Logistic Regression With Variables Subject To Post Randomization Method

Yong Ming Jeffrey Woo [*]**, Aleksandra B. Slavković**

Department of Statistics, The Pennsylvania State University

Abstract

An increase in quality and detail of publicly available databases increases the risk of disclosure of sensitive personal information contained in such databases. The goal of Statistical Disclosure Control (SDC) is to develop methodology that aims at minimizing disclosure risk while providing society with as much information as possible needed for valid statistical inference. The Post Randomization Method (PRAM) is a disclosure avoidance method, where values of categorical variables are perturbed via some known probability mechanism, and only the perturbed data are released thus raising issues regarding disclosure risk and data utility. In this paper, we propose a number of EM algorithms to obtain unbiased estimates of the logistic regression model with data subject to PRAM, and thus effectively account for the effects of PRAM and preserve data utility. The effect of the level of perturbation and sample size on the estimates are evaluated, and relevant standard error estimates are proposed.

Keywords: Statistical disclosure control, logistic regression, generalized linear models, EM algorithm

[*] corresponding author email: yjw102@psu.edu

1 Introduction

The goal of Statistical Disclosure Control (SDC), and related fields such as Privacy-Preserving Data-Mining, is to provide society with as much information as possible while individual information is sufficiently protected against public disclosure. Development of SDC methodology and its practice are important to official statistics for many reasons, a number of which are discussed by Willenborg and de Waal (1996), Fienberg and Slavković (2010), and Ramanayake and Zayatz (2010). For example, there are laws requiring the protection of the confidentiality of respondents' information by statistical offices such as Title 13 of the United States Code that outlines the role of the U.S. Census. Disclosure control is also needed to maintain trust amongst respondents in order to obtain adequate response rates. Concerns about threats to data privacy have even led to the abolition of the census in the past (e.g., Netherlands in 1971; see van der Laan (2000)).

Traditionally, publicly available data have been mostly in the aggregate form, but nowadays there exists substantial demand for high-quality detailed data products. Microdata are sets of records containing detailed information on individual respondents and many SDC methods have been developed for microdata. When SDC methods are applied to a dataset they lead to a publication of altered datasets, which would be available for a wider use and have reduced disclosure risk of sensitive information but also reduced data utility. The Post Randomization Method (PRAM) is a disclosure avoidance method originally proposed by Gouweleeuw et al. (1998). The main idea behind PRAM is to publish redacted data after the values of categorical variables in the original data have been misclassified by a known probability mechanism. This probability mechanism is described by a transition matrix, so called a PRAM matrix. While PRAM provides certain advantages compared to other SDC methodologies (e.g., unlike non-perturbative methods which lead to loss of detail, PRAM maintains detail in the variables; unlike most perturbative methods, the application of PRAM is probabilistic in nature and hence intruders cannot determine which records have been perturbed), it has seen a limited use in practice. One of the commonly-discussed issue in the literature is of model parameter estimation when data are subject to PRAM. Summary statistics from PRAMed data are typically biased and need to be adjusted to take the effects of PRAM into account. Current literature has focused on a specific subset of problems; for example, Gouweleeuw et al. (1998) proposed an unbiased moment estimator for frequency counts; van den Hout and van der Heijden (2002) proposed formulas to estimate odds ratios for data subject to PRAM; and van den Hout and Kooiman (2006) proposed an EM algorithm to estimate the linear regression model with covariates subject to PRAM.

In this paper, we focus on measuring data utility from the standpoint of statistical inference (e.g., see Slavković and Lee (2010)) and propose a way to obtain unbiased estimators of parameters in logistic regression models when data have been subjected to PRAM. We develop and implement EM-type algorithms to obtain asymptotically unbiased estimators, that is the maximum likelihood estimators of parameters in logistic regression models, when variables are subject to PRAM. The basic ideas are based on the "EM by method of weights" developed by Ibrahim (1990) for generalized linear models (GLMs) with covariates missing at random, and on the approach proposed by van den Hout and Kooiman (2006) for linear regression with covariates subject to PRAM. There is an extensive literature for missing data with either missing covariate or missing response variable. We extend these ideas by developing an EM-type algorithm that obtains unbiased estimates of logistic regression when both covariate **and** response variables are subject to PRAM. This is a more difficult problem than either case when covariates or response variables are subject to PRAM, and has received little attention in both PRAM and missing data literatures.

The outline of this paper is as follows. Section 2 presents the EM-type methodology to obtain estimates of the logistic regression model when variables are subject to PRAM and reports the results of simulation studies to evaluate the methodology for three different cases: (1) covariate subject to PRAM, (2) response variable subject to PRAM, and (3) both covariate and response variables subject to PRAM. Section 3 reports the results of simulation studies intended to evaluate the effects of varying the parameters of the logistic regression model on the proposed methodology. Section 4 applies the proposed methodology to data from the 1993 Current Population Survey, and Section 5 contains some discussion.

2 Estimating Logistic Regression Model with Variables Subject to PRAM

We first introduce the notation for the standard logistic regression model. Let $\mathbf{x} = (x_0, x_1, ..., x_p)^t$ denote the vector with p covariates, and let Y be the binary response variable. The logistic regression model is written as

$$E(Y|\mathbf{x}) = \frac{exp(\boldsymbol{\beta}^t\mathbf{x})}{1 + exp(\boldsymbol{\beta}^t\mathbf{x})} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ are the parameter coefficients. We fix $x_0 = 1$ so β_0 is the intercept. Let $\mu(\mathbf{x}) = P(Y = 1|\mathbf{x}, \beta)$, so $1 - \mu(\mathbf{x}) = P(Y = 0|\mathbf{x}, \beta)$.

The likelihood function is given by

$$L(\boldsymbol{\beta}) = \prod_i [\mu(\mathbf{x}_i)]^{y_i}[1 - \mu(\mathbf{x}_i)]^{1-y_i}, \tag{2}$$

the loglikelihood is

$$\ell(\boldsymbol{\beta}) = \sum_i^n \{y_i log[\mu(\mathbf{x}_i)]] + (1 - y_i)log[1 - \mu(\mathbf{x}_i)]]\}, \tag{3}$$

and the information matrix is given by

$$I(\boldsymbol{\beta})_{jk} = -\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k} = \sum_i x_{ij}x_{ik}\mu(\mathbf{x}_i)(1 - \mu(\mathbf{x}_i)). \tag{4}$$

Similar to the ideas in Ibrahim (1990) and van den Hout and Kooiman (2006) for covariates missing at random and for covariates subject to PRAM respectively, we develop and implement three EM-type algorithms to obtain unbiased estimators for parameters in the logistic regression model for three cases: (1) covariates subject to PRAM; (2) response variables subject to PRAM; and (3) both covariates and response variables subject to PRAM. The E-step consists of computing the weights with the derived formulas for the conditional distribution of the true data given the observed data and current parameter estimates. The M-step can be carried out using weighted regression with any standard statistical software, which makes implementation of the algorithms more convenient. Case 3 has not been studied as extensively as Cases (1) and (2) in either PRAM or missing data literature, and is more complex.

2.1 Categorical Covariate Subject to PRAM

Let X denote the categorical covariate to which PRAM is applied, with X^* denoting the observed, released version of X. The levels of X and X^* are $\{x_1, ..., x_J\}$. Let P_X be the J x J PRAM transition matrix that contains the transition probabilities, with $p_{Xjk} = P(x^* = x_k|X = x_j)$, $\pi^*_j = P(X^* = x_j)$ and $\pi_j = P(X = x_j)$.

2.1.1 EM Algorithm I

We present an EM algorithm to obtain unbiased estimates of (1) for Case (1) when covariates are subject to PRAM. This method is similar to the "EM by method of weights" proposed by Ibrahim (1990), which is used to estimate parameters in GLMs with missing covariates. Consider the joint distribution of (x_i, y_i), which can be specified via the conditional distribution of y_i given x_i and the distribution of x_i. Then the complete data loglikelihood can be expressed as

$$\begin{aligned} \ell(\phi; X, y) &= \sum_{i=1}^n \ell(\phi; x_i, y_i) \\ &= \sum_{i=1}^n \{\ell_{y_i|x_i}(\beta) + \ell_{x_i}(\pi)\} \end{aligned} \tag{5}$$

where $\phi = (\beta, \pi)$, and the distribution of X is multinomial with parameter π. The E-step can be written as

$$
\begin{aligned}
Q(\phi|\phi^{(v)}) &= \sum_{i}^{n} E(\ell(\phi; x_i, y_i)|data, \phi^{(v)}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{J} q_j(i)\ell(\phi; x_i, y_i) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{J} q_j(i)\{\ell_{y_i|x_i}(\beta) + \ell_{x_i}(\pi)\} \qquad (6)
\end{aligned}
$$

where $q_j(i)$ is the conditional probability of $X|$observed data, $\phi^{(v)}$ for subject i. The first part of (6) is the loglikelihood of the logistic regression model, and second part is the loglikelihood of a multinomial distribution.

The M-step maximizes (6). This can be done via a weighted logistic regression, by creating a "new" dataset, with each subject i having $(X_{new} = x_1), (X_{new} = x_2), ..., (X_{new} = x_J)$ with weights $q_j(i) = P(X_{new}(i) = x_j|y(i), x^*(i), \phi^{(v)})$, where x^* is the observed value of the PRAMed variable. Using Bayes' rule, the conditional distribution of X_{new} is

$$
P(X_{new} = x_j|X^* = x_k, Y, \phi^{(v)}) = \frac{p_{Xjk}P(Y|x_j, \beta^{(v)})\pi_j^{(v)}}{\sum_{l=1}^{J} p_{Xlk}P(Y|x_l, \beta^{(v)})\pi_l^{(v)}} \qquad (7)
$$

for $k, j \in \{1, ..., J\}$. An example on how the weighted logistic regression is carried out is shown below in Table (1).

Table 1: Methodology: Weighted logistic regression where Y is a response variable, X^* denotes observed value of covariate and X_{new} denotes a list of possible values of the covariate.

Y	X^*	X_{new}	Weight for X_{new}
$y(1)$	$x^*(1)$	0	$q_0(1)$
		1	$q_1(1)$
$y(2)$	$x^*(2)$	0	$q_0(2)$
		1	$q_1(2)$
$\vdots$			
$y(n)$	$x^*(n)$	0	$q_0(n)$
		1	$q_1(n)$

EM Algorithm I runs as follows:

EM Algorithm I: Initial values can be the estimates of β from the logistic regression on $Y \sim X^*$, where X^* is the PRAMed covariate. π^* can be used as the initial estimate of π.

E-step:
Compute $q_j^{(v)}(i)$ for $i = 1, ..., n$ and $j = 1, ..., J$.

M-step:
Carry out weighted logistic regression with weights $q_j^{(v)}(i)$, using standard software.
Update $\phi^{(v)}$
$\beta^{(v+1)} = \hat{\beta}$ from weighted logistic regression
$\pi^{(v+1)} = (q_1^{(v)}(+), ..., q_J^{(v)}(+))^t/n$, where $q_j^{(v)}(+) = \sum_i q_i^{(v)}$

With the updated $\phi^{(v+1)}$, a new dataset with new weights can be computed in the E-step, and the algorithm continues until convergence.

2.1.2 Comparison of Estimation With & Without Accounting for PRAM

To demonstrate the effect of PRAM on maximum likelihood estimates of parameter coefficients in logistic regression model (1), a simulation study was carried out, with one binary covariate, $\boldsymbol{\beta} = (1,2)^t$, x_{i1} sampled from $Bernoulli(0.4)$, and y_i sampled from $Bernoulli(\frac{exp(\boldsymbol{\beta}^t \mathbf{x})}{1+exp(\boldsymbol{\beta}^t \mathbf{x})})$. The estimates for the logistic regression of Y on X are reported in Table 2, labeled as "β_{real}". PRAM is then applied to x to obtain x^*, with the following transition matrix

$$P_X = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix},$$

where we vary the value of p.

Standard logistic regression is performed on y with x^*, while x is not released and is considered to be unobserved. We estimate β_1 and its approximate 95% confidence interval, $\hat{\beta_1} \pm 2SE(\hat{\beta_1})$. We ran 500 simulations, with $n = 100, 1000, 10000$, $p = 0.8, 0.9$. The mean estimates of β_1 and the coverage probabilities of the 95% CIs are computed and are reported in Table 2 in a row labeled "$\beta_{noadjust}$".

Next, using EM Algorithm I, we obtain the MLEs of model (1). The algorithm ran for 20 steps. The mean estimates of β_1 and the coverage probabilities of the 95% CIs are computed and reported in Table 2 in a row labeled "β_{adjust}". The inverse of the observed information matrix can be used to estimate the covariance matrix. Following the method described in Ibrahim et al. (2005) and Louis (1982), the estimated observed information matrix is given by

$$I(\hat{\beta}) = -\ddot{Q}(\hat{\beta}|\hat{\beta}^v) - \{[\sum_{i=1}^{n} \hat{q}_j(i) S_i(\hat{\beta}|x_i, y_i) S_i(\hat{\beta}|x_i, y_i)^t] - \sum_{i=1}^{n} \dot{Q}_i(\hat{\beta}|\hat{\beta}^v) \dot{Q}_i(\hat{\beta}|\hat{\beta}^v)^t\} \tag{8}$$

where $\ddot{Q}(\beta|\beta^v) = \sum_{i=1}^{n} \sum_{j=1}^{J} q_j(i) \frac{\partial^2 \ell(\beta|x_i,y_i)}{\partial\beta\partial\beta^t}$, $\dot{Q}_i(\beta|\beta^v) = \sum_{j=1}^{J} q_j(i) \frac{\partial \ell(\beta|x_i,y_i)}{\partial\beta}$, and $S_i(\beta|x_i, y_i) = \frac{\partial \ell(\beta|x_i,y_i)}{\partial\beta}$.

From Table 2, we can see that estimates of the logistic regression coefficients without accounting for PRAM are biased (see $\beta_{noadjust}$). The bias appears to increase when p decreases (i.e. higher level of perturbation). Actual coverage probabilities also decreases as sample size increases and as p decreases. When accounting for PRAM, the estimates appear to be close to their true values (see β_{adjust}). The estimates are less biased as sample size increases, for example, when $p = 0.9$, the bias is -0.0031, 0.0023 and 0.002 for $n = 100$, $n = 1000$ and $n = 10000$ respectively. Coverage probabilities decreases with higher level of perturbation, and larger sample sizes.

Table 2: Simulated ML estimates of (1) with accounting for PRAM. Average ML estimates. Coverage probabilities in parentheses

		$p = 0.9$	$p = 0.8$
$n = 100$	β_{real}	0.6062	0.6062
	$\beta_{noadjust}$	0.4403 (0.986)	0.2872 (0.872)
	β_{adjust}	0.6093 (1.000)	0.6914 (0.998)
$n = 1000$	β_{real}	0.4947	0.4947
	$\beta_{noadjust}$	0.3810 (0.960)	0.2822 (0.720)
	β_{adjust}	0.4924 (1.000)	0.5031 (0.926)
$n = 10000$	β_{real}	0.4995	0.4995
	$\beta_{noadjust}$	0.3897 (0.350)	0.2872 (0.02)
	β_{adjust}	0.5015 (0.890)	0.5042 (0.790)

Figures 1 and 2 display the plots of the true values of β_1 (in black), along with the estimates of β_1 (in red; left column without adjustment, right column using adjustment with EM algorithm), and the confidence intervals for the estimates via the EM algorithm (in dotted green), for $n = 10000$, $p = 0.90$ and $p = 0.80$ respectively. We only show plots for the first 50 simulations. When not adjusting for PRAM, the true values of β_1 barely fall within the confidence intervals, especially when $p = 0.8$. When the adjustment is made using EM algorithm I, the true values of β_1 are more likely to fall within the confidence intervals. Plots for $n = 100$ and $n = 1000$ are shown in Appendix A, Figures 11 - 14.

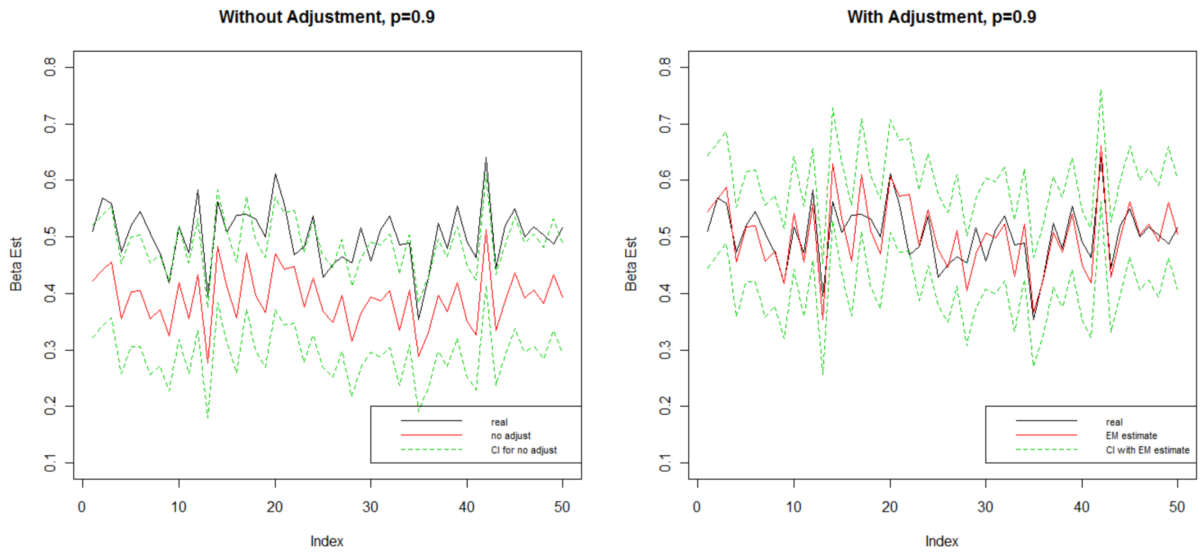

Figure 1: Plots of estimates with covariate subject to PRAM, with 95% CI, when $n = 10000$, $p = 0.90$

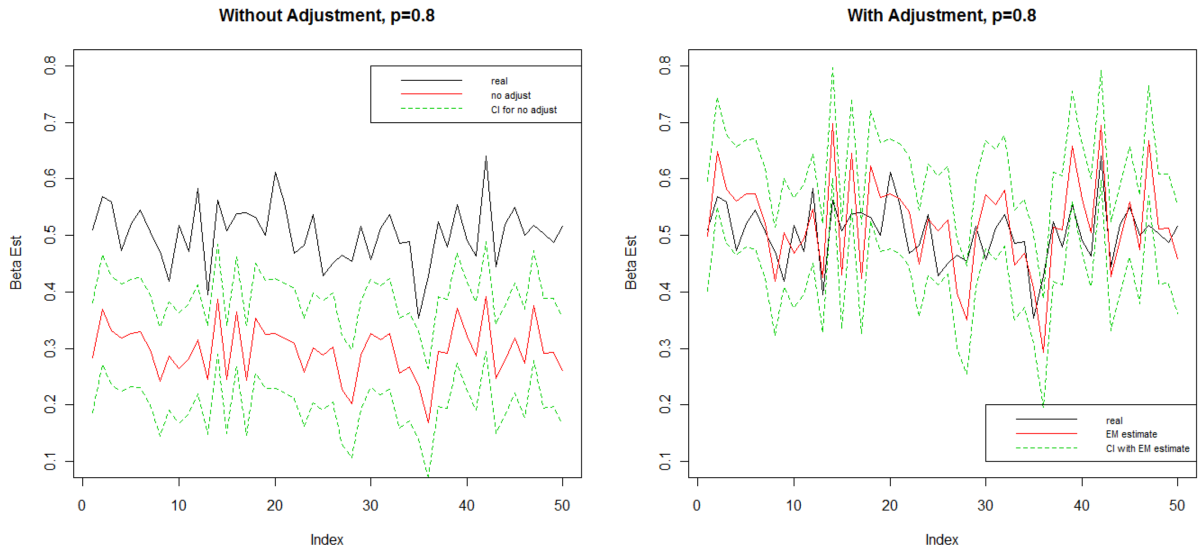

Figure 2: Plots of estimates with covariate subject to PRAM, with 95% CI, when $n = 10000$, $p = 0.80$

Additional simulations to study the effect of the proposed EM algorithm with varying π, β and multinomial covariates are discussed in Section 3. Next, we focus on presenting EM algorithms for cases (2) and (3).

2.2 Response Variable Subject to PRAM

Let Y denote the binary response variable to which PRAM is applied, with Y^* denoting the observed, unreleased version of Y. Let P_Y be the PRAM transition matrix that contains the transition probabilities, with $p_{Yjk} = P(Y^* = k|Y = j), k, j \in \{0, 1\}$.

2.2.1 EM Algorithm II

Following the method proposed in Section 2.1.1, the parameter π in the complete data loglikelihood (5) can be estimated directly since X is not subject to PRAM. Thus, the E-step simplifies to

$$Q(\phi|\phi^{(v)}) \quad = \quad \sum_{i=1}^{n}\sum_{j=1}^{J} r_j(i)\{\ell_{y_i|x_i}(\beta)\} \tag{9}$$

where $r_j(i)$ is the conditional probability of $Y|$observed data, $\beta^{(v)}$ for subject i.

The M-step maximizes (9). This can be done via a weighted logistic regression, by creating a "new" dataset, with each subject i having $(Y_{new} = 0), (Y_{new} = 1)$ with weights $r_j(i) = P(Y_{new}(i) = j|y^*(i), x(i), \beta^{(v)})$. Using Bayes' rule, the weights can be computed as

$$P(Y_{new} = j|Y^* = k, X, \beta^{(v)}) = \frac{p_{Yjk}P(Y = j|X, \beta^{(v)})}{\sum_l p_{Ylk}P(Y = l|X, \beta^{(v)})}. \tag{10}$$

EM Algorithm II runs as follows:

EM Algorithm II: Initial values can be the estimates of β from the logistic regression on $Y^* \sim X$, where Y^* is the PRAMed response variable.

E-step:
Compute $r_j^{(v)}(i)$ for $i = 1, ..., n$ and $j = 0, 1$.

M-step:
Carry out weighted logistic regression with weights $r_j^{(v)}(i)$, using standard software.
Update $\beta^{(v)}$
$\beta^{(v+1)} = \hat{\beta}$ from weighted logistic regression

With the updated $\beta^{(v+1)}$, a new dataset with new weights can be computed in the E-step, and the algorithm continues until convergence.

2.2.2 Comparison of Estimation With & Without Accounting for PRAM

To demonstrate the effect of PRAM on maximum likelihood estimates of parameter coefficients in logistic regression model (1), a simulation study was carried out, with one covariate, $\boldsymbol{\beta} = (1, 0.5)^t$, and x_{i1} sampled from $N \sim (1, 1)$, and y_i sampled from $Bernoulli(\frac{exp(\boldsymbol{\beta}^t\mathbf{x})}{1+exp(\boldsymbol{\beta}^t\mathbf{x})})$. Like before, we report the estimated parameters of the logistic regression of Y on X (see Table 3, row labeled as "β_{real}"). PRAM is then applied to y to obtain y^*, with the following transition matrix

$$P_Y = \left(\begin{array}{cc} p & 1-p \\ 1-p & p \end{array} \right),$$

where we vary the value of p.

Standard logistic regression is performed on y^* with x, while y is not released and is considered to be unobserved. We estimate β_1 and its approximate 95% confidence interval, $\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$. We ran 500 simulations, with

$n = 100, 1000, 10000$, $p = 0.8, 0.9$. The mean estimates of β_1 and the coverage probabilities of the 95% CIs are computed and are reported in Table 3 in a row labeled "$\beta_{noadjust}$".

Next, using EM Algorithm II, we obtain the MLEs of model (1). The algorithm ran for 20 steps. The mean estimates of β_1 and the coverage probabilities of the 95% CIs are computed and reported in Table 3 in a row labeled "β_{adjust}". The inverse of the observed information matrix can be used to estimate the covariance matrix. Note that

$$I(\boldsymbol{\beta}) = \sum_i I(\boldsymbol{\beta})_i$$

where $I(\boldsymbol{\beta})_i$ is the contribution of subject i to the information matrix. Following the method described in Louis (1982)

$$I_{Y^*|X} = I_{Y|X} - I_{Y|Y^*,X}.$$

From (4), it follows that

$$\begin{aligned} I(\boldsymbol{\beta})_i &= \mathbf{x}_i\mathbf{x}_i^t\hat{\mu}(\mathbf{x}_i)(1-\hat{\mu}(\mathbf{x}_i)) \\ &- \mathbf{x}_i\mathbf{x}_i^t[P(Y=1|Y^*,\mathbf{x}_i).(1-P(Y=1|Y^*,\mathbf{x}_i))]. \end{aligned} \tag{11}$$

From Table 3, we can see that estimates of the logistic regression coefficients without accounting for PRAM are biased (see $\beta_{noadjust}$). The bias appears to increase when p decreases. Actual coverage probabilities also decreases as sample size increases and as p decreases. When accounting for PRAM, the estimates appear to be close to their true values (see β_{adjust}). The estimates are least biased for $p = 0.9$ and $n = 10000$. Coverage probabilities decreases with higher level of perturbation, and larger sample sizes.

Table 3: Simulated ML estimates of (1) with accounting for PRAM. Average ML estimates. Coverage probabilities in parentheses

		$p = 0.9$	$p = 0.8$
$n = 100$	β_{real}	0.5361	0.5361
	$\beta_{noadjust}$	0.3550 (0.922)	0.2259 (0.754)
	β_{adjust}	0.5777 (0.970)	0.6077 (0.854)
$n = 1000$	β_{real}	0.4962	0.4962
	$\beta_{noadjust}$	0.3216 (0.464)	0.2117 (0.072)
	β_{adjust}	0.5015 (0.958)	0.5101 (0.746)
$n = 10000$	β_{real}	0.5016	0.5016
	$\beta_{noadjust}$	0.3229 (0.000)	0.2118 (0.000)
	β_{adjust}	0.5007 (0.956)	0.5007 (0.726)

Figures 3 and 4 display the plots of the true values of β_1, along with the estimates of β_1, and the confidence intervals for the estimates via the EM algorithm, for $n = 10000$, $p = 0.90$ and $p = 0.80$ respectively. Similar to the results from Section 2.1.2, the true values of β_1 barely fall within the confidence intervals when not using the EM algorithm. When the adjustment is made using EM algorithm II, the true values of β_1 are more likely to fall within the confidence intervals. Plots for $n = 100$ and $n = 1000$ are shown in Appendix B, Figures 15 - 18.

2.3 Covariate and Response Subject to PRAM

2.3.1 EM Algorithm III

The weighted logistic regression is done by creating a "new" dataset with each subject i having $(Y_{new} = 0)$, $(Y_{new} = 1)$ and $(X_{new} = x_1), (X_{new} = x_2), ..., (X_{new} = x_J)$ with weights $w_{ml}(i) = P(Y_{new}(i) = m, X_{new} = x_l|Y^*(i) = k, X^*(i) = x_j, \beta^{(v)})$. The weights can be computed as

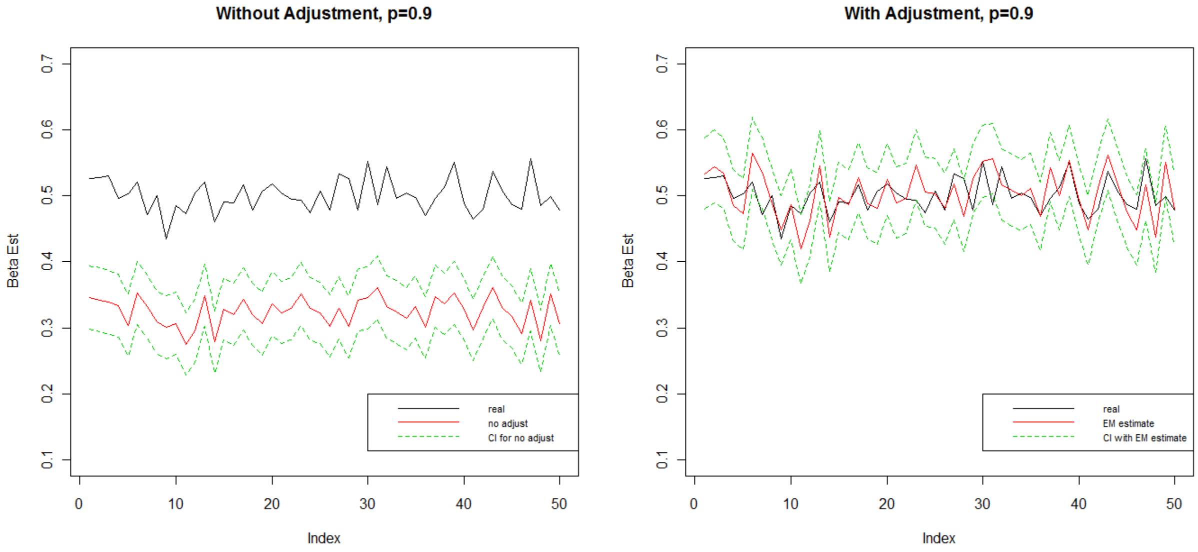

Figure 3: Plots of estimates with response subject to PRAM, with 95% CI, when $n = 10000$, $p = 0.90$

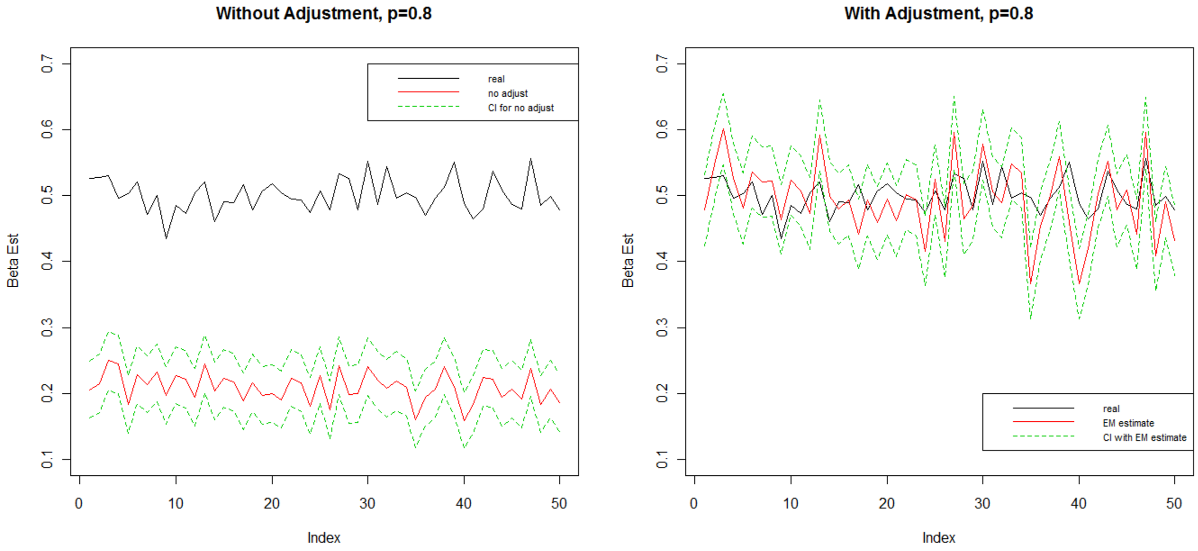

Figure 4: Plots of estimates with response subject to PRAM, with 95% CI, when $n = 10000$, $p = 0.80$

$$w_{ml} = \frac{p_{Ymk} P(Y = m|X = l, \beta^{(v)})}{\sum_a p_{Yak} P(Y = a|X = l, \beta^{(v)})} \frac{p_{Xlj} \pi(l) \sum_b p_{Ybk} P(Y = b|X = l, \beta^{(v)})}{\sum_c p_{Xcj} \pi(c) \sum_d p_{Ydk} P(Y = d|X = c, \beta^{(v)})}. \tag{12}$$

The weights (12) are more difficult to derive mathematically than the weights (7) and (10). See Appendix C for derivations of the weights.

EM Algorithm III runs as follows:

EM Algorithm III: Initial values can be the estimates of β from the logistic regression on $Y^* \sim X^*$, where Y^* and X^* are the PRAMed response variable and covariate. π can be estimated by $\hat{\pi} = (P_X^{-1})^t \pi^*$.

E-step:
Compute $w_j^{(v)}(i)$ for $i = 1, ..., n$ and $j = 0, 1$.

M-step:
Carry out weighted logistic regression with weights $w_j^{(v)}(i)$, using standard software.
Update $\beta^{(v)}$
$\beta^{(v+1)} = \hat{\beta}$ from weighted logistic regression

With the updated $\beta^{(v+1)}$, a new dataset with new weights can be computed in the E-step, and the algorithm continues until convergence.

2.3.2 Comparison of Estimation With & Without Accounting for PRAM

To demonstrate the effect of PRAM on maximum likelihood estimates of parameter coefficients in logistic regression model (1), a simulation study was carried out, with one binary covariate, $\boldsymbol{\beta} = (1, 2)^t$, and x_{i1} sampled from $Bernoulli(0.4)$, and y_i sampled from $Bernoulli(\frac{exp(\boldsymbol{\beta}^t \mathbf{x})}{1+exp(\boldsymbol{\beta}^t \mathbf{x})})$. The parameter estimates of the logistic regression of Y on X are reported in Table 4, labeled as "β_{real}". PRAM is then applied to y and x to obtain y^* and x^*, with the following transition matrix

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

Standard logistic regression is performed on y^* with x^*, while y and x are not released and are considered to be unobserved. We estimate β_1 and its approximate 95% confidence interval, $\hat{\beta_1} \pm 2SE(\hat{\beta_1})$. We ran 500 simulations, with $n = 100, 1000, 10000$. The mean estimates of β_1 and the coverage probabilities of the 95% CIs are computed and are reported in Table 4 in a row labeled "$\beta_{noadjust}$".

Next, using EM Algorithm III, we obtain the MLEs of model (1). The algorithm ran for 20 steps. The mean estimates of β_1 and the coverage probabilities of the 95% CIs are computed and reported in Table 4 in a row labeled "β_{adjust}". From Table 4, we can see that estimates of the logistic regression coefficients without accounting for PRAM are biased (see $\beta_{noadjust}$). Similar to the results from the previous sections, the bias appears to increase when p decreases. Actual coverage probabilities also decreases as sample size increases and as p decreases. When accounting for PRAM, the estimates appear to be close to their true values (see β_{adjust}). The estimates are less biased as sample size increases, for example, the bias is 1.0045, -0.0687 and 0.025 for $n = 100$, $n = 1000$ and $n = 10000$ respectively. Coverage probabilities decreases with higher level of perturbation, and larger sample sizes.

Table 4: Simulated ML estimates of (1) with accounting for PRAM. Average ML estimates. Coverage probabilities in parentheses

$n = 100$	β_{real}	4.4725
	$\beta_{noadjust}$	0.8872 (0.516)
	β_{adjust}	3.4680 (1.000)
$n = 1000$	β_{real}	2.0302
	$\beta_{noadjust}$	0.80912 (0.000)
	β_{adjust}	2.0989 (0.922)
$n = 10000$	β_{real}	1.9997
	$\beta_{noadjust}$	0.8003 (0.000)
	β_{adjust}	1.9747 (0.822)

Figure 5 displays the plots of the true values of β_1, along with the estimates of β_1, and the confidence intervals for the estimates via the EM algorithm, for $n = 10000$, $p = 0.90$. Similar to the previous results, the true values of β_1 fall outside the confidence intervals when the EM algorithm is not used. When the adjustment is made using EM algorithm III, the true values of β_1 are more likely to fall within the confidence intervals.

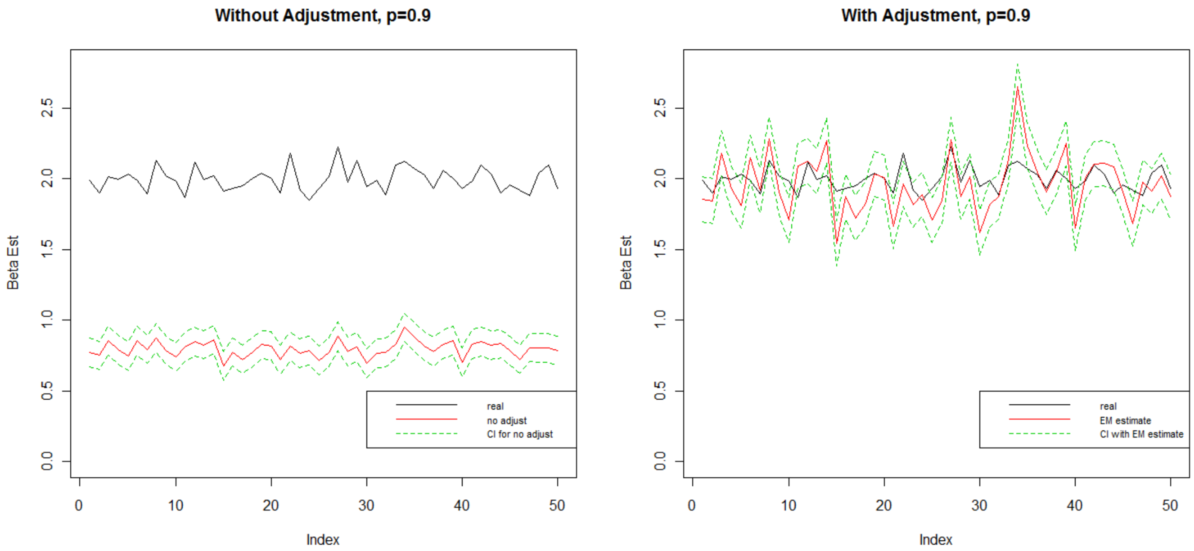

Figure 5: Plots of estimates with response and covariate subject to PRAM, with 95% CI, when $n = 10000$

3 Additional Simulations

In this section, we present additional results on performance of the proposed algorithms subject to varying probabilities of success for a binary response variable, varying distribution of a binary covariate, and perturbing multinomial rather than a binary covariate.

3.1 Adjusting Proportions of Covariate and Regression Parameters

Another simulation study was carried out, with one binary covariate, $\boldsymbol{\beta} = (1, \beta)^t$, x_{i1} sampled from $Bernoulli(\pi)$, and y_i sampled from $Bernoulli(\frac{exp(\boldsymbol{\beta}^t\mathbf{x})}{1+exp(\boldsymbol{\beta}^t\mathbf{x})})$ with varying the values of β and π to assess their effect on EM Algorithm I. The following values were used: $\beta = (-2, -0.5, 0.5, 2)$, and $\pi = (0.1, 0.2, 0.3, 0.4, 0.5)$. PRAM is then applied to x to obtain x^*, with the following transition matrix

$$P_X = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix},$$

where we vary the value of p.

Next, using EM Algorithm I, we obtain the MLEs of model (1). The algorithm ran for 20 steps. The mean estimates of β_1 and the coverage probabilities of the 95% CIs are computed and reported in Table 5. We ran 500 simulations. The results are displayed in Table 5 for $p = 0.9$ and $p = 0.8$. In terms of bias and coverage probabilities, the algorithm gives better estimates when the value of π goes closer to 0.5, when $\beta = -0.5, 0.5$, as well as when $p = 0.9$. For example, when $p = 0.9$, we get high probability coverages (greater than 0.95) when $\pi = 0.4, 0.5$ and $\beta = -0.5, 0.5$. This suggests that the EM algorithm works better when the distribution of the response variable and covariate is not skewed.

Table 5: Simulated ML estimates of (1) with accounting for PRAM, coverage probabilities in parentheses.

		β			
π	p	-2	-0.5	0.5	2
0.1	0.9	-2.0766 (0.760)	-0.5000 (0.898)	0.5785 (0.828)	2.4087 (0.630)
	0.8	-1.819038 (0.622)	-0.4564 (0.690)	0.5514 (0.642)	1.6592 (0.594)
0.2	0.9	-2.0344 (0.910)	-0.5070 (0.982)	0.5283 (0.954)	2.2101 (0.724)
	0.8	-1.9753 (0.700)	-0.5128 (0.760)	0.5483 (0.692)	1.9494 (0.640)
0.3	0.9	-2.0201 (0.948)	-0.51071 (0.996)	0.5051 (0.970)	2.1480 (0.788)
	0.8	-1.9908 (0.742)	-0.5159 (0.856)	0.5281 (0.780)	2.1353 (0.618)
0.4	0.9	-2.0174 (0.976)	-0.5037 (0.990)	0.5071 (0.986)	2.0640 (0.882)
	0.8	-2.0213 (0.784)	-0.5050 (0.866)	0.5213 (0.806)	2.1020 (0.642)
0.5	0.9	-2.0214 (0.982)	-0.5165 (0.994)	0.5131 (0.992)	2.0572(0.936)
	0.8	-2.0110 (0.822)	-0.5127 (0.844)	0.5216 (0.852)	2.1020 (0.716)

3.2 Covariate with More than Two Levels

Another simulation study was carried out, with one categorical covariate with 3 levels, $\boldsymbol{\beta} = (1, -1, 1)^t$, x_{i1} sampled from $multinomial(0.4, 0.3, 0.3)$, and y_i sampled from $Bernoulli(\frac{exp(\boldsymbol{\beta}^t \mathbf{x})}{1+exp(\boldsymbol{\beta}^t \mathbf{x})})$. The estimates of β_1 and β_2 from the logistic regression of Y on X are reported in Table 6, labeled as "$\beta_{i,real}$". PRAM is then applied to y to obtain y^*, with the following transition matrix

$$P_Y = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix},$$

where we vary the value of p.

Standard logistic regression is performed on y^* with x, while y is not released and is considered to be unobserved. We estimate β_1 and $\beta_2$and their approximate 95% confidence interval, $\hat{\beta}_i \pm 2SE(\hat{\beta}_i)$. We ran 500 simulations, with $n = 100, 1000, 10000$, $p = 0.8, 0.9$. The mean estimates of β_1 and β_2 and the coverage probabilities of the 95% CIs are computed and are reported in Table 6 in a row labeled "$\beta_{i,noadjust}$".

Next, using EM Algorithm II, we obtain the MLEs of model (1). The algorithm ran for 20 steps. The mean estimates of β_1 and β_2 and the coverage probabilities of the 95% CIs are computed and reported in Table 6 in a row labeled "$\beta_{i,adjust}$". Similar to the previous results, the estimates from the algorithm are less biased as sample size increases, for both β_1 and β_2. For example, for $p = 0.9$, the bias for β_2 is -0.1909, -0.0196, -0.0035 for $n = 100$, $n = 1000$ and $n = 10000$ respectively.

Table 6: Simulated ML estimates of β_1 & β_2 for(1) with accounting for PRAM. Average ML estimates. Coverage probabilities in parentheses.

$n = 100$	$\beta_{1,real}$	-1.0378	$\beta_{2,real}$	1.5883
	$\beta_{1,noadjust}$	-0.7762 (0.974)	$\beta_{2,noadjust}$	0.7781 (0.894)
	$\beta_{1,adjust}$	-1.0419 (0.972)	$\beta_{2,adjust}$	1.7792 (0.850)
$n = 1000$	$\beta_{1,real}$	-1.0043	$\beta_{2,real}$	1.0137
	$\beta_{1,noadjust}$	-0.7740 (0.800)	$\beta_{2,noadjust}$	0.6513 (0.502)
	$\beta_{1,adjust}$	-0.9988 (0.986)	$\beta_{2,adjust}$	1.0333 (0.944)
$n = 10000$	$\beta_{1,real}$	-0.9970	$\beta_{2,real}$	1.0024
	$\beta_{1,noadjust}$	-0.7739 (0.000)	$\beta_{2,noadjust}$	0.6424 (0.000)
	$\beta_{1,adjust}$	-0.9971 (0.984)	$\beta_{2,adjust}$	1.0059 (0.952)

Figure 6 displays the plots of the true values of β_1, along with the estimates of β_1 and the confidence intervals for the estimates via the EM algorithm for $n = 10000$, $p = 0.90$. Similar to the previous results, the algorithm works well since the true values of β_1 are likely to fall in the confidence intervals for the estimates from the EM algorithm.

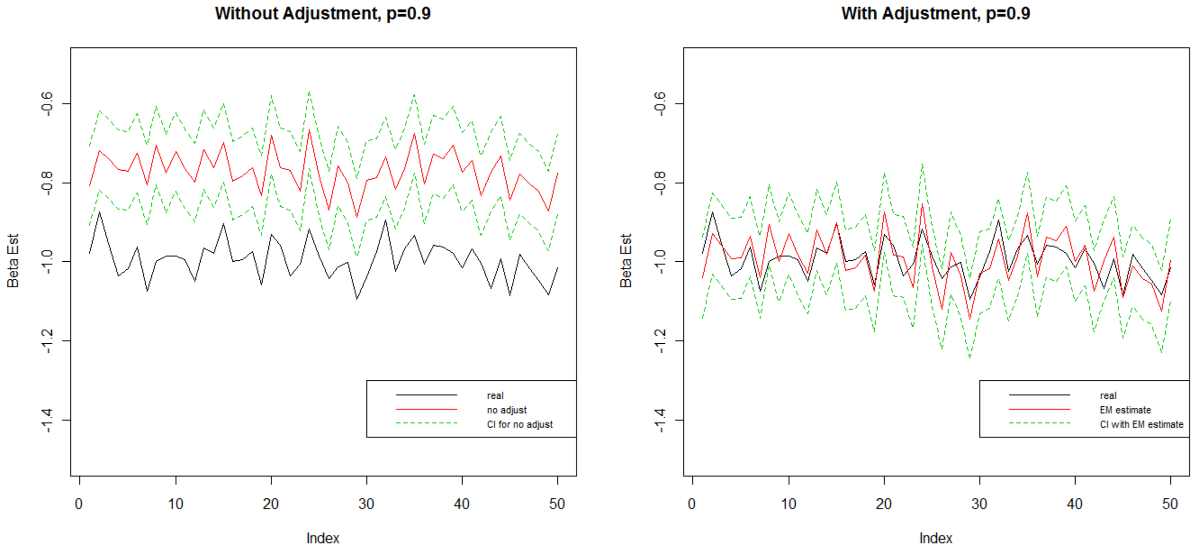

Figure 6: Plots of estimates with covariate subject to PRAM, with 95% CI, when $n = 10000$, $p = 0.90$

In the next simulation study, we had one categorical covariate with 3 levels, $\boldsymbol{\beta} = (1, -1, 1)^t$, x_{i1} sampled from $multinomial(0.4, 0.3, 0.3)$, and y_i sampled from $Bernoulli(\frac{exp(\boldsymbol{\beta}^t \mathbf{x})}{1+exp(\boldsymbol{\beta}^t \mathbf{x})})$. The estimate of β_1 and β_2 from the logistic regression of Y on X are reported in Table 7, labeled as "$\beta_{i,real}$". PRAM is then applied to x to obtain x^*, with the following transition matrix

$$P_X = \begin{pmatrix} p & \frac{1-p}{2} & \frac{1-p}{2} \\ \frac{1-p}{2} & p & \frac{1-p}{2} \\ \frac{1-p}{2} & \frac{1-p}{2} & p \end{pmatrix},$$

where we vary the value of p.

Standard logistic regression is performed on y with x^*, while x is not released and is considered to be unobserved. We estimate β_1 and β_2 and their approximate 95% confidence interval, $\hat{\beta}_i \pm 2SE(\hat{\beta}_i)$. We ran 500 simulations, with $n = 100, 1000, 10000$, $p = 0.8, 0.9$. The mean estimates of β_1 and β_2 and the coverage probabilities of the 95% CIs are computed and are reported in Table 7 in a row labeled "$\beta_{i,noadjust}$".

Next, using EM Algorithm I, we obtain the MLEs of model (1). The algorithm ran for 20 steps. The mean estimates of β_1 and β_2 and the coverage probabilities of the 95% CIs are computed and reported in Table 7 in a row labeled "$\beta_{i,adjust}$". Similar to the previous results, the estimates from the algorithm are less biased as sample size increases, for both β_1 and β_2.

Figure 7 displays the plots of the true values of β_1, along with the estimates of β_1 and the confidence intervals for the estimates via the EM algorithm for $n = 10000$, $p = 0.90$. Similar to the previous results, the algorithm appears to works well since the true values of β_1 are likely to fall in the confidence intervals for the estimates from the EM algorithm.

4 Application to 1993 CPS Dataset

We implement the methodology described in Section 2 on data from the 1993 Current Population Survey (CPS). The dataset contains 48842 records on 8 categorical variables. We perform logistic regression for *salary* ($0 =$ <\$50,000 or $1 =$ >\$50,000) on the covariates *sex* ($0 =$ Female or $1 =$ Male), *race* ($0 =$ Non White or $1 =$ White), and *marital*

Table 7: Simulated ML estimates of β_1 & β_2 for(1) with accounting for PRAM. Average ML estimates. Coverage probabilities in parentheses.

$n = 100$	$\beta_{1,real}$	-1.0378	$\beta_{2,real}$	1.5883
	$\beta_{1,noadjust}$	-0.7762 (0.974)	$\beta_{2,noadjust}$	0.7781 (0.894)
	$\beta_{1,adjust}$	-1.0419 (0.972)	$\beta_{2,adjust}$	1.7792 (0.850)
$n = 1000$	$\beta_{1,real}$	-1.0043	$\beta_{2,real}$	1.0137
	$\beta_{1,noadjust}$	-0.7740 (0.800)	$\beta_{2,noadjust}$	0.6513 (0.502)
	$\beta_{1,adjust}$	-0.9988 (0.986)	$\beta_{2,adjust}$	1.0333 (0.944)
$n = 10000$	$\beta_{1,real}$	-1.0034	$\beta_{2,real}$	1.0021
	$\beta_{1,noadjust}$	-0.8515 (0.038)	$\beta_{2,noadjust}$	0.7706 (0.006)
	$\beta_{1,adjust}$	-1.0040 (0.996)	$\beta_{2,adjust}$	1.0038 (0.980)

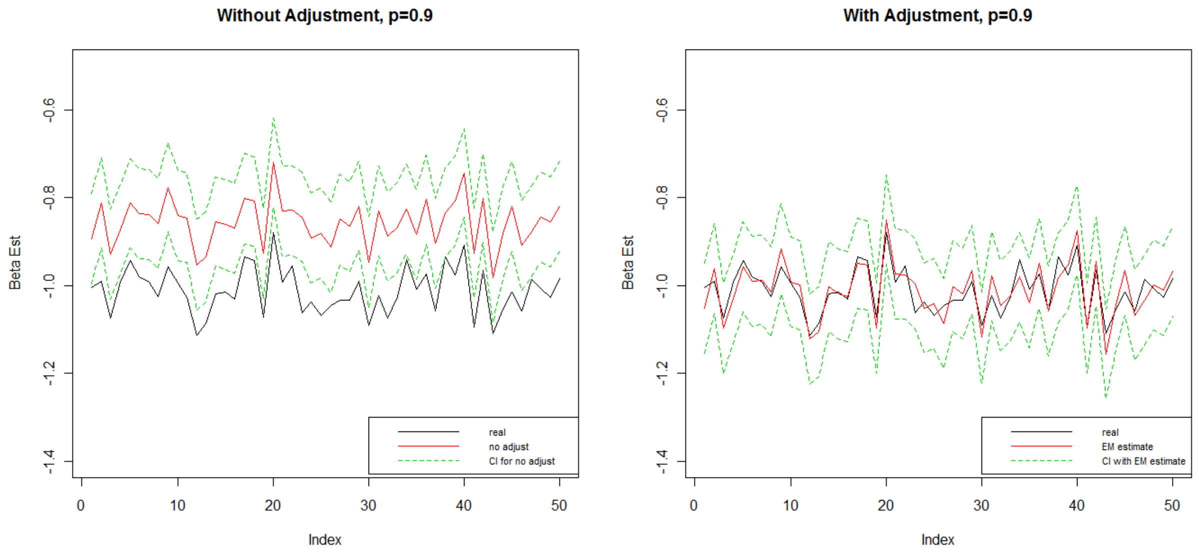

Figure 7: Plots of estimates with covariate subject to PRAM, with 95% CI, when $n = 10000$, $p = 0.90$

(0 = Married or 1 = Unmarried). The parameter estimates from fitting the logistic regression with the original data are displayed in the first line of Table 8, labeled as "O.D.".

We consider the following three cases: 1) *marital* subject to PRAM; 2) *salary* subject to PRAM; and 3) both *marital* and *salary* subject to PRAM. In each case, the following PRAM matrix was applied to the variables that were subject to PRAM

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

In each case, standard logistic regression is performed on the PRAMed data. We estimate β_i and their approximate 95% confidence intervals, $\hat{\beta}_i \pm 2SE(\hat{\beta}_i)$. We ran 500 simulations for each case. The mean estimates of β_i and the coverage probabilities of the 95% CIs are computed and are reported in Table 8 in the rows labeled "without".

Next, using the EM Algorithms I, II and III, we obtain the MLEs of model (1). The algorithm ran for 20 steps. The mean estimates of β_i and the coverage probabilities of the 95% CIs are computed and reported in Table 8 in the rows labeled "with". In terms of bias, EM Algorithm II appears to work best, followed by EM Algorithm I and EM Algorithm III. EM Algorithm III being least effective is no surprise, since PRAM is applied to more variables. For example, the bias of $\hat{\beta}_3$ is -0.0604, 0.0116, -0.1303 for algorithm I, II and III respectively.

Table 8: Parameter Estimates from Original Data (O.D.), PRAMed Data without EM algorithm, and PRAMed data with EM Algorithm. Average ML estimates. Coverage probabilities in parentheses Case 1: *marital* subject to PRAM; Case 2: *salary* subject to PRAM; Case 3: both *marital* and *salary* subject to PRAM

		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
	O.D.	-0.8585	0.2855	0.3925	-2.3166
Case 1	without	-1.4458 (0)	0.7398 (0)	0.4400 (0)	-1.6009 (0)
	with	-1.047738 (0.290)	0.3851 (0.124)	0.4249 (0.734)	-2.2562 (0.510)
Case 2	without	-0.5475 (0)	0.1550 (0)	0.2323 (0)	-1.4539(0)
	with	-0.7785 (0.592)	0.2138 (0.412)	0.3745 (0.946)	-2.3282 (0.842)
Case 3	without	-1.2807 (0)	0.7283 (0)	0.2721 (0)	-1.0581 (0)
	with	-1.2469 (0.128)	0.4372 (0.098)	0.444 (0.468)	-2.1863 (0.262)

Figures 8, 9 and 10 display the plots of the estimate of β_3 on the original data, along with the estimates of β_3, and the confidence intervals for the estimates via the EM algorithm, for case 1, case 2 and case 3 respectively. When not adjusting for PRAM, the estimates of β_3 from the original logistic regression fall outside the confidence intervals. Indeed, for case 3, the estimates for β_3 fall outside the range of the plot. When the adjustment is made using the EM algorithm, the value of β_3 from the original logistic regression is more likely to fall within the confidence intervals.

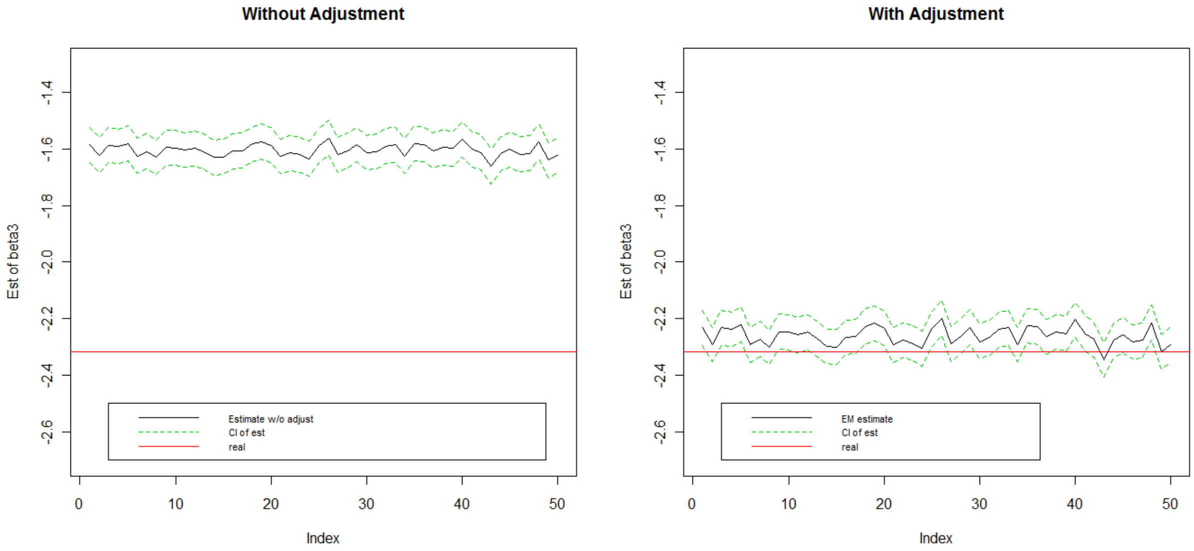

Figure 8: Plots of β_3 with CPS data, Case 1: *marital* subject to PRAM

4.1 Disclosure Risk Assessment

We need to evaluate both disclosure risk and data utility after the application of SDC methodology. This concept is illustrated via a risk-utility (R-U) map (for more details, see Duncan and Pearson (1991)).

This paper focuses on preserving data utility of microdata subject to PRAM when fitting a logistic regression on the PRAMed data. We now briefly discuss disclosure risk assessment and how a PRAM matrix may be chosen. Ideally, the chosen PRAM matrix is one that maximizes data utility under some predetermined levels of disclosure risk

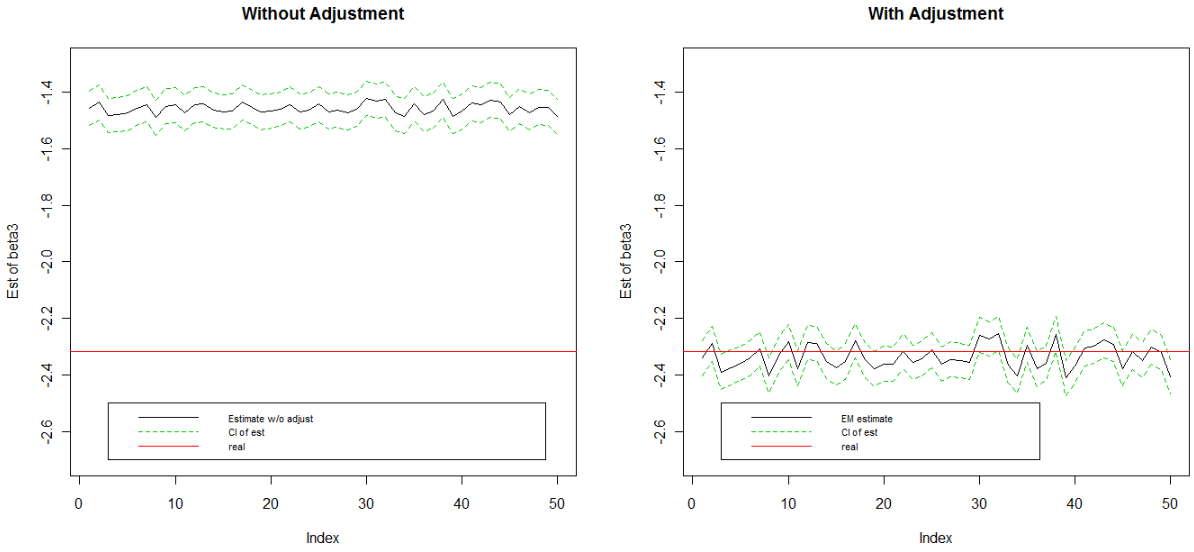

Figure 9: Plots of β_3 with CPS data, Case 2: *salary* subject to PRAM

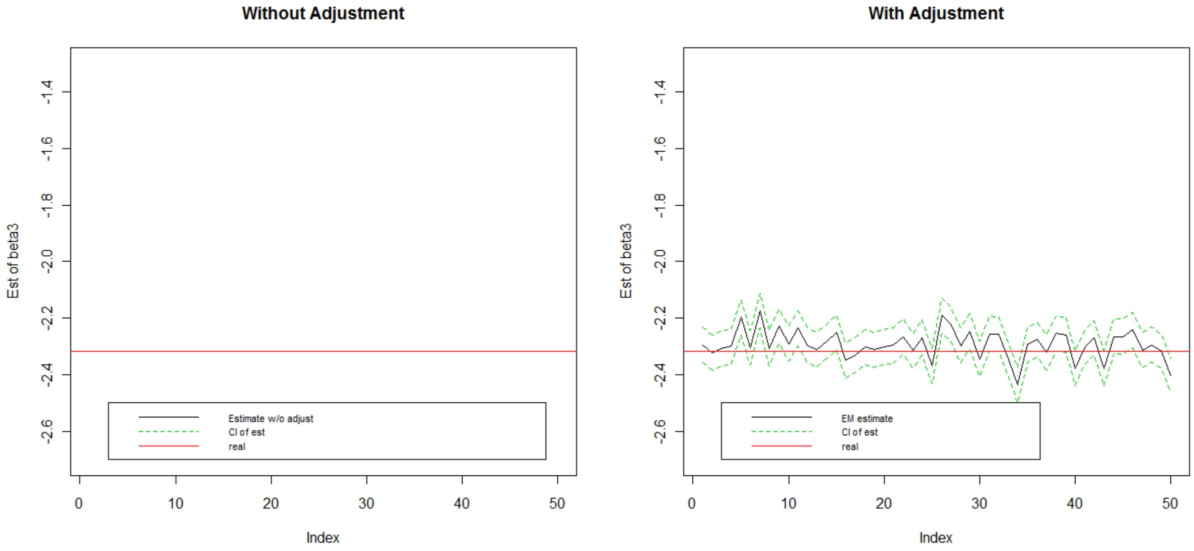

Figure 10: Plots of β_3 with CPS data, Case 3: *marital* and *salary* subject to PRAM

set by the statistical agency.

Since the estimates from the EM algorithm converge to the maximum likelihood estimates, data utility is preserved for all choices of PRAM matrices. Thus, the next step is to find a PRAM matrix that satisfies some level of disclosure control or risk set by the statistical agency. A traditional measure of disclosure risk was proposed by de Wolf and van Gelder (2004), which involves calculating the conditional probability that given a score k in the perturbed file, the original score was k as well, $P(X = k|X^* = k)$. In the context of PRAM, this conditional probability can be estimated by

$$\hat{R}_{PRAM}(k) = P(X = k|X^* = k) = \frac{p_{kk}T_X(k)}{\sum_l p_{lk}T_X(l)}, \tag{13}$$

where $T_X(l)$ are the frequency counts in the sample for score l. The numerator of (13) estimates the number of scores k in the original file that will remain as k in the perturbed file, and the denominator estimates the number of scores k in the original file that remain as k in the perturbed file plus the number of scores that were not k in the original file that take on the score k in the perturbed file. According to a traditional threshold rule, a record is safe when its score occurs more than a certain threshold d, so a safe record can be linked with at least d records (d to be determined by statistical office or rules). de Wolf and van Gelder (2004) suggest that a record is safe whenever

$$\hat{R}_{PRAM}(k) \leq \frac{T_\xi(k)}{d} \tag{14}$$

We use case 1, *marital* subject to PRAM as an example. The three-way table of counts and $\hat{R}_{PRAM}(k)$ for *marital*, *sex*, and *race* is displayed in Table 9. $\hat{R}_{PRAM}(k)$ for married females (both non-white and white) is much lower than unmarried females (non-white and white). This is expected since the number of unmarried females is greater than the number of married females in this example. Thus, even though both unmarried and married females had a 10% probability of being misclassified as married and unmarried females, respectively, the actual number of unmarried females misclassified as married females is much higher than married females being misclassified. This leads to an observation that a higher proportion of married females in the perturbed file were originally unmarried females in the original file. Depending on the disclosure rules set by the agency, such values of $\hat{R}_{PRAM}(k)$ may provide sufficient disclosure control. For example, if the agency decides that the threshold $d = 800$, $\hat{R}_{PRAM}(k)$ is less than $\frac{T_\xi(k)}{d}$ for all k, satisfying (14).

Table 9: Three Way Table for Marital, Sex and Race. $\hat{R}_{PRAM}(k)$, $\frac{T_\xi(k)}{d}$ in parentheses

Married		
	Non-White	White
Female	521 (0.6394, 0.65)	2288 (0.6572, 2.86)
Male	1990 (0.9029, 2.49)	18245 (0.9400, 22.81)

Unmarried		
	Non-White	White
Female	2644 (0.9786, 3.31)	10739 (0.9769, 13.42)
Male	1925 (0.8970, 2.41)	10490 (0.8380, 13.11)

5 Discussion

With increased computing power and built-in statistical packages, many researchers nowadays prefer to work with microdata instead of aggregated data. However, releasing microdata instead of aggregated data increases the risk of disclosure. The goal of SDC for microdata is that given an original microdata set **V**, a protected microdata set **V'** is released in its place so that disclosure risk is low and data utility is high. Most measures of data utility measure the distortion in the distribution in **V'**, when compared to the distribution in **V**. For examples, see Domingo-Ferrer and Torra (2001). In our examples, we propose comparing the coefficient estimates when fitting the regression model on the original data, on the data that has been subject to PRAM, and using the EM-type methodology on the data that has been subject to PRAM. In general, the estimates are biased when fitting the logistic regression model on the data that has been subject to PRAM; using the proposed EM-type methodology, we can obtain estimates that are much closer to their true values and asymptotically unbiased. The estimates are less biased with larger sample sizes, and when the distribution of the response variable and covariates are less skewed. In terms of coverage probabilities, the algorithm appears to work well since most of the coverage probabilities in our examples are high. The next step in this research is to extend these methodologies to handle a wider range of analyses including other models in the class of generalized linear models (GLMs).

Shlomo and Skinner (2010) claim that combining sampling with a perturbation method like PRAM offers greater protection than using either method on its own. PRAM itself guarantees ϵ-differential privacy (see Dwork (2006)), as long as the PRAM matrix does not contain zero elements. We could next evaluate the effect of using both sampling and PRAM on the regression coefficient estimates using our EM-algorithm. We can compare the estimates after both sampling and PRAM has been applied to a dataset.

There are other SDC methodologies that can be applied to categorical variables in microdata. Data swapping is used by agencies like the U.S. Census Bureau, and their approach guarantees that marginals involving the matching variables remain the same. However, the effect on regression analysis is ambiguous(e.g., see Fienberg and McIntyre (2004)). Reiter (2005) carried out an empirical study using fully synthetic data with the 2000 Current Population Study, and found that the coverage probabilities for the logistic regression are extremely low. An interesting next step would be to compare performance of synthetic data methodology to our proposed EM algorithms for PRAM.

References

P. de Wolf and I. van Gelder. An Empirical Evaluation of PRAM. Technical Report Discussion Paper 04012, Statistics Netherlands, Voorburg/Heerlen, September 2004.

J. Domingo-Ferrer and V. Torra. Disclosure Control Methods and Information Loss for Microdata. In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, chapter 5, pages 91–110. North-Holland: Elsevier, 2001.

G.T. Duncan and R.W. Pearson. Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future. *Statistical Science*, 6:219–239, 1991.

C. Dwork. Differential Privacy. In *International Colloquium on Automata, Languages and Programming (ICALP)*, pages 1–12. Springer, 2006.

S.E. Fienberg and J. McIntyre. Data Swapping: Variations on a Theme by Dalenius and Reiss. In *Privacy in Statistical Databases*, 3050, pages 14–29. Springer Berlin Heidelberg, 2004.

S.E. Fienberg and A.B. Slavković. *Data Privacy and Confidentiality*. International Encyclopedia of Statistical Science. Springer-Verlag, 2010.

J. Gouweleeuw, P. Kooiman, L. Willenborg, and P. de Wolf. Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*, 14(4):463–478, 1998.

J. Ibrahim. Incomplete Data in Generalized Linear Models. *Journal of the American Statistical Association*, 85(411): 765–769, 1990.

J. Ibrahim, M.H. Chen, S.R Lipsitz, and A.H. Herring. Missing-Data Methods for Generalized Linear Models: A Comparitive Review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

T.A. Louis. Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society*, 44(2):226–233, 1982.

A. Ramanayake and L. Zayatz. Balancing Disclosure Risk with Data Quality. Statistical Research Division Research Report Series 2010-04, U.S. Census Bureau, 2010.

J.P. Reiter. Releasing Multiply-Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. *Journal of the Royal Statistical Society*, 168:185–205, 2005.

N. Shlomo and C.J. Skinner. Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. *Annals of Applied Statistics*, 4(3):1291–1310, 2010.

A.B. Slavković and J. Lee. Synthetic Two-Way Contingency Tables that Preserve Conditional Frequencies. *Statistical Methodology*, 7:225–239, 2010.

A. van den Hout and P. Kooiman. Estimating the Linear Regression Model with Categorical Covariates Subject to Randomized Response. *Computational Statistics & Data Analysis*, 50:3311–3323, 2006.

A. van den Hout and P. van der Heijden. Randomized Response, Statistical Disclosure Control and Misclassification: A Review. *International Statistical Review*, 70(2):269–288, 2002.

P. van der Laan. The 2001 Census in the Netherlands: Integration of Registers and Surveys. In *The Census of Population: 2000 and Beyond*, Manchester, UK, 2000. Cathie Marsh Centre for Census and Survey Research, Faculty of Economics and Social Studies, University of Manchester.

L. Willenborg and T. de Waal. *Statistical Disclosure Control in Practice*. Springer, New York, NY, 1996.

Appendices

A More Figures from Section 2.1.2

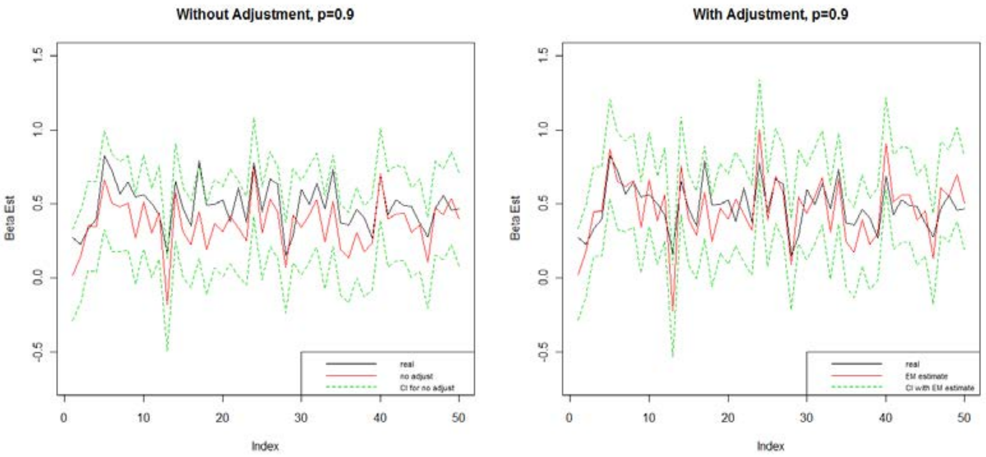

Figure 11: Plots of estimates with covariate subject to PRAM, with 95% CI, when $n = 1000$, $p = 0.90$

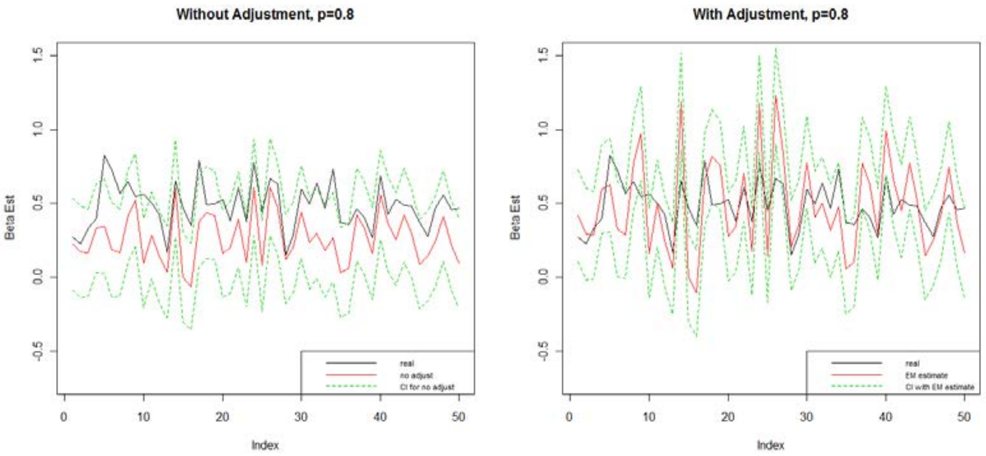

Figure 12: Plots of estimates with covariate subject to PRAM, with 95% CI, when $n = 1000$, $p = 0.80$

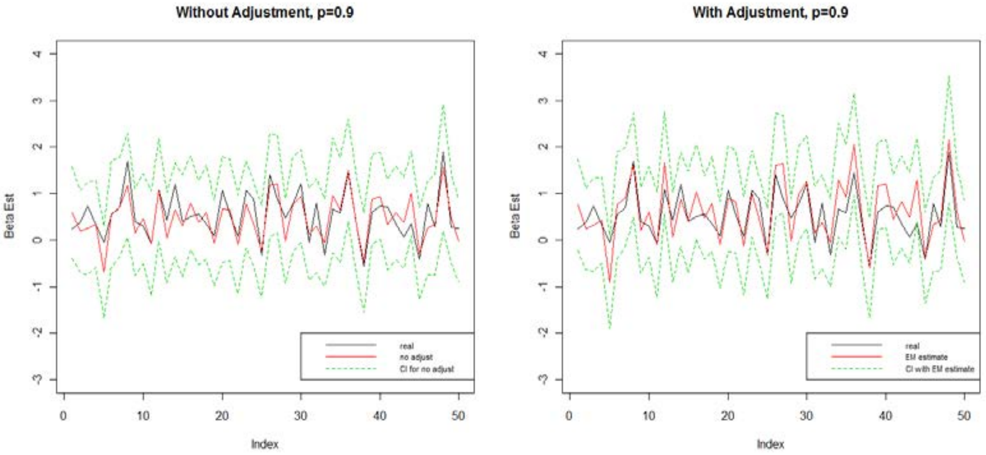

Figure 13: Plots of estimates with covariate subject to PRAM, with 95% CI, when $n = 100$, $p = 0.90$

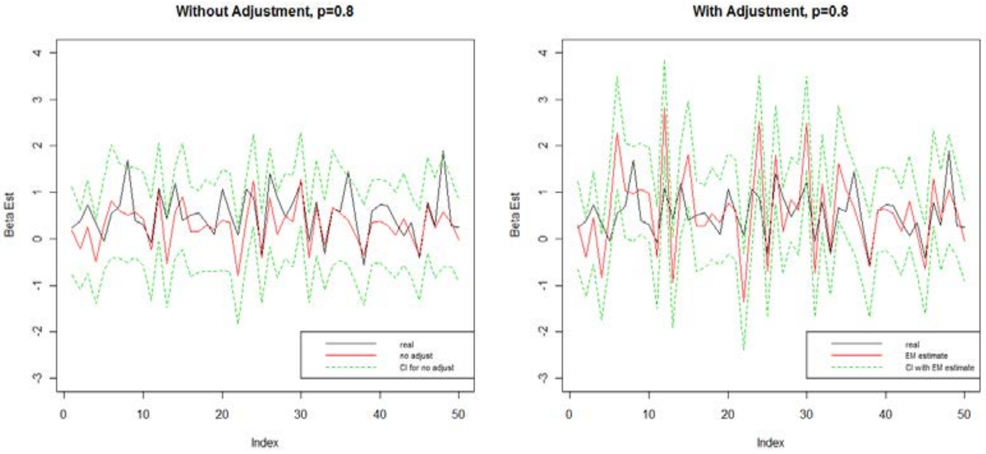

Figure 14: Plots of estimates with covariate subject to PRAM, with 95% CI, when $n = 100$, $p = 0.80$

B More Figures from Section 2.2.2

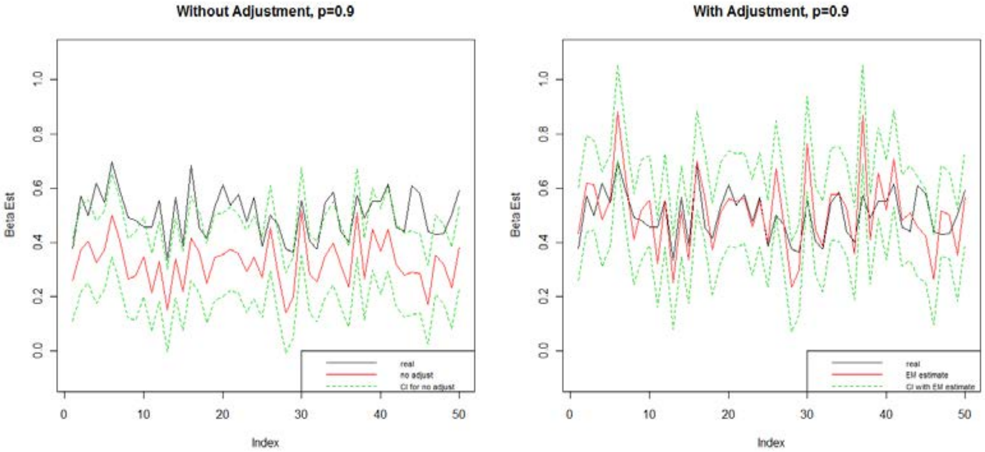

Figure 15: Plots of estimates with response subject to PRAM, with 95% CI, when $n = 1000$, $p = 0.90$

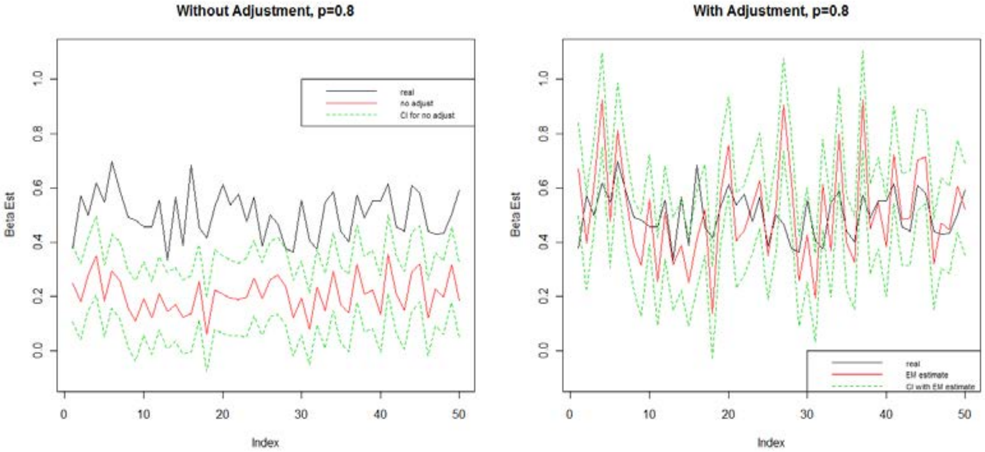

Figure 16: Plots of estimates with response subject to PRAM, with 95% CI, when $n = 1000$, $p = 0.80$

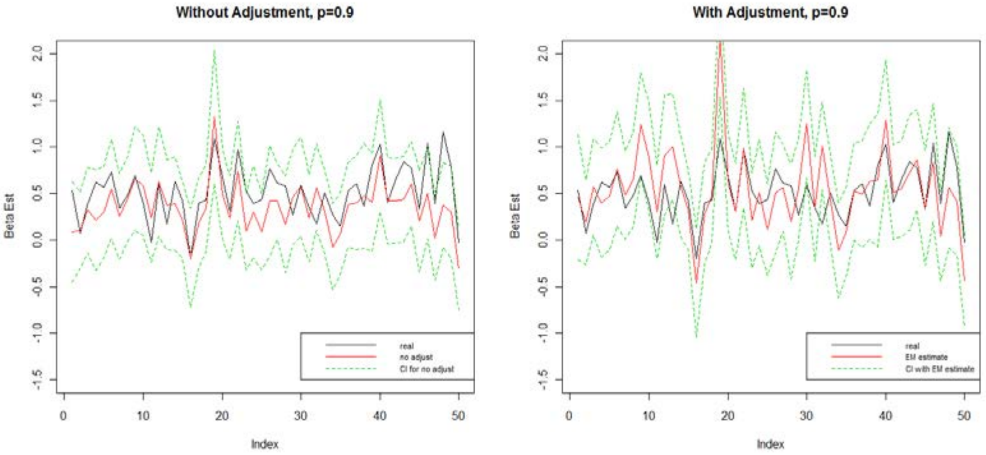

Figure 17: Plots of estimates with response subject to PRAM, with 95% CI, when $n = 100$, $p = 0.90$

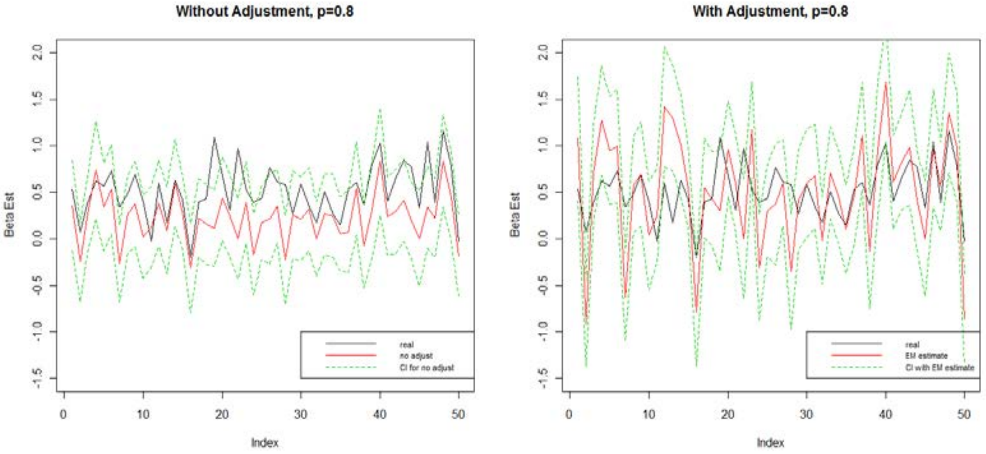

Figure 18: Plots of estimates with response subject to PRAM, with CI, when $n = 100$, $p = 0.80$

C Derivation of Weights for EM Algorithms I, II, III

C.1 Weights for EM Algorithm I

$$\begin{aligned}
P(X_{new}=x_j|X^*=x_k,Y,\phi^{(v)}) &= \frac{P(Y,X_{new}=x_j,X^*=x_k|\phi^{(v)})}{P(Y,X^*=x_k|\phi^{(v)}))} \\
&= \frac{P(Y|X_{new}=x_j,X^*=x_k,\phi^{(v)})P(X_{new}=x_j,X^*=x_k,\phi^{(v)})}{\sum_l^J P(Y,X^*=x_k,X=x_l|\phi^{(v)}))} \\
&= \frac{P(Y|X_{new}=x_j,\phi^{(v)})P(X^*=x_k|X_{new}=x_j)P(X_{new}=x_j)}{\sum_l^J P(Y|X_{new}=x_l,X^*=x_k,\phi^{(v)})P(X^*=x_k|X_{new}=x_l)P(X_{new}=x_l)} \\
&= \frac{p_{Xjk}P(Y|x_j,\beta^{(v)})\pi_j^{(v)}}{\sum_{l=1}^J p_{Xlk}P(Y|x_l,\beta^{(v)})\pi_l^{(v)}}
\end{aligned}$$

C.2 Weights for EM Algorithm II

$$\begin{aligned}
P(Y_{new}=j|Y^*=k,X,\beta^{(v)}) &= \frac{P(Y_{new}=j,Y^*=k|X,\beta^{(v)})}{P(Y^*=k|X,\beta^{(v)})} \\
&= \frac{P(Y_{new}=j|X,\beta^{(v)})P(Y^*=k|Y_{new}=j)}{\sum_l^J P(Y^*=k,Y=l|X,\beta^{(v)})} \\
&= \frac{p_{Yjk}P(Y=j|X,\beta^{(v)})}{\sum_l P(Y^*=k|Y=l)P(Y=l|X,\beta^{(v)})} \\
&= \frac{p_{Yjk}P(Y=j|X,\beta^{(v)})}{\sum_l p_{Ylk}P(Y=l|X,\beta^{(v)})}
\end{aligned}$$

C.3 Weights for EM Algorithm III

$$P(Y=m,X=l|Y^*=k,X^*=j,\beta) = P(Y=m|X=l,Y^*=k,X^*=j,\beta)P(X=l|Y^*=k,X^*=j,\beta)$$

The first part is

$$\begin{aligned}
P(Y=m|X=l,Y^*=k,X^*=j,\beta) &= \frac{P(Y=m,X=l,Y^*=k,X^*=j|\beta)}{P(X=l,Y^*=k,X^*=j|\beta)} \\
&= \frac{P(Y^*=k|Y=m)P(Y=m|X=l,\beta)P(X^*=j|X=l)P(X=l)}{\sum_a P(Y=a,X=l,Y^*=k,X^*=j|\beta)} \\
&= \frac{P_{Ymk}P(Y=m|X=l,\beta)P(X^*=j|X=l)P(X=l)}{\sum_a P(Y^*=k|Y=a)P(Y=a|X=l,\beta)P(X^*=j|X=l)P(X=l)} \\
&= \frac{P_{Ymk}P(Y=m|X=l,\beta)}{\sum_a P_{Yak}P(Y=a|X=l,\beta)}.
\end{aligned}$$

The second part is

$$\begin{aligned}
P(X=l|Y^*=k,X^*=j,\beta) &= \frac{P(X=l,Y^*=k,X^*=j|\beta)}{P(Y^*=k,X^*=j|\beta)} \\
&= \frac{\sum_b P(Y=b,X=l,Y^*=k,X^*=j|\beta)}{\sum_c\sum_d P(Y=d,X=c,Y^*=k,X^*=j|\beta)} \\
&= \frac{\sum_b P(Y^*=k|Y=b)P(Y=b|X=l,\beta)P(X^*=j|X=l)P(X=l)}{\sum_c\sum_d P(Y^*=k|Y=d)P(Y=d|X=c,\beta)P(X^*=j|X=c)P(X=c)} \\
&= \frac{p_{Xlj}\pi(l)\sum_b p_{Ybk}P(Y=b|X=l,\beta)}{\sum_c p_{Xcj}\pi(c)\sum_d p_{Ydk}P(Y=d|X=c,\beta)}.
\end{aligned}$$

Hence

$$P(Y=m, X=l|Y^*=k, X^*=j, \beta) = \frac{p_{Ymk} P(Y=m|X=l, \beta^{(v)})}{\sum_a p_{Yak} P(Y=a|X=l, \beta^{(v)})} \frac{p_{Xlj} \pi(l) \sum_b p_{Ybk} P(Y=b|X=l, \beta^{(v)})}{\sum_c p_{Xcj} \pi(c) \sum_d p_{Ydk} P(Y=d|X=c, \beta^{(v)})}.$$