

Sensitivity analysis in Bayesian nonignorable selection model for binary responses

Seong Mi Choi¹ · Dal Ho Kim²

^{1,2}Department of Statistics, Kyungpook National University

Received 31 October 2013, revised 19 November 2013, accepted 26 November 2013

Abstract

We consider a Bayesian nonignorable selection model to accommodate the selection bias. Markov chain Monte Carlo methods is known to be very useful to fit the nonignorable selection model. However, sensitivity to prior assumptions on parameters for selection mechanism is a potential problem. To quantify the sensitivity to prior assumption, the deviance information criterion and the conditional predictive ordinate are used to compare the goodness-of-fit under two different prior specifications. It turns out that the ‘MLE’ prior gives better fit than the ‘uniform’ prior in viewpoints of goodness-of-fit measures.

Keywords: Bayesian, binary, CPO, DIC, goodness-of-fit, MCMC, nonignorable, selection bias, sensitivity.

1. Introduction

A serious concern of many government agencies is selection bias in survey data. In many complex sample surveys, units are sampled with different probabilities of selection, and these are incorporated in the survey weights (i.e., reciprocals of selection probabilities). Our interest is to infer about a finite population proportion using binary responses from a possibly biased sample. Kwak and Kim (2012) studied Bayesian estimation for the finite population proportions in multinomial data without selection bias.

In the absence of any selection bias or survey weights, a simple ignorable selection model, which assumes that the binary responses are independent and identically distributed Bernoulli random variables, can be used. However, this ignorable selection model is inappropriate when there is a selection bias in the sample. A nonignorable selection model allows for binary responses with selection bias mechanism through the inclusion of some correlation between the binary responses and selection probabilities.

To incorporate the selection bias into the ignorable selection model, Malec *et al.* (1999) use a hierarchical Bayesian model to estimate a finite population proportion when there are binary data. Difficulty in including the selection probabilities directly in the model forces them to make an ad hoc adjustment to the likelihood function and use an empirical

¹ Ph.D. candidate, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea.

² Corresponding author: Professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail: dalkim@knu.ac.kr

Bayes approach. Nandram and Choi (2010) have incorporated selection probabilities into a nonignorable nonresponse model to analyze continuous data using a full Bayesian analysis. Nandram *et al.* (2013) have developed the likelihood which includes the specific relation between the selection probabilities and a binary characteristic. After putting a proper informative prior on parameters for selection mechanism, they used Markov chain Monte Carlo (MCMC) methods to fit Bayesian nonignorable selection model.

A critical problem in the formulation of Bayesian nonignorable selection model could be the specification of the prior distribution for parameters for selection mechanism. To investigate the sensitivity to the prior assumptions, the deviance information criterion and the conditional predictive ordinate are used to compare the goodness-of-fit under two different prior specifications. Recently Roos and Held (2011) discussed the sensitivity analysis in Bayesian generalized linear mixed models for binary data in infinite population setup.

In this paper, we consider sensitivity investigations for Bayesian nonignorable selection models for binary responses with selection bias mechanism. In Section 2, we describe the Bayesian nonignorable selection model and the sensitivity analysis using MCMC. In Section 3, we provide numerical results for sensitivity analysis using the simulated data sets. Some concluding remarks are given in Section 4.

2. Sensitivity analysis using MCMC

2.1. Modeling

Let y_1, \dots, y_N denote the finite population values with the corresponding selection probabilities $\pi_i, i = 1, \dots, N$. Let $P = \sum_{i=1}^N y_i/N$ denote the finite population proportion. A sample S of size n is taken from the population; also let \bar{S} denote the set of nonsampled values. Let the sampled values be y_1, \dots, y_n . Our target is to predict P when a sample information $(y_i, \pi_i), i = 1, \dots, n$ is available.

We assume that the sample selection probabilities (π_1, \dots, π_n) have support over the set $\pi_u^*, u = 1, \dots, U$. That is, $\pi_i, i = 1, \dots, n$, have a histogram where the midpoints of the categories are the π_u^* . Throughout these π_u^* are assumed known and the π_i are assumed to be random quantities. The distribution of the selection probabilities, given the binary response y_i , is

$$P(\pi_i = \pi_u^* | \boldsymbol{\theta}, y_i = y) = \theta_{uy}, \quad u = 1, \dots, U, \quad y = 0, 1, \quad i = 1, \dots, n$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = (\theta_{10}, \dots, \theta_{U0}, \theta_{11}, \dots, \theta_{U1})'$ and

$$y_i | p \stackrel{iid}{\sim} \text{Bernoulli}(p), \quad i = 1, \dots, N.$$

Note that p is the proportion of ones in the entire superpopulation.

Nandram *et al.* (2013) clearly expressed the probability mass function of a single observation

$$P(Y = y, \pi = \pi_u^* | \boldsymbol{\theta}, p) = \frac{\pi_u^* \theta_{uy} p^y (1-p)^{1-y}}{\sum_y \theta_{uy} p^y (1-p)^{1-y}} \quad (2.1)$$

by updating the derivation in Malec *et al.* (1999). Then we easily show that

$$P(Y = y | \pi = \pi_u^*, \boldsymbol{\theta}, p) = \frac{\theta_{uy} p^y (1-p)^{1-y}}{\sum_y \theta_{uy} p^y (1-p)^{1-y}} \quad (2.2)$$

and

$$P(Y = y | \boldsymbol{\theta}, p) = \frac{\sum_u \pi_u^* \theta_{uy} p^y (1-p)^{1-y}}{\sum_y \sum_u \pi_u^* \theta_{uy} p^y (1-p)^{1-y}}. \quad (2.3)$$

The sampled data actually come from the probability mass function in (2.3) and the entire population is described by $P(Y = y | p) = p^y (1-p)^{1-y}$, $y = 0, 1$, thereby showing how the selection bias enters into the model.

Let $\mathbf{y} = (y_1, \dots, y_n)'$ and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)'$. Since the sampling units are independent, the joint density of the entire sample is

$$P(\mathbf{y}, \boldsymbol{\pi} | \boldsymbol{\theta}, p) = \frac{\prod_{u=1}^U (\pi_u^* \theta_{u0})^{g_{u0}} \prod_{u=1}^U (\pi_u^* \theta_{u1})^{g_{u1}}}{[p \sum_u \pi_u^* \theta_{u1} + (1-p) \sum_u \pi_u^* \theta_{u0}]^n} p^s (1-p)^{(n-s)}, \quad (2.4)$$

where $s = \sum_{i \in S} y_i$, g_{u0} is the cell count for category u for $y = 0$ and g_{u1} is the cell counts for category u for $y = 1$. Note that $\sum_{u=1}^U g_{u0} = n - s$, $\sum_{u=1}^U g_{u1} = s$ and $\sum_{u=1}^U (g_{u0} + g_{u1}) = n$. This likelihood includes the selection bias.

A priori we assume that p , $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are independent, and we take

$$p \sim \text{Uniform}(0, 1)$$

$$\boldsymbol{\theta}_0 | \tau \sim \text{Dirichlet}(\boldsymbol{\theta}_0^{(0)} \tau) \text{ and } \boldsymbol{\theta}_1 | \tau \sim \text{Dirichlet}(\boldsymbol{\theta}_1^{(0)} \tau),$$

where $\boldsymbol{\theta}_0^{(0)}$ and $\boldsymbol{\theta}_1^{(0)}$ are to be specified. Finally, we put

$$p(\tau) = \frac{1}{(1 + \tau)^2}, \tau \geq 0,$$

which is called a shrinkage prior (see Gelman, 2006).

2.2. MCMC

Using Bayes' theorem, the joint posterior density of $p, \boldsymbol{\theta}_1, \boldsymbol{\theta}_0, \tau$ given the data, $\boldsymbol{\pi}, \mathbf{y}$, is

$$\begin{aligned} h(p, \boldsymbol{\theta}_1, \boldsymbol{\theta}_0, \tau | \boldsymbol{\pi}, \mathbf{y}) &\propto \frac{\prod_{u=1}^U (\pi_u^* \theta_{u0})^{g_{u0}} \prod_{u=1}^U (\pi_u^* \theta_{u1})^{g_{u1}}}{[p \sum_u \pi_u^* \theta_{u1} + (1-p) \sum_u \pi_u^* \theta_{u0}]^n} p^s (1-p)^{n-s} \\ &\times \frac{\prod_{u=1}^U \theta_{u0}^{\theta_{u0}^{(0)} \tau - 1}}{D(\boldsymbol{\theta}_0^{(0)} \tau)} \frac{\prod_{u=1}^U \theta_{u1}^{\theta_{u1}^{(0)} \tau - 1}}{D(\boldsymbol{\theta}_1^{(0)} \tau)} \frac{1}{(1 + \tau)^2}, \end{aligned} \quad (2.5)$$

where $\sum_{u=1}^U g_{u0} = n - s$ and $\sum_{u=1}^U g_{u1} = s$. For convenience, we drop $\boldsymbol{\pi}$ from the conditioning.

Using the joint posterior density in (2.5) and assuming the π_u^* are fixed and known, to perform the Gibbs sampler, we need the conditional posterior densities, given by

$$\begin{aligned} h_1(\boldsymbol{\theta}_0 | \boldsymbol{\theta}_1, p, \tau, \mathbf{y}) &\propto \frac{\prod_{u=1}^U (\pi_u^* \theta_{u0})^{g_{u0}}}{[a_1 p + a_0 (1-p)]^n} \prod_{u=1}^U \theta_{u0}^{\theta_{u0}^{(0)} \tau - 1}, \\ h_2(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_0, p, \tau, \mathbf{y}) &\propto \frac{\prod_{u=1}^U (\pi_u^* \theta_{u1})^{g_{u1}}}{[a_1 p + a_0 (1-p)]^n} \prod_{u=1}^U \theta_{u1}^{\theta_{u1}^{(0)} \tau - 1}, \end{aligned}$$

$$h_3(p|\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \tau, \mathbf{y}) \propto \frac{1}{[a_1 p + a_0(1-p)]^n} p^s (1-p)^{n-s},$$

and

$$h_4(\tau | \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, p, \mathbf{y}) \propto \left(\frac{\Gamma(\tau)}{1+\tau} \right)^2 \prod_{u=1}^U \left\{ \frac{\theta_{u0}^{\theta_{u0}\tau-1} \theta_{u1}^{\theta_{u1}\tau-1}}{\Gamma(\theta_{u0}\tau) \Gamma(\theta_{u1}\tau)} \right\},$$

where $a_y = \sum_u \pi_u^* \theta_{uy}$, $y = 0, 1$.

Because of the accept-reject method for drawing p within the Gibbs chain, this Gibbs sampler is a bit slow. So we use an alternative procedure that avoids this computational inefficiency. We make a one-to-one transformation from p to q via $q = \frac{a_1 p}{a_1 p + a_0(1-p)}$. To accelerate the Gibbs sampler, we integrate out q from the joint posterior. So we generate samples for p independently with other parameters using the simple accept-reject algorithm. Then we execute the Gibbs sampler for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_0, \tau$ using a grid method. Here we use $(1, \dots, 1)'$ for $\boldsymbol{\theta}_1$ and 5 for τ as initial values in Gibbs chain. For details, see Nandram *et al.* (2013).

Once we obtain the sample of $p^{(h)}$, $h = 1, \dots, H$, from the Gibbs sampler, we obtain a sample of P , say $P^{(h)}$, by drawing $\sum_{i=1}^N y_i$ from $\text{Binomial}(N, p^{(h)})$ and dividing the result by N . Thus, we have obtained a Rao-Blackwellized estimator of the posterior density of P . The posterior mean is given by $\hat{P} = \sum_{h=1}^H P^{(h)}/H$. Also the posterior standard deviation can be obtained easily.

2.3. Sensitivity analysis

Sensitivity analysis is a technique used to determine how different priors will impact a posterior analysis under given set of scenario. We study the sensitivity to the prior assumptions on the parameters for accommodating selection mechanism. This is a key assumption in a Bayesian nonignorable selection model.

We consider two different types of prior specifications. First, $\boldsymbol{\theta}_y | \tau \stackrel{iid}{\sim} \text{Dirichlet}(\boldsymbol{\theta}_y^{(0)}\tau)$, $y = 0, 1$ and τ has a proper density $p(\tau) = \frac{1}{(1+\tau)^2}$, $\tau \geq 0$, which are given in Section 2. The hyperparameters $\boldsymbol{\theta}_y^{(0)}$ can be specified by the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}_y$. Malec *et al.* (1999) used the MLE to obtain a Bayes empirical Bayes analysis. In specific, $\hat{\boldsymbol{\theta}}_y$ has the form

$$\hat{\theta}_{uy} = \frac{g_{uy}\pi_u^*}{\sum_u g_{uy}\pi_u^*}, u = 1, \dots, U, y = 0, 1.$$

Using this MLE we have $\boldsymbol{\theta}_y | \tau \stackrel{iid}{\sim} \text{Dirichlet}(\hat{\boldsymbol{\theta}}_y\tau)$, which we call a ‘MLE’ prior. Second, we can choose $\boldsymbol{\theta}_y^{(0)} = (1, \dots, 1)'$, $y = 0, 1$ and some constant for τ . So we have a prior $\boldsymbol{\theta}_y \stackrel{iid}{\sim} \text{Dirichlet}(1, \dots, 1)$, which we call a ‘uniform’ prior.

A simple sensitivity analysis by the comparison of the posterior densities and some posterior quantities was considered by Nandram *et al.* (2013). Our interest is the sensitivity analysis in viewpoints of some goodness-of-fit procedures. We will compare the measures under two prior specifications.

The deviance information criterion (DIC) is a well-known Bayesian model choice criterion for comparing complex hierarchical models (Spiegelhalter *et al.*, 2002). Lower DIC values correspond to better models.

Recall that the likelihood is given by

$$P(\mathbf{y}_s, \boldsymbol{\pi} \mid \boldsymbol{\theta}, p) = \frac{\prod_{u=1}^U (\pi_u^* \theta_{u0})^{g_{u0}} \prod_{u=1}^U (\pi_u^* \theta_{u1})^{g_{u1}}}{[p \sum_u \pi_u^* \theta_{u1} + (1-p) \sum_u \pi_u^* \theta_{u0}]^n} p^s (1-p)^{(n-s)}.$$

Let $\theta_{uy}^{(h)}$, $u = 1, \dots, U$, $y = 0, 1$ and $p^{(h)}$, $h = 1, \dots, H$ denote the iterates from the MCMC output under the Bayesian nonignorable selection model. Let $\bar{\theta}_{uy} = \sum_{h=1}^H \theta_{uy}^{(h)} / H$, $u = 1, \dots, U$, $y = 0, 1$ and $\bar{p} = \sum_{h=1}^H p^{(h)} / H$. Letting $D(\bar{\boldsymbol{\theta}}, \bar{p}) = -2 \log\{P(\mathbf{y}_s, \boldsymbol{\pi} \mid \bar{\boldsymbol{\theta}}, \bar{p})\}$ and $\bar{D} = -2 \sum_{h=1}^H \log P(\mathbf{y}_s, \boldsymbol{\pi} \mid \boldsymbol{\theta}^{(h)}, p^{(h)}) / H$, the DIC is given by

$$DIC = 2\bar{D} - D(\bar{\boldsymbol{\theta}}, \bar{p}).$$

A cross-validatory (or “leave-one-out”) approach was suggested by Gelfand *et al.* (1992). The leave-one-out cross-validatory density under the Bayesian nonignorable selection bias model is given by

$$P(Y = y, \pi = \pi_u^* \mid (\mathbf{y}, \boldsymbol{\pi})_{(uy)}) = \int f(Y = y, \pi = \pi_u^* \mid (\mathbf{y}, \boldsymbol{\pi})_{(uy)}, \Omega) \pi(\Omega \mid (\mathbf{y}, \boldsymbol{\pi})_{(uy)}) d\Omega,$$

where $(\mathbf{y}, \boldsymbol{\pi})_{(uy)}$ denotes the data without the (uy) th observation, $u = 1, \dots, U$, $y = 0, 1$, and $\Omega = (\boldsymbol{\theta}, p)$. This is called the conditional predictive ordinate (CPO). Note that the likelihood of a single observation is given by

$$P(Y = y, \pi = \pi_u^* \mid \Omega) = \frac{\pi_u^* \theta_{uy} p^y (1-p)^{1-y}}{\sum_{u=1}^U \pi_u^* \sum_{y=0}^1 \theta_{uy} p^y (1-p)^{1-y}}, u = 1, \dots, U, y = 0, 1.$$

Using MCMC output, the CPO can be estimated by

$$\widehat{CPO}_{uy} = \sum_{h=1}^H w_h P(Y = y, \pi = \pi_u^* \mid \Omega^{(h)}), u = 1, \dots, U, y = 0, 1$$

where $w_h = \frac{1}{f(Y=y, \pi=\pi_u^* \mid \Omega^{(h)})} / \sum_{h=1}^H \frac{1}{f(Y=y, \pi=\pi_u^* \mid \Omega^{(h)})}$, $h = 1, \dots, H$.

A pointwise comparison of CPO can be given in plot, along with a reference line marking where the values are equal for the two models. Larger CPO values are indicative of better model fit. One might even view the product of CPO values as a “pseudo marginal likelihood,” which gives a cross-validatory summary measure of fit (Carlin and Louis, 2009). For computational convenience, we use the mean logarithmic CPO defined as

$$\overline{LCPO}_1 = -\frac{1}{2U} \sum_{u=1}^U \sum_{y=0}^1 \log(\widehat{CPO}_{uy}).$$

Alternatively we can use the weighted mean of logarithmic CPO's,

$$\overline{LCPO}_2 = -\frac{1}{n} \sum_{u=1}^U \sum_{y=0}^1 g_{uy} \log(\widehat{CPO}_{uy}).$$

with the cell counts g_{uy} for category $u = 1, \dots, U$ and $y = 0, 1$ as weights. The mean logarithmic CPO can be identified as the cross-validated logarithmic score (Gneiting and Raftery, 2007), which measures the predictive quality of a model. Lower values of the mean logarithmic CPO indicate a better model.

3. Numerical results

We illustrate the results in preceding section with the analysis of simulated data sets. We generate a finite population of size $N = 500$ with selection probability π . Then a sample of size $n = 50$ is taken from this finite population. The outline of the generation is as follows.

- Set $\tau = 100$, $p = 0.5$, $f = n/N$, $a = 0.95$, $\mu_0 = af$, $\mu_1 = f/a$.
- Generate $u \sim U(0, 1)$. If $u \leq p$ then we set $y_i = 1$; otherwise set $y_i = 0$, $i = 1, \dots, N$.
- If $y_i = 1$ then we generate $\pi_i \sim \text{Beta}(\mu_1\tau, (1 - \mu_1)\tau)$; if $y_i = 0$ then we generate $\pi_i \sim \text{Beta}(\mu_0\tau, (1 - \mu_0)\tau)$ for $i = 1, \dots, N$.
- Sample n units by systematic PPS sampling with probabilities $\pi_i = \frac{n\pi_i}{\sum_{i=1}^N \pi_i}$.

We control some correlation between the binary responses and the selection probabilities by changing the a values. If we set $a = 0.95, 0.90, 0.80, 0.70$ in the simulation scheme, then the correlations are obtained by $|\rho| = 0.208, 0.327, 0.606, 0.762$, respectively. So we have four simulated data sets with true $P = 0.482$.

In order to determine π_u^* in given data set, we order π_i 's from smallest to largest $\pi_{(1)}, \dots, \pi_{(n)}$. Let the quantiles be $t_1 = \pi_{0.20n}$, $t_2 = \pi_{0.40n}$, $t_3 = \pi_{0.60n}$, $t_4 = \pi_{0.80n}$ and let $t_0 = \pi_{(1)}$ and $t_5 = \pi_{(n)}$. We define $\pi_u^* = (t_{u-1} + t_u)/2$, $u = 1, \dots, 5$ (i.e., the midpoints). Here we set $U = 5$.

Using the simulated data sets, we compare the point estimates for the finite population proportion as well as the goodness-of-fit measures for the sensitivity analysis under two different prior specifications. We report the posterior mean (PM) and the posterior standard deviation (PSD) as the basic statistics. Also we calculate the DIC and two mean logarithmic CPO for investigating the sensitivity to the prior assumptions. The results are reported in Table 3.1.

From Table 3.1, we may find that the MLE prior provides the better estimates than the uniform prior in the sense that estimates are closer to true $P = 0.482$ and have smaller values of PSD in all cases except $a = 0.8$. Moreover, the MLE prior has smaller \overline{LCPO}_2 values than the uniform prior in all cases. But the MLE prior has slightly larger \overline{LCPO}_1 value than the uniform prior in case of $a = 0.7$. This might cause misleading since \overline{LCPO}_1 is defined only for the histogram with the midpoint values π_u^* whereas \overline{LCPO}_2 is extended to all sample values.

Table 3.1 Comparison of DIC, \overline{LCPO}_1 , \overline{LCPO}_2 under two different prior specifications

$a(\rho)$	prior	PM	PSD	DIC	\overline{LCPO}_1	\overline{LCPO}_2
0.70	Uniform	.519	.076	201.519	3.272	0.173
(0.762)	MLE	.506	.075	195.340	3.656	0.152
0.80	Uniform	.449	.074	229.986	2.708	0.262
(0.606)	MLE	.436	.072	220.341	2.645	0.250
0.90	Uniform	.490	.071	233.095	2.635	0.237
(0.327)	MLE	.480	.070	223.353	2.560	0.222
0.95	Uniform	.497	.074	234.431	2.610	0.226
(0.208)	MLE	.488	.071	224.417	2.510	0.206

Moreover, we make a pointwise comparison of CPO values under two different prior specifications using the four different a values. The results are shown in Figure 3.1. Notice that

the dashed line represents a reference line making where the values are equal for the two models. An inspection of Figure 3.1 reveals that the MLE prior has larger CPO values than the uniform prior in most cases. Notice that larger CPO values are indicative of better model fit in pointwise comparisons.

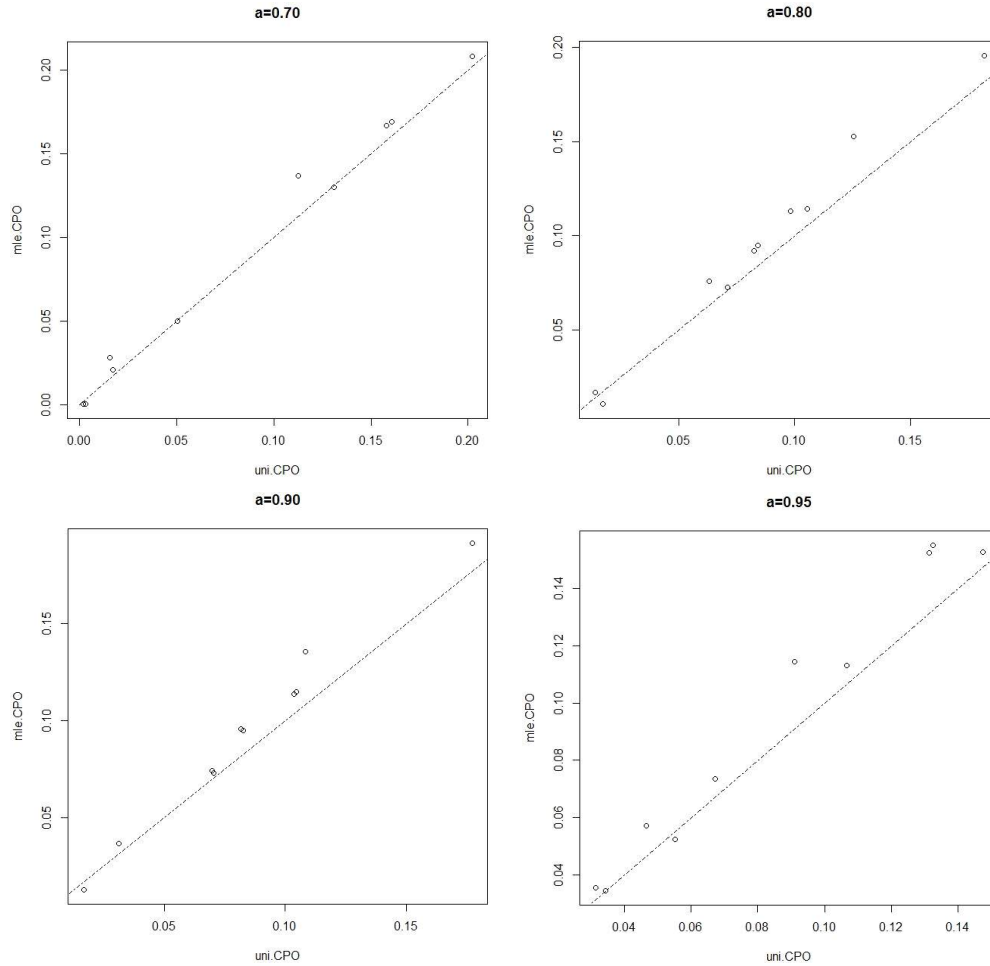


Figure 3.1. A pointwise comparison of CPO values under two different prior specifications

4. Concluding remarks

We have investigated the sensitivity to two different prior specifications on parameters for selection mechanism under Bayesian nonignorable selection models. The data analysis using four simulated data sets reveals that the MLE prior provides better fit than the uniform prior in viewpoints of goodness-of-fit measures. But we need extra work for the specification of the hyperparameters using MLE.

References

- Carlin, B. P. and Louis, T. A. (2009). *Bayesian methods for data analysis*, 3rd Ed., Chapman & Hall/CRC, Boca Raton.
- Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based method (with discussion). In *Bayesian Statistics 4*, edited by J.M. Bernardo, et al., Oxford University Press, Oxford, 147-167.
- Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515-533.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359-378.
- Kwak, S. and Kim, D. (2012). Bayesian estimation for finite population proportions in multinomial data. *Journal of the Korean Data & Information Science Society*, **23**, 587-593.
- Malec, D., Davis, W. W. and Cao, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine*, **18**, 3189-3200.
- Nandram, B., Bhatta, D., Bhadra, D. and Shen, G. (2013). Bayesian predictive inference of a finite population proportion under selection bias. *Statistical Methodology*, **11**, 1-21.
- Nandram, B. and Choi, J. W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, **105**, 120-135.
- Roos, M. and Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, **6**, 259-278.
- Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002). Bayesian measure of model complexity and fit. *Journal of the Royal Statistical Society B*, **64**, 583-616.