appendix B of the main paper with the standard Metropolis-Hastings approach, we simulated a dataset with 1,000 genes based on the hierarchical model defined by equations (1)–(6) in the main paper, and used the hyper-parameters estimated from the alfalfa experiment as the true hyperparameter values. After simulating the data, we used our proposed empirical Bayes method to analyze the dataset by either sampling from posteriors using the Metropolis-Hastings algorithm (implemented in C for computing r_n in iteration n and in R for other parts) or by using the approximation of Appendix B in the main paper (implemented in R). The system running time based on the Metropolis-Hastings approach was about 4.77 hours, but the system running time based on the approximation was only about 15 seconds. When there are tens of thousands of genes in an experiment, such as the alfalfa and maize experiments, analyzing data based on the Metropolis-Hastings algorithm requires even more computing power though the increment may not be linear. We plotted the estimated posterior probabilities for exhibiting HPH, LPH, and MPH for these 1,000 genes in Figures 1(a), 1(b), and 1(c) based on these two different strategies. The plots suggest that the two approaches yield estimated posterior probabilities that are nearly identical for almost all genes.

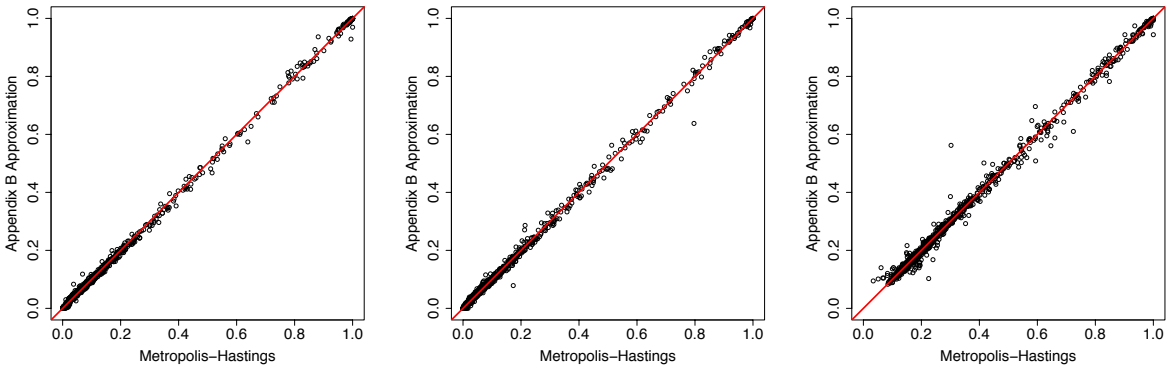
[Figure 1 about here.]

ADDITIONAL FIGURES

[Figure 2 about here.]

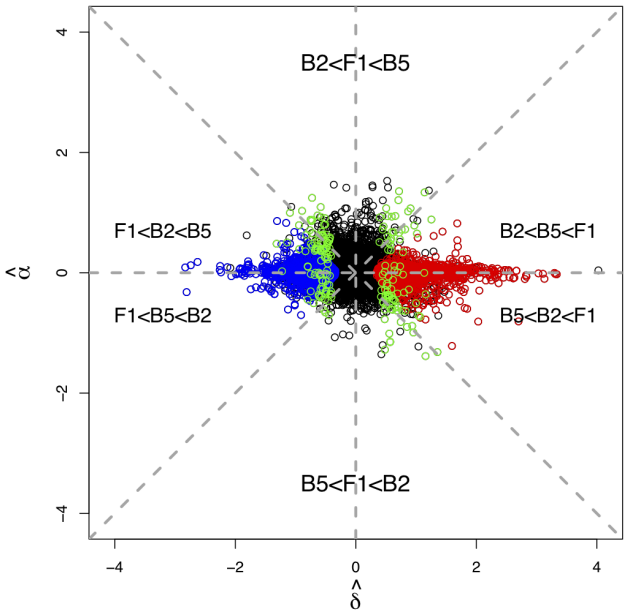
[Figure 3 about here.]

[Figure 4 about here.]

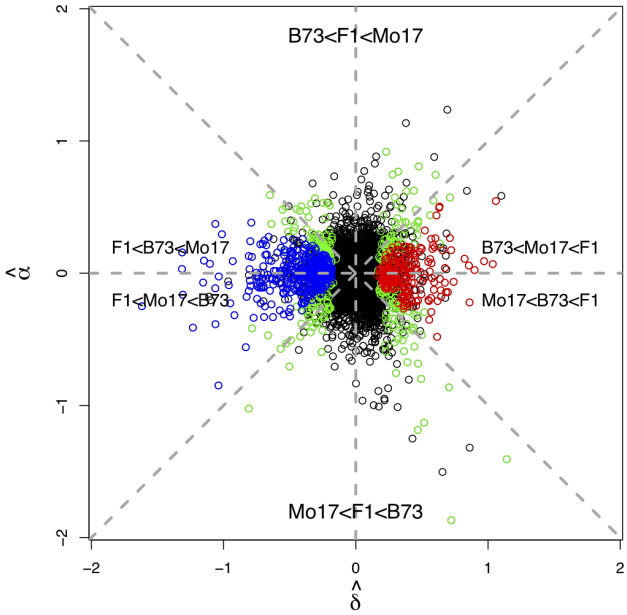


(a) (b) (c)

Figure 1: Plots for comparing estimated posterior probabilities of exhibiting (a) HPH, (b) LPH, and (c) MPH for 1,000 simulated genes using the approximation described in Appendix B of the main paper vs. sampling from the posterior via the Metropolis-Hastings algorithm.



(a) Alfalfa dataset.



(b) Maize dataset.

Figure 2: Segmentation of scatter plots of $\widehat{\alpha}_j$ vs. $\widehat{\delta}_j$ $(j = 1, \cdots, J)$ based on analysis results. Genes showing significant ($\widehat{\text{FDR}} \leq 0.05$) evidence of HPH (red), LPH (blue), MPH (red, blue, or green) are plotted, as well as genes showing no significant evidence of any form of gene expression heterosis (black).

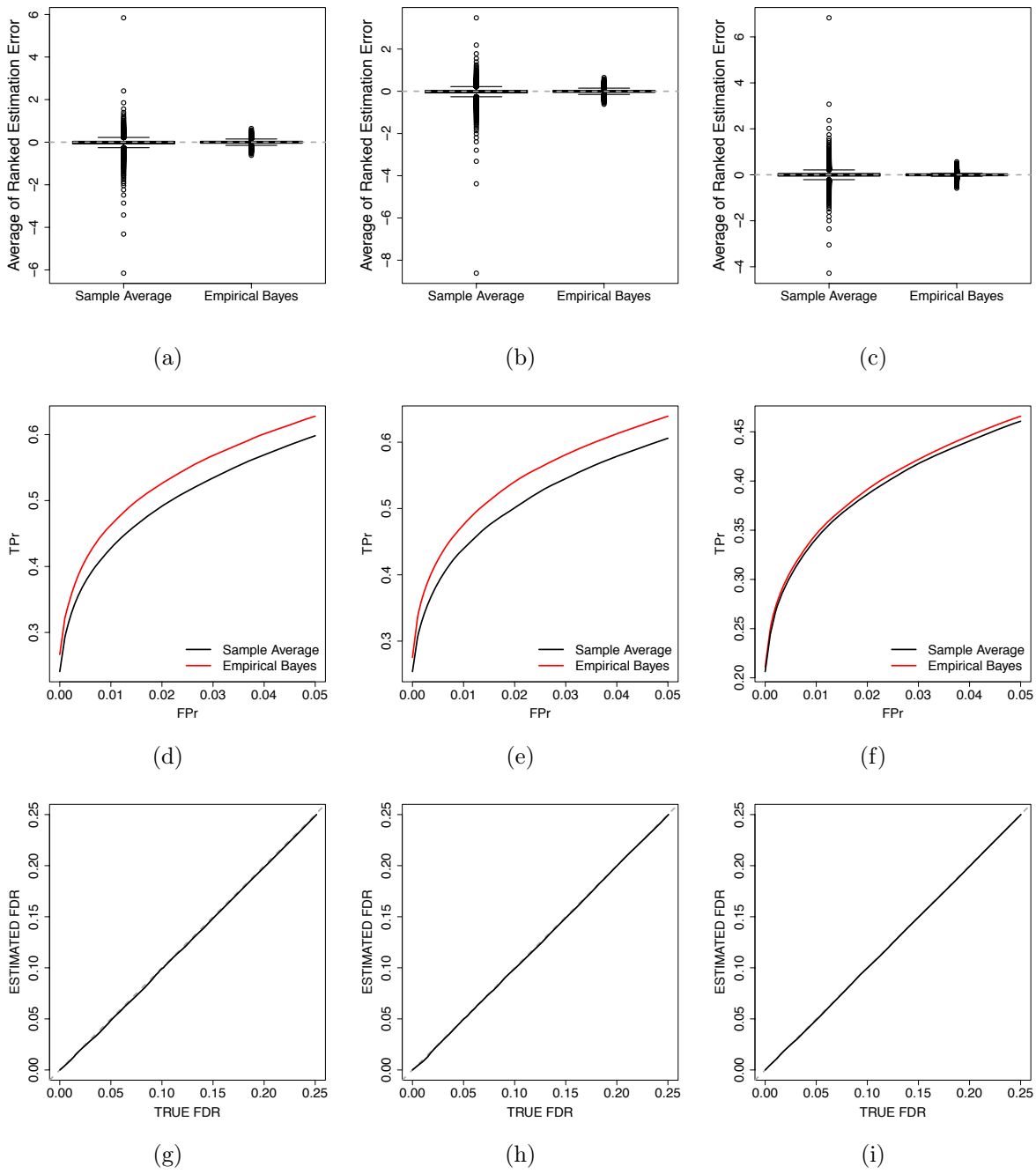


Figure 3: Plots for the simulation study 5.2 based on the maize data. Top row: box plots of ranked estimation errors averaged over 100 simulated datasets. Middle row: ROC curves averaged over 100 simulated datasets. Bottom row: estimated FDRs based on posterior probabilities versus true FDRs. Left column: HPH. Middle column: LPH. Right column: MPH.

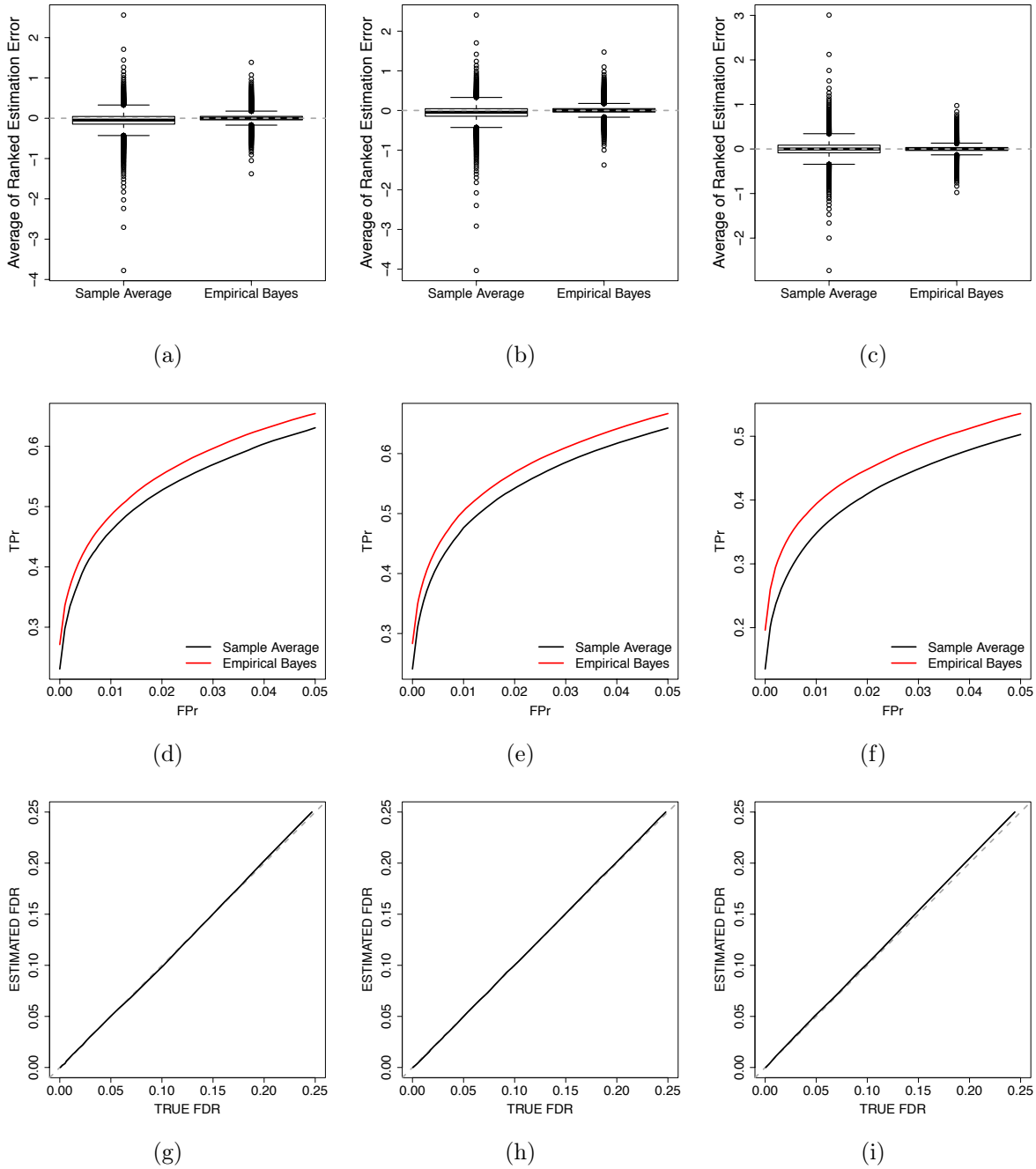


Figure 4: Plots for the simulation study 5.3 based on hypothetical probability models that violate the model assumptions of the empirical Bayes method. Top row: box plots of ranked estimation errors averaged over 100 simulated datasets. Middle row: ROC curves averaged over 100 simulated datasets. Bottom row: estimated FDRs based on posterior probabilities versus true FDRs. Left column: HPH. Middle column: LPH. Right column: MPH.