SMIKe vs. Data Swapping and PRAM for Statistical Disclosure Control in Microdata: a Simulated Study

Fang Liu and Roderick J. A. Little
Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48105-2029, U.S.A.

KEY WORDS: Statistical Disclosure Control (SDC), Information Loss; Disclosure Risk; Protection.

ABSTRACT: Selective Multiple Imputation of Keys (SMIKe) is a tool for statistical disclosure control (SDC) in microdata. It is model-based with emphasis on satisfactory protection, low-level information loss and valid statistical inferences. SMIKe releases multiple sets of the modified data, rather than just one set as in data swapping and post randomization (PRAM). This article compares these three SDC techniques in a simulation study. The results suggest that SMIKe is a more efficient SDC technique than data swapping and PRAM: information loss on statistical inferences in SMIKe can be reduced to a low level and the provided protection on sensitive cases is superior to that in data swapping and PRAM. Valid statistical inferences can be obtained from SMIKe-modified data using Reiter's (2003) corrected combining rules, while data swapping and PRAM give estimates of parameters with large biases and incorrect coverage probabilities.

1. Introduction

Selective multiple imputation of keys (SMIKe) (Little and Liu, 2003a; Liu and Little, 2002) is a form of probabilistic imputation of categorical key variables. It bears some resemblance to data swapping methods where the key variables of paired cases are switched. However, simple data swapping has two obvious shortcomings: relationships between swapped and nonswapped variables are distorted by swapping, the method offers only limited protection. SMIKe has some similarity to PRAM as well, since both methods change the values of categorical key variables probabilistically. However, the probability matrix that PRAM uses is the same for all cases and entries in that matrix are chosen by the data producer. In contrast, SMIKe employs *empirically*-based probabilities that differ from case to case. In SMIKe, only the values of key variables in a subset of cases – namely, sensitive cases mixed with a subset of nonsensitive cases – are imputed (sensitive cases are defined as cases in cells formed by the key variables with $\leq s$ cases, where s is called the sensitivity threshold, the corresponding cells are called sensitive cells; nonsensitive cases come from nonsensitive cells with size $> s$). Thus, instead of releasing samples of the imputed population data set, we release the sample data with values of key variables for some cases replaced by multiple imputations. The selective aspect of SMIKe limits information loss, and reduces sensitivity of inferences to misspecification of the imputation model. SMIKe method can also be extended to handle microdata with both continuous and categorical key variables (Little, Liu and Raghu, 2003).

Besides SMIKe, we have developed another Bayesian SDC method called multiple and stochastic swapping of keys (MaSSK). MaSSK is based on the same principles as SMIKe, but replaces imputation by switching to preserve the original key cell counts. Specifically, MASSK swaps key information between selectively paired cases according to some swapping probabilities that are derived from a Bayesian model on the original data and releases multiple swapped data sets. For more details on the MASSK method, see Liu (2003). This article focuses on the comparison of disclosure control by SMIKe, data swapping and PRAM in a simulated microdata set.

2. SMIKe

Basically, there are five steps in SMIKe: 1) selection of sensitive cases and a mixing set of nonsensitive cases for which the values of the key variables are deleted and imputed, 2) construction of an imputation model for keys, 3) multiple imputation of keys, 4) statistical inferences based on the SMIKed data and assessment of information loss and protection, 5) release of the imputed data sets. Here, we elaborate these steps with continuous $\mathbf{Y}$, for details in general situations, see Little and Liu (2003a).

2.1 Selection

For each sensitive case i, a mixing set $\mathcal{M}_i$ is selected of nonsensitive cases that have similar values of the nonkey variables based on some suitable measure. (Little and Liu, 2003a) describe some alternative methods for selecting mixing sets. We use here a version called local selection, which selects a nonsen-

sitive cell that is close to a sensitive case in terms of the non-key variables, and then randomly picks n_{mix} cases from that selected cell. The closeness of sensitive case i to nonsensitive cell k is measured by the Mahalanobis distance $(\bar{\mathbf{y}}_k - \mathbf{y}_i)^T S^{-1}(\bar{\mathbf{y}}_k - \mathbf{y}_i)$, where S^{-1} is the sample within-cell covariance matrix. This is a natural measure if the nonkey variables are approximately normal; other measures can be developed for non-normal variables. If the closest cell does not contain enough cases, cases from other close cells are including in $\mathcal{M}_i$. For simplicity, the size of mixing set n_{mix} is the same for all sensitive cases, but it can be varied according to the level of sensitivity of the sensitive case. After selection a mixing set for each of the sensitive case, the values of key variables are deleted for the set $\mathcal{M}$ consisting of all the sensitive cases and their mixing sets (K^* cells and n^* cases).

2.2 Construction of the Imputation Model

The general location model (Olkin and Tate, 1961) is one possible imputation model for categorical $\mathbf{X}$ and continuous $\mathbf{Y}$. The model is defined in terms of the marginal distribution of x and conditional distribution of $\mathbf{Y}$ given x

$$\begin{aligned} p(x_i = k) &= \pi_k, \text{ where } k = 1, \ldots, K^*; \sum_k \pi_k = 1 \\ p(\mathbf{Y}_i | x_i) &\overset{indep}{\sim} N_{(p)}(\boldsymbol{\mu}_{x_i}, \Sigma) \text{ for } i = 1, \ldots, n^*. \end{aligned}$$

For the local selection method described above, $p(x)$ and $p(\mathbf{Y}|x)$ can both be estimated using the data in $\mathcal{M}$. However, since the selection of mixing sets is based on x, $p(\mathbf{Y}|x)$ can also be estimated without selection bias using the larger set (say $\mathcal{C}$ with K^* cells and n^{**} cases) of sensitive cases and all nonsensitive cases from the cells in the mixing sets, yielding improved model efficiency. The general location model requires the assumptions of normality and constant covariance of $\mathbf{Y}$ given x. A transformation on $\mathbf{Y}$ might improve the fit of the model if the assumptions are not satisfied. Another possible model is the extended general location model (Liu and Rubin, 1998), where covariance matrix does not have to be constant across cells and normal distribution may be replaced by other (like t) distributions. If the cases in a data set are not independent, as in multistage samples, we may need to modify the above model to incorporate the correlations.

2.3 Imputation of Keys

Denote the parameters in the model by $\boldsymbol{\theta} = \{\pi_1, \ldots, \pi_{K^*-1}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{K^*}, \Sigma\}$ and suppose set $\mathcal{C}$ is used for model building. If Jeffreys' priors are used,

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K^*} \pi_k^{-\frac{1}{2}} |\Sigma|^{-\frac{p+1}{2}},$$

then the posterior distributions of $\boldsymbol{\theta}$ is

$$\begin{aligned} [\boldsymbol{\pi}|x, \mathbf{y}] &\sim Dirichlet(n_1^* + \frac{1}{2}, \ldots, n_{K^*}^* + \frac{1}{2}) \\ [\Sigma|\boldsymbol{\pi}, x, \mathbf{y}] &\sim Inv - Wishart(S, n^{**} - K^*) \\ [\boldsymbol{\mu}_k|\boldsymbol{\pi}, \Sigma, x, \mathbf{y}] &\sim N_{(p)}(\bar{\mathbf{y}}_k, \Sigma/n_k^{**}) \text{ for } k = 1, \ldots, K^*, \end{aligned} \quad (1)$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{K^*})^T$, $\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{pk})^T$, S is the pooled sample covariance matrix of n^{**} cases, and $\bar{\mathbf{y}}_k$ is the sample mean of $\mathbf{y}$ in cell k from $\mathcal{C}$. The full conditional posterior predictive distribution of $\tilde{x}_i$ for case $i = 1, \ldots, n^*$ is given by

$$p(\tilde{x}_i = k|\boldsymbol{\theta}, x, \mathbf{y}) = \frac{\pi_k exp(\psi_{ik})}{\sum_{k'=1}^{K^*} \pi_{k'} exp(\psi_{ik'})} \quad (2)$$

where

$$\begin{aligned} k &= 1, \ldots, K^* \\ \psi_{ik'} &= \mathbf{y}_i^T \Sigma^{-1} \boldsymbol{\mu}_{k'} - \frac{1}{2} \boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_{k'} \end{aligned}$$

The imputation process involves drawing $\boldsymbol{\theta}$ from distributions in Eqn. (1) and imputing $\tilde{x}$ given drawn $\boldsymbol{\theta}$ and $\mathbf{Y}$ from Eqn. (2). As with multiple imputation in missing data problems, there are proper and improper versions of SMIKe. In proper SMIKe, a new set of $\boldsymbol{\theta}$ is drawn for each set of $\tilde{x}$; In improper SMIKe, only one set of $\boldsymbol{\theta}$ is drawn for D sets of $\tilde{x}$. Proper SMIKe is theoretically superior to improper SMIKe since it correctly propagates uncertainty about $\boldsymbol{\theta}$ in the imputations. We show simulation results for proper SMIKe here, although results for improper SMIKe were similar.

2.4 Statistical Inferences and Information Loss

Statistical inferences, information loss and disclosure risk can be based on D multiply-imputed data sets. For inferences for a scalar parameter of interest ϕ, let $\hat{\phi}_d$ denote an estimate of ϕ from the d^{th} imputed data set and V_d denote an estimate of the variance of $\hat{\phi}_d$, $(d = 1, \ldots, D)$. Then the MI estimate of ϕ is

$$\bar{\phi} = \sum_{d=1}^{D} \hat{\phi}_d / D, \quad (3)$$

and the MI estimate of the variance of $\bar{\phi}$ is given by

$$\begin{aligned} T &= W + \frac{1}{D} B, \text{ where} \quad (4) \\ W &= \sum_{d=1}^{D} V_d / D, \text{ and } B = \sum_{d=1}^{D} (\hat{\phi}_d - \bar{\phi})^2 / (D-1), \end{aligned}$$

where T, W and B are respectively called the total, within and between variance of $\hat{\phi}$, and $1/D$ is a correction factor for small D. Note that the combining rule for T in Eqn. (4) differs from the rule for missing data, namely $T = W + (1 + \frac{1}{D})B$ (Rubin, 1987). The reason for the difference is that in SMIKe the parameters are drawn from the complete data prior to masking rather than the incomplete data with values of x masked. Since the posterior distribution of the parameters is based on more information than in standard missing-data application of MI, the standard combination rule overestimates the variance of $\bar{\phi}$ and results in conservative inferences. Reiter (2003) derives Eqn. (4) for the situation where all the values of a subset of variables are multiply imputed. The method remains valid in our case, providing the MI predictive distribution takes into account the selection of mixing sets. The corresponding measure of information loss γ is given by:

$$\gamma = \frac{B/D}{T}. \tag{5}$$

2.5 Disclosure Risk and Protection

Assessment of disclosure risk is difficult since it requires conjectures about the behavior of data intruder. In SMIKe the difficulties are compounded by the release of multiple imputed data sets. Little and Liu (2003a) proposes two methods to measure disclosure risk in SMIKed data. For simplicity, we use the simpler one here, though results from the more complex measure are similar. It measures disclosure risk $R^d(\text{smike})$ the d^{th} imputed data set, then takes an average to get an overall measure of $R = \sum_d R^d(\text{smike})/D$. The relative gain in protection is then given by

$$P = 1 - R(\text{smike})/R(\text{ori}), \tag{6}$$

where $R(\text{ori})$ is the disclosure risk in original data set. The larger P is, the greater the reduction in disclosure risk.

3. Data Swapping and PRAM

Data swapping is a model-free SDC technique that switches key information between a pair of cases. We consider two versions of data swapping. The first is what people usually refer to as data swapping, and we call it random data swapping (RDS). In RDS, a swapping rate $r\%$ is pre-specified and $n \times r\%$ of the total cases are randomly picked from the data to be swapped with another randomly picked $n \times r\%$ cases. Sanil, Gomatam and Karr (2002) suggest $r\%$ between $1 \sim 10\%$. In our setting, $r\% = s_2$, that is, the fraction of cases under swapping goes up with the amount of sensitivity in the data. Obviously, RDS leaves some sensitive cases unprotected due to the random selection of cases for swapping. For full protection on sensitive cases, we also consider an alternative version, deterministic data swapping (DDS), where all sensitive cases are required to be swapped into different cells.

PRAM is another model-free SDC technique that transforms categorical key information of every case according to some pre-specified Makov matrix. For PRAM, the Markov matrix A is chosen according to the method proposed in Gouweleeuw, Willenborg and de Wolf (1998). With K categories in the key, A is a $K \times K$ matrix. Entry a_{kl} for $k, l = 1, \ldots, K$ represent the probability that a case whose original key is k is transformed to category l, thus $\sum_{l=1}^{K} a_{kl} = 1$ given k. A does not have to be symmetric and is constant for all cases. Let $T(k)$ denote the frequency of cases in cell k in original data and assume without loss of generality that $T(k) \geq T(K) > 0$ for $k = 1, \ldots, K$, (thus $T(K) = 1$ in this simulation) then as suggested by the authors, a simple choice of A is:

$$a_{kl} = \begin{cases} 1 - (\theta\, T(K)/T(k)) & \text{if } l = k \\ \theta\, T(K)/((K-1)T(k)) & \text{if } l \neq k\,, \end{cases} \tag{7}$$

where k is the original key of a case in the data set, $l\,(= 1, \ldots, K)$ is the candidate cell after transformation. With $T(K) = 1$ and $\theta \in (0, 1)$, the probability that cases in cell k remain in their original cell is $a_{kk} = 1 - \theta/T(k)$. If $T(k) = 1$, then $a_{kk} = 1 - \theta$. Therefore, if we want to obtain good protection for unique cases, then the likelihood of the uniqueness being transformed out of their original cells should be high and θ should be close to 1. The probability that cases remain in original cell k increases with $T(k)$. For an instance, with $\theta = 0.999$ for cases in a 3-case cell, the probability of remaining in original cell is 0.667, which we think is unsatisfactorily high. Hence, PRAM focuses more on the protection of unique cases than on non-unique cases. In this study, we fix $\theta = 0.999$ for all eight sensitivity levels; a value close to its upper limit of 1 seems needed to provide adequate protection under this method.

4. Simulation Scenario

Each of the 500 simulated data set has sample size $n = 750$, four categorical key variables $\mathbf{X} = \{X_1, X_1, X_3, X_4\}$ and two continuous nonkey variables $\mathbf{Y} = \{y_1, y_1\}$. The probabilities of population key cells are presented in Table 1 (these are the predicted probabilities from a loglinear model with four-way interaction fitted to the right panel of the table

Table 1: *Population Key Cell Structure* ($\boldsymbol{\pi} \cdot 100$)

		X_3	1		2		3	
X_1	X_2	X_4	1	2	1	2	1	2
1	1		2.64	1.38	2.17	1.64	1.61	0.93
	2		2.37	2.43	0.86	1.33	0.44	0.83
2	1		1.33	0.85	0.43	0.13	0.64	0.07
	2		0.36	0.56	0.20	0.15	0.14	0.07
3	1		3.26	3.23	1.15	1.03	0.80	0.60
	2		2.10	2.56	0.33	0.38	0.11	0.17
4	1		2.95	4.53	1.41	1.69	1.35	0.97
	2		1.64	4.08	0.70	1.41	0.34	0.65
5	1		2.55	1.75	0.73	0.68	1.02	0.39
	2		1.40	1.14	0.40	0.45	0.32	0.24
6	1		4.76	3.28	1.06	0.49	1.02	0.11
	2		2.01	2.44	0.42	0.36	0.32	0.10
7	1		3.03	1.70	1.08	0.69	0.69	0.30
	2		1.27	0.91	0.69	0.58	0.30	0.27

in page 160 in the book by Agresti (1990)). Each case is independently simulated from the following general location model:

$$\begin{aligned} x_i &\sim \text{multinomial}\,(\pi_1, \ldots, \pi_K), \quad \text{where } K = 84 \\ \mathbf{y}_i | x_i &\sim N_{(2)}(\boldsymbol{\mu}_{x_i}, \Sigma), \end{aligned}$$

where $x_i = \{1, \ldots, K = 84\}$, $i = 1, \ldots, n$ and Σ is the covariance matrix with $\sigma_1^2 = 1.0$, $\sigma_2^2 = 1.44$ and $\sigma_{12} = 1.02$ (the values of 84 sets of mean $\boldsymbol{\mu} = (\mu_1, \mu_2)^t$ are not presented here).

We vary the assumed level of sensitivity in the data by applying the SDC methods with eight different values of s ($s = 3, 4, 5, 6, 7, 8, 9, 10$). For each values of s, the sensitivity level of a sample is measured using two indices: s_1 = proportion of sensitive cells among all key cells, s_2 = proportion of sensitive cases in sample data. Increasing s results in more sensitive cases and more modifications of the data by SMIKe and the other two SDC techniques. Parameters of interest are the coefficients from two multiple linear regressions: y_1 on $(\mathbf{X}, y_2)$ and y_2 on $(\mathbf{X}, y_1)$

5. Results

The measures of protection are presented in Table 2. The first two rows under "Sensitivity Index" are the averaged s_1 and s_2 over the 500 simulated data sets. In DDS, $P = 1$, since all sensitive cases are swapped to some other-than-their-original cells. From the table, we can see that RDS and PRAM are inferior to SMIKe in protection, across all sensitivity levels.

Results on estimated bias and confidence interval coverage are shown for 4 regression coefficients out of 24 for simplicity, results for the other parameters are given in Liu (2003), and are similar. Estimated bias, averaged over the 500 simulated data sets, is plotted in Figure 2(a). SMIKe produces small estimated bias, as expected since by MI theory the method yields consistent estimates under a correctly-specified model. Data swapping (either RDS or DDS) has the most severe estimated bias, and PRAM also has much larger estimated bias compared to that in SMIKe. In SMIKe and data swapping, estimated bias goes up with (s_1, s_2), and different parameters response differently to the increasing sensitivity level.

Figure 2(b) presents (CP(SDC)-CP(original)), where CP denotes empirical coverage rate of the 95Data swapping and PRAM have coverages far below the nominal 95biased estimation and the failure to reflect imputation uncertainty in the variance. In data swapping, CP can be close to zero, yielding CP(swap)-CP(original) $\simeq 0 - 0.95 = -0.95$, as seen in the plots. In contrast, SMIKe with the adjusted MI estimate of variance gives coverages close to the nominal 95Level. Note estimated bias and CP for PRAM are the same for eight sensitivity levels, since varying sensitivity level actually does not affect PRAM – all cases have their keys transformed at all levels of s according to a fixed transformation matrix A (the parameter θ is fixed at 0.999 in Eqn. (7)).

Information loss is shown only for SMIKe, since equivalent measures are not available for data swapping and PRAM. Figure 1 displays the information loss in SMIKe of the 4 parameters across the 8 sensitivity levels. We conclude that information loss increases with the sensitivity threshold, but is generally small.

6. Conclusion

In this paper, we have discussed SMIKe as an SDC tool in microdata and compared it with two model-free approaches – data swapping and PRAM. From MI theory and the simulation results, it can be seen that model-based SMIKe provides valid statistical inferences if the imputation model is well- specified and the imputation process is proper. SMIKe makes good use of the original information, and limits the imputation task by modifying only part of the original data. It provides a tool for balancing the gain in disclosure control against the loss of information. In contrast, data swapping and PRAM do not measure

Table 2: *Protection in Data Sets Modified by SMIKe, RDS, DDS and PRAM*

		Sensitivity Index							
SDC	$\bar{s}_1$:	0.253	0.325	0.388	0.446	0.497	0.543	0.584	0.619
Technique	$\bar{s}_2$:	0.054	0.088	0.123	0.162	0.202	0.243	0.284	0.324
SMIKe		0.921	0.904	0.884	0.882	0.883	0.888	0.893	0.898
DDS		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
RDS		0.755	0.708	0.677	0.664	0.661	0.669	0.681	0.697
PRAM		0.629	0.640	0.638	0.628	0.630	0.625	0.599	0.577

or propagate this information loss, and some inferences based on the modified data are invalid. While more work is needed on developing SMIKe for realistic survey data sets, we think the method is also feasible in practice. Data users are provided with a set of imputed rectangular data sets that can be analyzed using standard statistical software, and final inferences can be combined using the comparatively simple adjusted MI methods of analysis.

References

Agresti, Alan (1990), *Categorical Data Analysis,* New York: John Wiley: 160.

Gouweleeuw, P.K., Willenborg, L.C.R.J. and de Wolf, P.-P. (1998), "Post Randomization for Statistical Disclosure Control: Theory and Implementation," *Journal of Official Statistics* (14): 463-478.

Little, R.J.A. and Liu, F. (2003,a), "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata," *to be submitted*

Little, R.J.A. and Liu, F. (2003,b), "Multiple and Stochastic Swapping of Key Variable for Statistical Disclosure Control in Microdata," *to be submitted*

Little, R.J.A., Liu, F. and Raghunathan T.E. (2003) "Statistical Disclosure Techniques Based on Multiple Imputation," *to be submitted*

Liu, F. (2003), "Bayesian Methods for Statistical Disclosure Control in Microdata," *Ph.D. Dissertation*, Department of Biostatistics, the University of Michigan, Ann Arbor.

Liu, F. and Little, R.J.A. (2002), "Selective Multiple Imputation of Keys in Microdata – an Introdution," *2002 Proceedings of the American Statistical Association*, Section of Survey Research Methodology [CD-ROM], Alexandria, VA: American Statitical Association.

Liu, C.H. and Rubin, D.B. (1998), "Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data," *Biometrika*, 85(3): 673-688.

Olkin, I. and Tate, R.F. (1961), "Multivariate Correlation Models with Mixture Discrete and Continuous Variables," *Annals of Mathematical Statistics*, (32): 448-465

Reiter, J.P. (2003), "Inferences for partially Synthetic, Public Use Microdata Sets," Unpublished manuscripts.

Rubin, D.B.(1987), *Multiple Imputation for Nonresponse in Survey*, New York: John Wiley and Sons.

Sanil, A., Gomatam, S. and Karr, F. A.(2002), "NISS WebSwap: A Web Service for Data Swapping," Technical report for Digital Government Project as National Institute of Statistical Sciences.

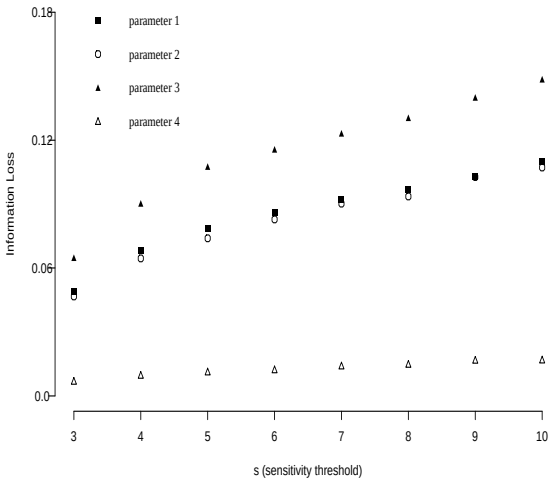
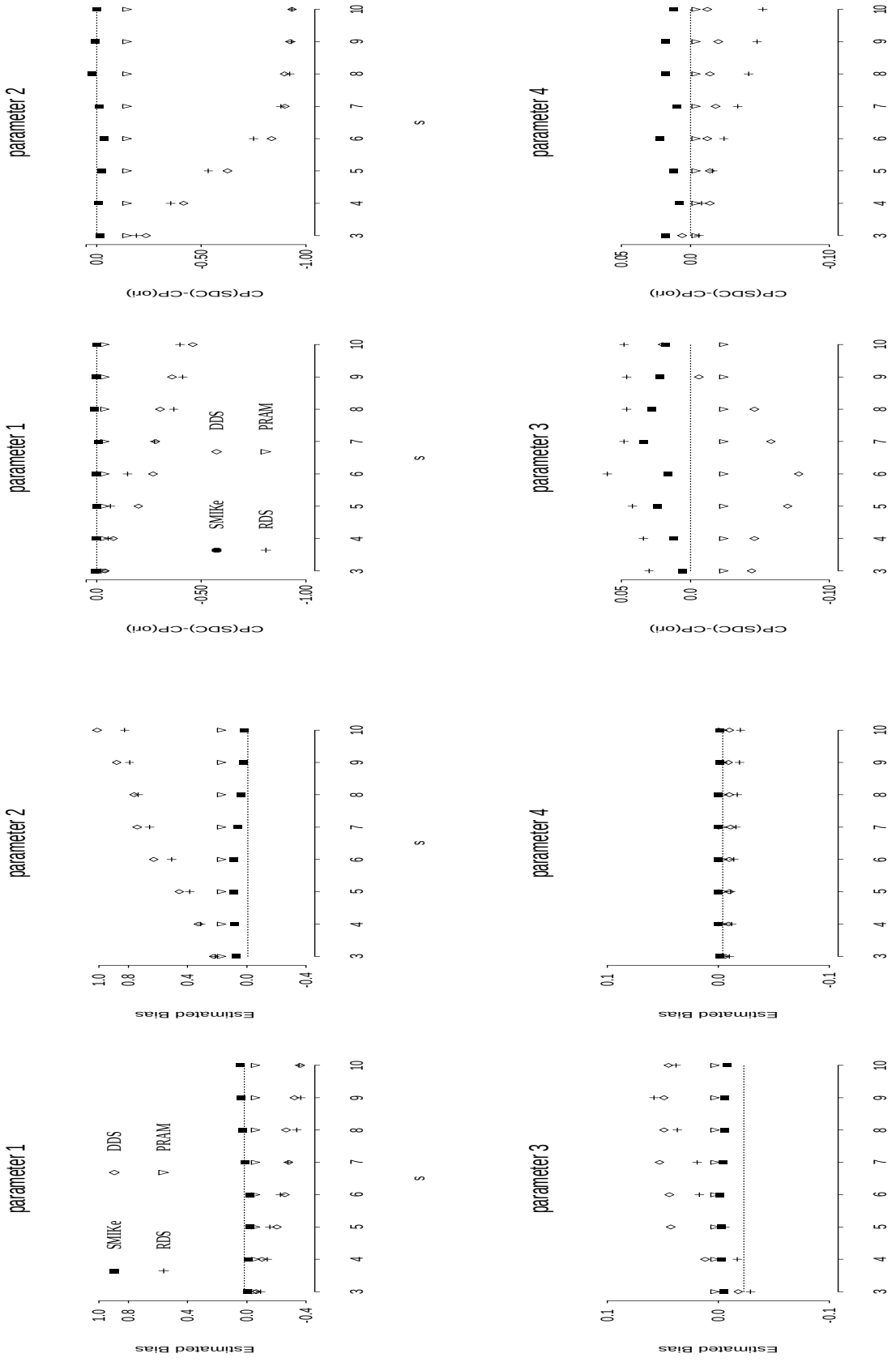


Figure 1: Information Loss for the Estimation of 4 Parameters from SMIKe-imputed Data



(a) *Estimated Bias ("- - " represents bias from original data)*

(b) *CP(SDC)-CP(ori) of nominal 95% CI*