SUPPLEMENT TO “BAYESIAN SEMIPARAMETRIC ANALYSIS FOR TWO-PHASE STUDIES OF GENE-ENVIRONMENT INTERACTION”

BY JAEIL AHN, BHRAMAR MUKHERJEE, STEPHEN B. GRUBER AND MALAY GHOSH

SUPPLEMENTARY MATERIAL

1. Appendix.

1.1. *Dunson and Xing (2009) Update:.* We describe the posterior sampling steps in relation to parameters in $P(\boldsymbol{W}|\boldsymbol{\theta})$, $\boldsymbol{\theta} = \{\boldsymbol{\psi}, \boldsymbol{V}, \alpha\}$, by following Dunson and Xing (2009). They introduce a vector of latent variables $\boldsymbol{u} = \{u_1, \ldots, u_N\}$, $u_u > 0$. The joint distribution of $\boldsymbol{u}, \boldsymbol{w}|\boldsymbol{V}, \boldsymbol{\psi}, \alpha$ is defined as,

$$\prod_{u=1}^{N} \left\{ \sum_{h \in A_{u\nu}} \prod_{j=1}^{p} \prod_{l=1}^{d_j} \psi_{hl}^{(j)\, I(w_{uj}=l)} \right\}, \tag{1.1}$$

where $A_{u\nu} = \{h : \nu_h > z_u\}$ and $\nu_h = V_h \prod_{l<h}(1 - V_l)$. The joint posterior is then the product of the augmented data likelihood (1.1) and respective priors on $\boldsymbol{V}, \boldsymbol{\psi}, \alpha$. The augmented data Gibbs sampling is based on the following steps.

(a) We start with the simplest one. Update u_u for $u = 1, \ldots, N$, by sampling from $U(0, \nu_{z_u})$

(b) Next step is regarding $\boldsymbol{\psi}_h^{(j)}$. Note that we can find $h^* = max\{z_1, \ldots, z_N\}$ such that we do not need to compute any conditionals $h > h^*$ later on. And we notice that

$$\pi(\boldsymbol{\psi}_h^{(j)}|\cdot) \propto Dirichlet(a_{j1}, \ldots, a_{jd_j}) \times \prod_{u=1}^{N} \prod_{l=1}^{d_j} \psi_{hl}^{(j)\, I(w_{uj}=l)}.$$

Due to the conjugate prior, the posterior conditional for j-th response when $z_u = h$ is given as

$$Dirichlet\left(a_{j1} + \sum_{u:z_u=h} I(w_{uj} = 1), \ldots, a_{jd_j} + \sum_{u:z_u=h} I(w_{uj} = dj) \right).$$

(c) The conditionals with respect to V_h is

$$\pi(V_h|\cdot) \propto (1 - V_h)^{\alpha-1} \times \prod_{u=1}^{N} I(u_u < V_h) \prod_{l<h}(1 - V_l).$$

If we focus on the latter part, we can obtain a beta$(1, \alpha)$ distribution truncated at

$$\left[max_{u:z_u=h} \left\{ \frac{u_u}{\prod_{l<h}(1-V_l)} \right\}, 1 - max_{u:z_u>h} \left\{ \frac{u_u}{V_{z_u} \prod_{l<z_u, l\neq h}(1-V_l)} \right\} \right].$$

(d) Update z_u from the multinomial full conditional as given,

$$Pr(z_u = h|\cdot) = \frac{I(\nu_h > u_u) \prod_{j=1}^{p} \psi^{(j)}_{hw_{uj}}}{\sum_{l \in A_{u\nu}} \prod_{j=1}^{p} \psi^{(j)}_{lw_{uj}}}, \qquad u = 1, \ldots, N.$$

As discussed in Walker (2007), we will be in trouble without latent variables $\boldsymbol{u}$ in that the choice of z_u can be infinite. Since the number of subjects in the dataset itself is finite, the cardinality of the set $A_{i\nu}$ is finite. To validate this argument, we need to find the smallest k^* such that $\sum_{h=1}^{k^*} \nu_h > 1 - \min\{u_1, \ldots, u_N\}$. Noticing that $\sum_{h=1}^{\infty} \nu_h$ is monotonically increasing and $\sum_{h=1}^{\infty} \nu_h = 1$, we can compute the desired k^*.

(e) In the last step, we update α from

$$Gamma \left(a_\alpha + h^*, b_\alpha - \sum_{h=1}^{h^*} \log(1 - V_h) \right)$$

Above steps are equivalent to Dunson and Xing (2009) except we set the maximum of the number of mixtures k such that $argmin \sum_{h=1}^{k} \nu_h > 0.99$ to avoid possible large number of mixtures, theoretically up to the data size N. This can lead to the bias which has little influence overall. In such a situation, we adjust V_h and ν_h to satisfy $\sum_h{}^k \nu_h = 1$. The practical gain from this method is that we can resort to known posterior sampling distributions and can anticipate facilitating the process.

TABLE 1

Stratified contingency table containing frequency of different configurations defined by levels of $RS1800775$ on CETP (G_1) and $RS1056836$ on CYP1B1 (G_2), statins (E), and disease status. The frequencies are based on completely observed 2,334 subjects at phase II

	CETP ($RS1800775$)						CYP1B1 ($RS1056836$)				
	Control		Case				Control		Case		
	E=0	E=1	E=0	E=1	Total		E=0	E=1	E=0	E=1	Total
G=0(A/A)	347	67	354	21	789	(C/C)	294	72	347	20	733
G=1(A/C)	373	97	454	44	968	(G/C)	471	97	522	51	1141
G=2(C/C)	233	54	267	33	587	(G/G)	188	49	206	27	470
Total	953	218	1075	98	2344		953	218	1075	98	2344

TABLE 2

Simulation results under exposure enriched sampling with all $E = 1$ in phase I data are selected in phase II for both cases and controls. We consider two association scenarios: 1) $G_1 \perp E$, $G_1 \perp G_2$, and $G_2 \perp E$ association, 2) $G_1 \perp E$, $G_1 \sim G_2$, and $G_2 \sim E$. The results are based on 200 replicated datasets, each with 1,000 cases and 1,000 controls in phase I and 800 cases and 800 controls in phase II. The approaches listed, TPFB, TPFB$_{\mathrm{emp}}$, WL, PL, UML, CML, and EB where each represents Two-phase full Bayes (with empirically obtained prior variance), Weighted likelihood, Pseudolikeliohod, Unconstrained Maximum likelihood, Constrained Maximum Likelihood, and Empirical Bayes respectively. The CML imposes G_1-E and G_1-G_2 independence, however, no constraints on G_2-E association. We set $(\beta_E, \beta_{G_1G_2}, \beta_{G_1E}, \beta_{G_2E}) = (-1.5, 0, \log(2), \log(2))$ for all scenarios. The rows with the two smallest sum(MSE) are in bold.

		$G_1 \perp E, G_1 \perp G_2, G_2 \perp E$					$G_1 \perp E, G_1 \sim G_2, G_2 \sim E$				
Stratified sampling (a)†		E	G_1 x G_2	G_1 x E	G_2 x E	Sum‡	E	G_1 x G_2	G_1 x E	G_2 x E	Sum‡
		$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (0, 0, 0)$					$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (\log(2), 0, \log(1.5))$				
TPFB	Bias	0.019	0.011	-0.022	-0.070		-0.116	0.232	-0.018	0.190	
	MSE	(0.079)	(0.035)	(0.101)	(0.095)	(**0.309**)	(0.109)	(0.062)	(0.093)	(0.111)	(**0.376**)
TPFB$_{\mathrm{emp}}$	Bias	0.019	0.038	-0.086	-0.109		-0.086	0.034	-0.048	0.126	
	MSE	(0.079)	(0.015)	(0.105)	(0.090)	(**0.294**)	(0.103)	(0.041)	(0.069)	(0.109)	(**0.323**)
WL	Bias	-0.030	-0.009	0.049	0.030		-0.074	0.006	0.047	0.071	
	MSE	(0.097)	(0.050)	(0.147)	(0.134)	(0.428)	(0.132)	(0.048)	(0.123)	(0.137)	(0.439)
PL	Bias	-0.030	-0.010	0.050	0.030		-0.074	0.004	0.047	0.071	
	MSE	(0.097)	(0.049)	(0.146)	(0.134)	(0.426)	(0.132)	(0.048)	(0.123)	(0.137)	(0.439)
UML	Bias	-0.049	-0.010	0.050	0.030		-0.095	0.004	0.047	0.071	
	MSE	(0.099)	(0.049)	(0.146)	(0.134)	(0.428)	(0.135)	(0.048)	(0.123)	(0.137)	(0.443)
CML	Bias	-0.042	-0.007	0.042	0.009		-0.080	0.705	0.023	0.061	
	MSE	(0.086)	(0.023)	(0.088)	(0.123)	(0.320)	(0.117)	(0.274)	(0.080)	(0.128)	(0.600)
EB	Bias	-0.044	-0.010	0.042	0.014		-0.083	0.112	0.035	0.062	
	MSE	(0.089)	(0.030)	(0.104)	(0.126)	(0.349)	(0.119)	(0.059)	(0.091)	(0.129)	(0.397)

†All subjects with $E = 1$ in case and control are sub-sampled for phase II.

‡The combined MSEs over all four parameters.

TPFB uses the informative prior $N(0, 10^{-2})$ on G-G and G-E associations in the model (**??**)

TPFB$_{\mathrm{emp}}$ uses the prior $N(0, \hat{\theta}^2)$ on G-G and G-E associations in the model (**??**) where $\hat{\theta}^2$ is empirically estimated G-G or G-E association parameter under controls.

TABLE 3

Simulation results under **NO** *exposure enriched sampling in phase II. A random sample of cases and controls from phase I are selected for genotyping in phase II. We consider two association scenarios 1) $G_1 \perp E$, $G_1 \perp G_2$, and $G_2 \perp E$ association, 2) $G_1 \perp E$, $G_1 \sim G_2$, and $G_2 \sim E$. The results are based on 1,000 replicated datasets, each with 1,000 cases and 1,000 controls in phase I and 600 cases and 600 controls in phase II. The approaches listed, TPFB, TPFB$_{\mathrm{emp}}$, WL, PL, UML, CML, and EB where each represents Two-phase full Bayes (with empirically obtained prior variance), Weighted likelihood, Pseudolikeliohod, Unconstrained Maximum likelihood, Constrained Maximum Likelihood, and Empirical Bayes respectively. The CML imposes G_1-E and G_1-G_2 independence, however, no constraints on G_2-E association. We set $(\beta_E, \beta_{G_1G_2}, \beta_{G_1E}, \beta_{G_2E}) = (-1.5, 0, \log(2), \log(2))$ for all scenarios. The rows with the two smallest sum(MSE) are in bold.*

		$G_1 \perp E, G_1 \perp G_2, G_2 \perp E$					$G_1 \perp E, G_1 \sim G_2, G_2 \sim E$				
Random sampling †		E	G_1 x G_2	G_1 x E	G_2 x E	Sum (MSE)‡	E	G_1 x G_2	G_1 x E	G_2 x E	Sum (MSE)‡
		$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (0, 0, 0)$					$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (\log(2), 0, \log(1.5))$				
TPFB	Bias	-0.055	0.019	-0.007	-0.024		-0.119	0.151	-0.025	0.197	
	(MSE)	(0.138)	(0.059)	(0.199)	(0.182)	(0.579)	(0.173)	(0.070)	(0.139)	(0.207)	(0.589)
TPFB$_{\mathrm{emp}}$	Bias	-0.019	0.011	-0.051	-0.055		-0.056	0.025	-0.055	0.092	
	(MSE)	(0.138)	(0.038)	(0.188)	(0.179)	(**0.543**)	(0.167)	(0.056)	(0.154)	(0.206)	(**0.583**)
WL	Bias	-0.070	0.013	0.037	0.019		-0.056	-0.007	-0.013	0.071	
	(MSE)	(0.157)	(0.070)	(0.240)	(0.225)	(0.692)	(0.167)	(0.052)	(0.180)	(0.204)	(0.603)
PL	Bias	-0.070	0.013	0.038	0.020		-0.056	-0.006	-0.013	0.071	
	(MSE)	(0.157)	(0.070)	(0.240)	(0.225)	(0.692)	(0.167)	(0.052)	(0.179)	(0.204)	(0.602)
UML	Bias)	-0.065	0.013	0.038	0.020		-0.065	-0.006	-0.013	0.071	
	(MSE)	(0.163)	(0.070)	(0.240)	(0.225)	(0.698)	(0.191)	(0.052)	(0.179)	(0.204)	(0.626)
CML	Bias	-0.056	0.006	0.007	0.015		-0.087	0.697	0.026	0.069	
	(MSE)	(0.139)	(0.036)	(0.138)	(0.204)	(**0.517**)	(0.181)	(0.512)	(0.108)	(0.190)	(0.991)
EB	Bias	-0.059	0.010	0.024	0.017		-0.084	0.069	0.009	0.068	
	(MSE)	(0.143)	(0.048)	(0.163)	(0.207)	(0.561)	(0.181)	(0.061)	(0.124)	(0.191)	(**0.558**)

†In terms of stratified sampling, 600 cases and 600 control are randomly selected for phase II, respectively.

‡The combined MSEs over all four parameters.

TPFB uses the informative prior $N(0, 10^{-2})$ on G-G and G-E associations in the model.

TPFB$_{\mathrm{emp}}$ uses the prior $N(0, \hat{\theta}^2)$ on G-G and G-E associations in the model where $\hat{\theta}^2$ is empirically estimated G-G or G-E association parameter under controls.

Supplement: C source code
(http://www.umich.edu/∼jaeil/tp.zip). Zipped C-code for MECC data analysis

References.

[1] DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes Modeling of Multivariate Categorical Data. *Journal Amer. Stat. Assoc.* **104** 1042–1051.

[2] WALKER, S. G. (2007). Sampling the Dirichlet Mixture Model With Slices. *Simulation and Computation* **36** 45–54.