Web-based Supplementary Materials for "A Bayesian Hierarchical Model for Classification with Selection of Functional Predictors" by Hongxiao Zhu, Marina Vannucci, and Dennis D. Cox

Hongxiao Zhu

Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, Houston, TX 77230, U.S.A.

hzhu1@mdanderson.org

Marina Vannucci and Dennis D. Cox

Department of Statistics, Rice University, 6100 Main St. MS-138, Houston, Texas 77005, U.S.A.

Web Appendix A

A constructed example on how the device difference can mislead the classification in an unbalanced design

Systematic effects such as device artifacts can mislead the classification, especially in an unbalanced design. Here is a toy example. In the following table, we list the counts of objects measured by two devices for a binary classification problem. If the device difference is used to predict the classes, for example, by classifying all objects measured by device one to class one, the misclassification rate will be $(5 + 50)/365 = 15\%$, which seems quite good but is obviously biased since the device difference is purely artificial. Unfortunately, most classification algorithms can hardly recognize the sources of variation and may end up differentiating the objects based on the device difference. We refer the variations caused by device or other experimental difference as "random batch effects".

True class	Device one	Device two
Class one	300	50
Class two	5	10

Web Appendix B

Integrating $\mathbf{b}_l$, $\mathbf{b}_0$, $\boldsymbol{\alpha}$ out sequentially from $\pi(\boldsymbol{\alpha}, \mathbf{b}_1, \ldots, \mathbf{b}_L, \mathbf{b}_0, \sigma_b^2, \boldsymbol{\tau} | \mathbf{Z}_l, \mathbf{Y}_l, l = 1, \ldots, L)$

From equation (10) and the associated priors in equation (2) and (9) in the text, we have

$$
\begin{aligned}
&\pi(\boldsymbol{\alpha}, \mathbf{b}_1, \ldots, \mathbf{b}_L, \mathbf{b}_0, \sigma_b^2, \boldsymbol{\tau} | \mathbf{Z}_l, \mathbf{Y}_l, l = 1, \ldots, L) \\
&\propto \prod_l |\mathbf{K}_l^{-1}|^{-1/2} \exp\left\{ -\frac{1}{2} \sum_l \left(\mathbf{b}_l^T \mathbf{K}_l \mathbf{b}_l - 2\mathbf{b}_l^T \mathbf{M}_l + \mathbf{M}_l^T \mathbf{K}_l^{-1} \mathbf{M}_l \right) \right\} \\
&\cdot \exp\left\{ \frac{1}{2} \sum_l \left[\mathbf{M}_l^T \mathbf{K}_l^{-1} \mathbf{M}_l - (\mathbf{Z}_l - \mathbf{S}_l \boldsymbol{\alpha})^T (\mathbf{Z}_l - \mathbf{S}_l \boldsymbol{\alpha}) \right] - \frac{1}{2} \mathbf{b}_0^T \left[L(\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1} + (\sigma_0^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1} \right] \mathbf{b}_0 \right\} \\
&\cdot \exp\left\{ -\frac{1}{2} \boldsymbol{\alpha}^T (\sigma_1^2 I)^{-1} \boldsymbol{\alpha} \right\} \left(\prod_l |\mathbf{K}_l^{-1}|^{1/2} \right) |\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}}|^{-L/2} |\sigma_0^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}}|^{-1/2} \pi(\sigma_b^2) \pi(\boldsymbol{\tau}),
\end{aligned}
$$

where $\mathbf{K}_l = \mathbf{C}_l^T \mathbf{C}_l + (\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1}$ and $\mathbf{M}_l = \mathbf{C}_l^T (\mathbf{Z}_l - \mathbf{S}_l \boldsymbol{\alpha}) + (\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1} \mathbf{b}_0$, $l = 1, \ldots, L$. From above, we find the conditional distribution $\mathbf{b}_l | \boldsymbol{\alpha}, \mathbf{b}_0, \sigma_b^2, \boldsymbol{\tau}, \mathbf{Z}_l, \mathbf{Y}_l \sim N(\boldsymbol{\mu}_l, \mathbf{V}_l)$, where $\boldsymbol{\mu}_l = \mathbf{K}_l^{-1} \mathbf{M}_l$ and $\mathbf{V}_l = \mathbf{K}_l^{-1}$, for $l = 1, \ldots, L$. The $\mathbf{b}_l$'s can be integrated out from the above conditional posterior since the first $2L$ factors construct L normal density kernels. After integrating out $\mathbf{b}_l$'s, we can expand $\mathbf{M}_l^T \mathbf{K}_l^{-1} \mathbf{M}_l$ and combine the terms with $\mathbf{b}_0$, which gives the following:

$$
\begin{aligned}
&\pi(\boldsymbol{\alpha}, \mathbf{b}_0, \sigma_b^2, \boldsymbol{\tau} | \mathbf{Z}_l, \mathbf{Y}_l, l = 1, \ldots, L) \\
&\propto |\mathbf{K}_0^{-1}|^{-1/2} \exp\left\{ -\frac{1}{2} \left(\mathbf{b}_0^T \mathbf{K}_0 \mathbf{b}_0 - 2\mathbf{b}_0^T \mathbf{M}_0 + \mathbf{M}_0^T \mathbf{K}_0^{-1} \mathbf{M}_0 \right) \right\} \\
&\cdot \exp\left\{ \frac{1}{2} \mathbf{M}_0^T \mathbf{K}_0^{-1} \mathbf{M}_0 + \frac{1}{2} \sum_l (\mathbf{Z}_l - \mathbf{S}_l \boldsymbol{\alpha})^T (\mathbf{C}_l \mathbf{K}_l^{-1} \mathbf{C}_l^T - I)(\mathbf{Z}_l - \mathbf{S}_l \boldsymbol{\alpha}) \right\} \\
&\cdot \exp\left\{ -\frac{1}{2} \boldsymbol{\alpha}^T (\sigma_1^2 I)^{-1} \boldsymbol{\alpha} \right\} |\mathbf{K}_0^{-1}|^{1/2} \left(\prod_l |\mathbf{K}_l^{-1}|^{1/2} \right) |\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}}|^{-L/2} |\sigma_0^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}}|^{-1/2} \pi(\sigma_b^2) \pi(\boldsymbol{\tau}),
\end{aligned}
$$

where $\mathbf{K}_0 = (\sigma_0^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1} + L(\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1} - (\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1} (\sum_l \mathbf{K}_l^{-1}) (\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1}$ and $\mathbf{M}_0 = (\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1} \sum_l \mathbf{K}_l^{-1} \mathbf{C}_l^T (\mathbf{Z}_l - \mathbf{S}_l \boldsymbol{\alpha})$. It is easy to see from above that $\mathbf{b}_0 | \boldsymbol{\alpha}, \sigma_b^2, \boldsymbol{\tau}, \mathbf{Z}_l, \mathbf{Y}_l \sim N(\boldsymbol{\mu}_0, \mathbf{V}_0)$, where $\boldsymbol{\mu}_0 = \mathbf{K}_0^{-1} \mathbf{M}_0$ and $\mathbf{V}_0 = \mathbf{K}_0^{-1}$. We can further integrate $\mathbf{b}_0$ out since the first two factors form a normal density kernel. After integrating out $\mathbf{b}_0$, we can expand the term $\mathbf{M}_0^T \mathbf{K}_0^{-1} \mathbf{M}_0$, combine terms of $\boldsymbol{\alpha}$ and factor out a normal kernel for $\boldsymbol{\alpha}$, from where we obtain that $\boldsymbol{\alpha} | \sigma_b^2, \boldsymbol{\tau}, \mathbf{Z}_l, \mathbf{Y}_l, \forall l \sim N(\boldsymbol{\mu}_\alpha, \mathbf{V}_\alpha)$, where $\boldsymbol{\mu}_\alpha = \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{M}}$, $\mathbf{V}_\alpha = \tilde{\mathbf{K}}^{-1}$,

$$
\tilde{\mathbf{K}} = \sum_l \mathbf{S}_l^T \mathbf{S}_l + (\sigma_1^2 I)^{-1} - \sum_l \mathbf{S}_l^T \mathbf{C}_l \mathbf{K}_l^{-1} \mathbf{C}_l^T \mathbf{S}_l - (\sum_l \mathbf{K}_l^{-1} \mathbf{C}_l^T \mathbf{S}_l)^T (\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1} \mathbf{K}_0^{-1} (\sigma_b^2 \boldsymbol{\Sigma}_{\boldsymbol{\tau}})^{-1} (\sum_l \mathbf{K}_l^{-1} \mathbf{C}_l^T \mathbf{S}_l),
$$

and

$$\tilde{\mathbf{M}} = \sum_l \mathbf{S}_l^T \mathbf{Z}_l - \sum_l \mathbf{S}_l^T \mathbf{C}_l \mathbf{K}_l^{-1} \mathbf{C}_l^T \mathbf{Z}_l - (\sum_l \mathbf{K}_l^{-1} \mathbf{C}_l^T \mathbf{S}_l)^T (\sigma_b^2 \mathbf{\Sigma}_{\boldsymbol{\tau}})^{-1} \mathbf{K}_0^{-1} (\sigma_b^2 \mathbf{\Sigma}_{\boldsymbol{\tau}})^{-1} (\sum_l \mathbf{K}_l^{-1} \mathbf{C}_l^T \mathbf{Z}_l).$$

We finally can integrate out $\boldsymbol{\alpha}$ to obtain the marginal conditional posterior of σ_b^2 and $\boldsymbol{\tau}$, conditional on values of $\mathbf{Z}_l$'s and $\mathbf{Y}_l$'s, which gives

$$\begin{aligned}
&\pi(\sigma_b^2, \boldsymbol{\tau} | \mathbf{Z}_l, \mathbf{Y}_l, l = 1, \ldots, L) \\
\propto\ &\exp\left\{ \frac{1}{2} \tilde{\mathbf{M}}^T \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{M}} + \frac{1}{2} (\sum_l \mathbf{K}_l^{-1} \mathbf{C}_l^T \mathbf{Z}_l)^T (\sigma_b^2 \mathbf{\Sigma}_{\boldsymbol{\tau}})^{-1} \mathbf{K}_0^{-1} (\sigma_b^2 \mathbf{\Sigma}_{\boldsymbol{\tau}})^{-1} (\sum_l \mathbf{K}_l^{-1} \mathbf{C}_l^T \mathbf{Z}_l) \right\} \\
\cdot\ &\exp\left\{ \frac{1}{2} \sum_l \mathbf{Z}_l^T \mathbf{C}_l \mathbf{K}_l^{-1} \mathbf{C}_l^T \mathbf{Z}_l \right\} |\tilde{\mathbf{K}}|^{-1/2} |\mathbf{K}_0|^{-1/2} (\prod_l |\mathbf{K}_l|^{-1/2}) |\sigma_b^2 \mathbf{\Sigma}_{\boldsymbol{\tau}}|^{-L/2} |\sigma_0^2 \mathbf{\Sigma}_{\boldsymbol{\tau}}|^{-1/2} \pi(\sigma_b^2) \pi(\boldsymbol{\tau}),
\end{aligned}$$

where $\tilde{\mathbf{K}}, \tilde{\mathbf{M}}, \mathbf{K}_0$ and $\mathbf{K}_l$'s are defined in the above derivation.

Web Appendix C
More details on setting priors and other parameters in MCMC algorithms

In our proposed model, besides the truncation parameters p_j and the weights $\{w_k^j\}_{k=1}^{\infty}$ discussed in Section 3, there are several other priors that need to be set, including σ_1^2, σ_0^2, (d_1, d_2), ω_j's and (ν_1, ν_0).

The σ_1^2 and σ_0^2 are scaling parameters in the covariance of $\boldsymbol{\alpha}$ and $\beta_j^0(t)$'s. We usually set them between 10 and 100. Larger values also work but don't have significant influence to the posterior estimation of $\boldsymbol{\alpha}$ and $\beta_j^0(t)$'s. The parameter ω_j reflects the prior belief on the probability that the jth functional predictor is selected. If no further information is available on the preference of selecting certain functional predictor, we can set ω_j's to be a constant across all j's, and set this constant be the proportion of functional predictors we expect to select. It is harder to make choices on d_1 and d_2, which are inverse-gamma priors for the scaling parameter σ_b^2. Our suggestion is to set up a mean and variance for the inverse-gamma prior and solve for d_1 and d_2. For example, if one set the inverse-gamma prior for σ_β^2 to have mean 1 and variance 80, the resulting solution is $d_1 = 2.01$, $d_2 = 0.9$. On the choice of (ν_1, ν_0), since we have scaling parameters σ_b^2 and σ_0^2 for γ_{τ_j}, we usually fix $\nu_1 = 1$ and set ν_0 near zero(e.g, $\nu_0^2 = 10^{-6}$).

Other parameters need to be determined in the two MCMC algorithms include δ, ζ, ξ and a. Parameter δ affects the acceptance rate of σ_b^2. An empirical value of δ between 0.5 and 2 yields

acceptance rate approximated between 20% and 60%. The parameter ζ in Algorithm 2 determines the probability of mutation, which we usually set to be 0.5. The other parameter ξ determines the swapping probability in step 3 of Algorithm 1 and in the mutation step in Algorithm 2. Experiments show that adjusting values of ξ will not improve the acceptance rate of $\boldsymbol{\tau}$ significantly, so we usually set it to be 0.5. In Algorithm 2, we also need to determine temperature ladder by a geometric ratio a. The initial value of a is usualy set to be $3 - 5$.

Table 1: (Web Table) Real Data Application: The acceptance rates for the EMC algorithm based on two different function approximation methods. M-H denotes the Metropolis-Hastings update. The vector values correspond to the acceptance rates of all chains at the temperature ladder stated in the text.

Acceptance rate	Method using cosine basis expansion	Method using FPC's
M-H for σ_b^2	$(60, 45, 32, 27, 17, 13, 11, 10, 10) \times 10^{-2}$	$(59, 44, 31, 26, 16, 13, 10, 9, 9) \times 10^{-2}$
Mutation for $\boldsymbol{\tau}$	$(27, 18, 9, 4, 1, .8, .6, .6, .6) \times 10^{-2}$	$(27, 17, 7, 3, 0.9, .5, .9, .6, .6) \times 10^{-2}$
Crossover for $\boldsymbol{\tau}$	0.11	0.14
Exchange for $\boldsymbol{\tau}$	0.08	0.11

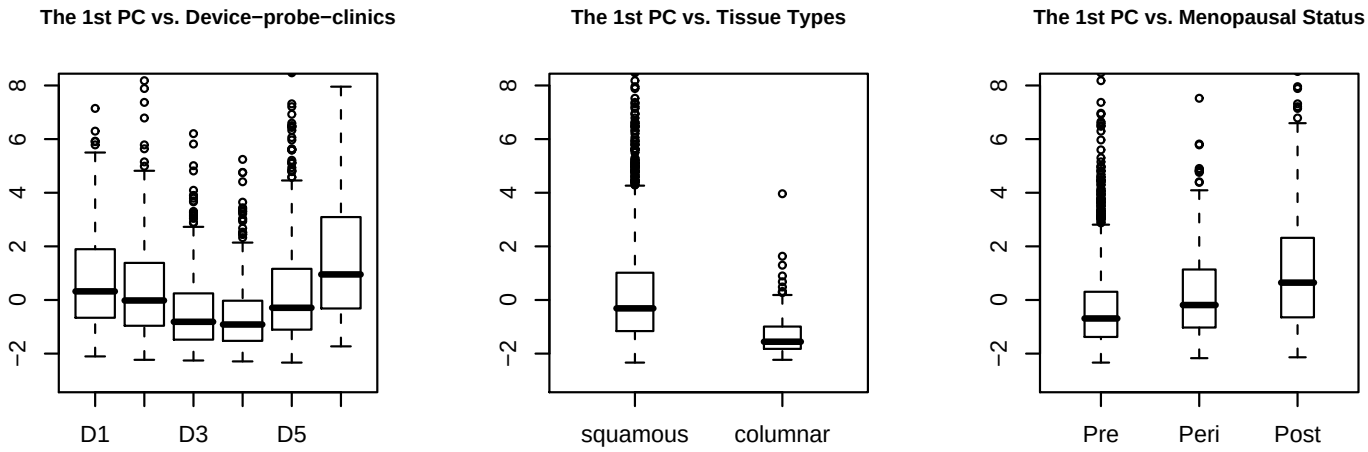

Figure 1: (Web figure) The box-plot of the first functional principle component scores of one spectral curve (measured at excitation 340 nm) versus six device-probe-clinic combinations (left), two tissue types (middle) and three menopausal states (right). Systematic differences across different levels of these factors can be seen obviously. Note that here we only used observations from the normal class, which excludes the possibility that the differences are caused by unbalanced proportions of diseased cases in each level of the factors.

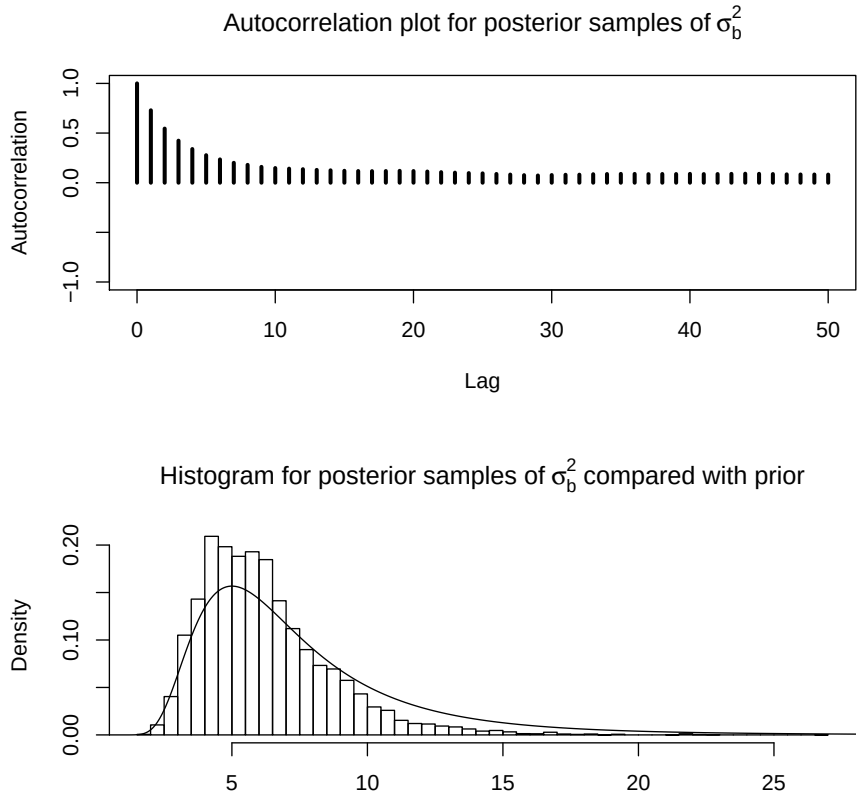

Figure 2: (Web figure) Result of Simulation 1: The autocorrelation plot for posterior samples of σ_b^2 and the corresponding histogram plot. On the bottom panel, the curve on top of the histogram is the prior density of σ_b^2.