A SUMMARY OF NOTATIONS

Table 5: Summary of notations

$g(x)$	True density
$a(x)$	Auxiliary density
$b(x)$	Noise distribution
n	Sample size
$\tilde{p}(x;\theta)$	Unnormalized model
$p(x;\theta)$	Normalized model
$q(x;\tau)$	One-parameter extended model
$x_{\rm obs}\,, x_{\rm mis}$	Observed data and missing data
$r(x)$	$q(x;\tau)/a(x)$ or $\tilde{p}(x;\theta)/a(x)$
$\Pr(\delta\|x,\phi)$	Selection probability
$\pi(\delta\|x,\phi)$	Propensity score model
η	(θ,ϕ)
ζ	(τ,ϕ)
M_{sc}	Loss function of score matching
Z_{sc}	Estimating equation of score matching
M_{nc}, M_{nc1}, M_{nc2}	Loss function of NCE
Z_{nc}, Z_{nc1}, Z_{nc2}	Estimating equation of NCE
$p(x;\theta)$	Normalized model of $\tilde{p}(x;\theta)$
$t(x_{\rm mis};\eta)$	Posterior $p(x;\theta)\pi(\delta\|x;\phi)$
θ_0	True θ
$x_i^{(*k)}$	Imputed data
$t(x)^{\otimes 2}$	$t(x)t(x)^\top$
$\hat{\eta}_p$	Initial estimator
$\hat{\eta}_{sc}$	Estimator by FISCORE and MISCORE
$\hat{\eta}_{nc,f}$	Estimator by FINCE and MINCE
μ	Baseline measure
$c_s(x;\theta)$	$\nabla_{x^s}\log\tilde{p}(x;\theta)$

B PROOF

To keep the clarity of the main points of this section, we will not specify regularity conditions. For details, see Chapter 5 in van der Vaart (1998).

Proof of Theorem 2 amd 1. Direct calculation based on the original theory of M-estimator. ■

Proof of Theorem 3. First, we discuss the general derivation without using a specific form of $z_{sc}(\theta)$ so that it can be applied to NCE case. Then, we derived the specific formula for FISCORE.

we have

$$\bar{Z}_{sc}(\theta|\hat{\theta}_p) = Z_{sc,\rm obs}(\theta) + \mathrm{E}[Z_{sc,\rm mis}|\mathbf{x}_{\rm obs};\hat{\theta}_p],$$

where $Z_{sc,\rm mis} = Z_{sc}(\theta) - Z_{sc,\rm obs}(\theta)$. By Taylor expansion, we have

$$\mathrm{E}[Z_{sc,\rm mis}|\mathbf{x}_{\rm obs};\hat{\theta}_p] \quad = \mathrm{E}[Z_{sc,\rm mis}(\theta_0)|\mathbf{x}_{\rm obs};\theta_0] + \mathrm{E}[Z_{sc,\rm mis}(\theta)\nabla_{\theta^\top}\log p(\mathbf{x}_{\rm mis}|\mathbf{x}_{\rm obs};\theta)|\mathbf{x}_{\rm obs};\theta_0]|_{\theta_0}(\hat{\theta}_p-\theta_0) + \mathrm{o}_p(n^{-1/2}).$$

Therefore,

$$\begin{aligned}\bar{Z}_{sc}(\theta_0|\hat{\theta}_p) &= Z_{sc,\rm obs}(\theta_0) - \mathcal{I}_{2,sc}(\hat{\theta}_p-\theta_0) + \mathrm{o}_p(n^{-1/2}) \\ &= -\mathcal{I}_{1,sc}(\hat{\theta}_{sc}-\theta_0) - \mathcal{I}_{2,sc}(\hat{\theta}_p-\theta_0) + \mathrm{o}_p(n^{-1/2}) \\ &= (-\mathcal{I}_{1,sc}-\mathcal{I}_{2,sc})(\hat{\theta}_{sc}-\theta_0) - \mathcal{I}_{2,sc}(\hat{\theta}_p-\hat{\theta}_{sc}) + \mathrm{o}_p(n^{-1/2}),\end{aligned} \tag{12}$$

where

$$\begin{aligned}
\mathcal{I}_{1,sc} &= \mathrm{E}[\nabla_{\theta^\top} Z_{sc,\mathrm{obs}}(\theta_0)], \\
\mathcal{I}_{2,sc} &= -\mathrm{E}[\mathrm{E}[Z_{sc,\mathrm{mis}}(\theta_0)\nabla_{\theta^\top} \log p(\mathbf{x}_{\mathrm{mis}}|\mathbf{x}_{\mathrm{obs}};\theta_0)]] \\
&= -\mathrm{E}[\mathrm{E}[z_{sc,\mathrm{mis}}(\theta_0)\nabla_{\theta^\top} \log p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta_0)]] \\
&= -\mathrm{E}[z_{sc,\mathrm{mis}}(\theta_0)\nabla_{\theta^\top} \log p(x_{\mathrm{mis}}|x_{\mathrm{obs}})] \\
&= -\mathrm{E}[\mathrm{cov}[z_{sc,\mathrm{mis}}(\theta_0), \nabla_{\theta^\top} \log \tilde{p}(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta_0)]|x_{\mathrm{obs}};\theta_0].
\end{aligned}$$

From the first line to the second line (12), we used $\mathrm{E}[Z_{sc,\mathrm{mis}}(\theta_0)|\mathbf{x}_{\mathrm{obs}};\theta_0] = 0$ and Theorem 2.

In addition, since $\hat{\theta}_{sc,\infty}$ is the solution to $\bar{Z}_{sc}(\theta|\hat{\theta}_p)$. Then,

$$\begin{aligned}
0 &= \bar{Z}_{sc}(\hat{\theta}_{sc,\infty}|\hat{\theta}_p) \\
&= \bar{Z}_{sc}(\theta_0|\hat{\theta}_p) + \mathrm{E}[\nabla_{\theta^\top} Z_{sc}(\theta_0)](\hat{\theta}_{sc,\infty} - \theta_0) + \mathrm{o}_p(n^{-1/2}) \\
&= \bar{Z}_{sc}(\theta_0|\hat{\theta}_p) + \mathcal{I}_{3,sc}(\hat{\theta}_{sc,\infty} - \theta_0) + \mathrm{o}_p(n^{-1/2}),
\end{aligned}$$

where

$$\mathcal{I}_{3,sc} = \mathrm{E}[\nabla_{\theta^\top} Z_{sc}(\theta_0)].$$

Therefore, we get

$$\begin{aligned}
(\hat{\theta}_{sc,\infty} - \theta_0) &= -\mathcal{I}_{3,sc}^{-1}\{(-\mathcal{I}_{1,sc} - \mathcal{I}_{2,sc})(\hat{\theta}_{sc} - \theta_0) - \mathcal{I}_{2,sc}(\hat{\theta}_p - \hat{\theta}_{sc})\} + \mathrm{o}_p(n^{-1/2}), \\
&= (\hat{\theta}_{sc} - \theta_0) + \mathcal{I}_{3,sc}^{-1}\mathcal{I}_{2,sc}(\hat{\theta}_p - \hat{\theta}_{sc}) + \mathrm{o}_p(n^{-1/2}).
\end{aligned}$$

From the first line to the second line of the last equation, we used the relation $\mathcal{I}_{3,sc} = \mathcal{I}_{1,sc} + \mathcal{I}_{2,sc}$. This is proved by

$$\begin{aligned}
\mathcal{I}_{1,sc} + \mathcal{I}_{2,sc} &= \mathrm{E}[\nabla_{\theta^\top}(\mathrm{E}[Z_{sc}(\theta)|\mathbf{x}_{\mathrm{obs}};\theta])] - \mathrm{E}[\mathrm{E}[Z_{sc,\mathrm{mis}}(\theta_0)\nabla_{\theta^\top} \log p(\mathbf{x}_{\mathrm{mis}}|\mathbf{x}_{\mathrm{obs}};\theta_0)|\mathbf{x}_{\mathrm{obs}};\theta_0]] \\
&= \mathrm{E}[\nabla_{\theta^\top} Z_{sc}(\theta_0)] = \mathcal{I}_{3,sc}.
\end{aligned}$$

We go back to the special case of FISCORE.

Noting $m_{sc}(\theta) = \sum_{s=1}^{d_x} 0.5c_s^2(x) + \nabla_{x^s}(c_s(x))$, the term $z_{sc}(\theta)$ is

$$z_{sc}(\theta) = \sum_{s=1}^{d} \left\{c_s(x)\nabla_\theta(c_s(x)) + \nabla_{x^s}(\nabla_\theta c_s(x))\right\}.$$

Then, we have

$$\begin{aligned}
&\mathrm{E}[\nabla_{\theta^\top} z_{sc,\mathrm{obs}}(\theta)]|_{\theta_0} = \mathrm{E}[\nabla_{\theta^\top}\{\mathrm{E}[z_{sc}(\theta)|x_{\mathrm{obs}};\theta]\}]|_{\theta_0} \\
=&\mathrm{E}[\nabla_{\theta^\top} z_{sc}(\theta)]|_{\theta_0} + \mathrm{E}[\mathrm{E}[z_{sc}(\theta)\{\nabla_{\theta^\top} \log p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta)\}|x_{\mathrm{obs}};\theta_0]]|_{\theta_0} \\
=&\mathrm{E}[\nabla_{\theta^\top} z_{sc}(\theta)]|_{\theta_0} + \mathrm{E}[z_{sc}(\theta)\{\nabla_{\theta^\top} \log p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta)\}]|_{\theta_0},
\end{aligned}$$

where $\nabla_\theta \log p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta)$ is

$$\nabla_\theta \log \tilde{p}(x;\theta) - \mathrm{E}[\nabla_\theta \log \tilde{p}(x;\theta)|x_{\mathrm{obs}};\theta].$$

So, the above is equal to

$$\mathrm{E}[\nabla_{\theta^\top} z_{sc}(\theta)]|_{\theta_0} + \mathrm{E}[\mathrm{cov}[z_{sc}(\theta), \nabla_\theta \log \tilde{p}(x;\theta)|x_{\mathrm{obs}}]]|_{\theta_0}.$$

In addition,

$$\begin{aligned}
\mathrm{E}[\nabla_{\theta^\top} z_{sc}(\theta)]|_{\theta_0} &= \mathrm{E}\left[\sum_{s=1}^{d}\left\{\nabla_\theta c_s(x)\nabla_{\theta^\top} c_s(x) + c_s(x)\nabla_{\theta\theta^\top} c_s(x) + \nabla_{x^s}(\nabla_{\theta\theta^\top} c_s(x))\right\}\right]|_{\theta_0} \\
&= \mathrm{E}\left[\sum_{s=1}^{d}\{\nabla_\theta c_s(x)\}^{\otimes 2}\right]|_{\theta_0}.
\end{aligned}$$

From the second line to the third line, we used a partial integration trick, which is a core concept of score matching. ■

Proof of Corollary 1.

Clear from Theorem 4.

■

Proof of Corollary 2.

First, we calculate $\mathcal{J}_{1,nc}$. By noting the sampling mechanism of full data is a stratified sampling, this is calculated as follows:

$$n^{-1}(\mathrm{var}_q[\mathrm{E}[z_{nc1}(x;\tau_0)|x_{\mathrm{obs}}]] + \mathrm{var}_a[z_{nc2}(y;\tau_0)]).$$

Next, we calculate $\mathcal{I}_{1,nc}$:

$$\begin{aligned}\mathcal{I}_{1,nc} &= \mathrm{E}[\nabla_{\tau^\top} Z_{nc,\mathrm{obs}}(\tau)]|_{\tau_0} = \mathrm{E}[\nabla_{\tau^\top} z_{nc,\mathrm{obs}}(\tau)]|_{\tau_0} = \mathrm{E}[\nabla_{\tau^\top}\{\mathrm{E}[z_{nc}(x,y;\tau)|x_{\mathrm{obs}};\tau]\}]|_{\tau_0} \\ &= \mathrm{E}[\nabla_{\tau^\top} z_{nc}(x,y;\tau)]|_{\tau_0} + \mathrm{E}[z_{nc}(x,y;\tau)\{\nabla_{\tau^\top}\log\bar{q}(x_{\mathrm{mis}}|x_{\mathrm{obs}};\tau)\}]|_{\tau_0} \quad (13)\\ &= \mathcal{I}_{3,nc} - \mathcal{I}_{2,nc}. \quad (14)\end{aligned}$$

where

$$\bar{q}(x_{\mathrm{mis}}|x_{\mathrm{obs}};\tau) = q(x_{\mathrm{mis}},x_{\mathrm{obs}};\tau)/\int q(x_{\mathrm{mis}},x_{\mathrm{obs}};\tau)\mu(\mathrm{d}x_{\mathrm{mis}}).$$

By some algebra, the first term in (14) is

$$\mathcal{I}_{3,nc} = \mathrm{E}[\nabla_{\tau^\top} z_{nc}(x,y;\tau)]|_{\tau_0} = \mathrm{E}\left[\frac{\nabla_\tau \log q(x;\tau_0)^{\otimes 2}}{1+r}\right]|_{\tau_0}.$$

In addition, the second term in (14) is

$$\begin{aligned}\mathcal{I}_{2,nc} &= -\mathrm{E}[z_{nc}(x,y;\tau)\{\nabla_{\tau^\top}\log\bar{q}(x_{\mathrm{mis}}|x_{\mathrm{obs}};\tau)\}]|_{\tau_0} \\ &= -\mathrm{E}[\mathrm{E}[z_{nc1}(x;\tau)|x_{\mathrm{obs}}]\{\nabla_{\tau^\top}\log\bar{q}(x_{\mathrm{mis}}|x_{\mathrm{obs}};\tau)\}]|_{\tau_0} \\ &= -\mathrm{E}[\mathrm{cov}[z_{nc1}(x;\tau),\nabla_\tau \log q(x;\tau)|x_{\mathrm{obs}}]] \\ &= \mathcal{I}_{3,nc} - \mathrm{E}\left[\mathrm{E}\left[\frac{\nabla_\tau \log q(x;\tau_0)}{1+r}|x_{\mathrm{obs}}\right]\mathrm{E}\left[\nabla_{\tau^\top}\log q(x;\tau_0)|x_{\mathrm{obs}}\right]\right].\end{aligned}$$

Therefore, adding the first and the second term in (14), we get

$$\mathcal{I}_{1,nc} = \mathrm{E}\left[\mathrm{E}\left[\frac{\nabla_\tau \log q(x;\tau_0)}{1+r}|x_{\mathrm{obs}}\right]\mathrm{E}\left[\nabla_{\tau^\top}\log q(x;\tau_0)|x_{\mathrm{obs}}\right]\right].$$

■

Proof of Corollary 3. By some algebra, as in the proof of Corollary 2, we obtain

$$\begin{aligned}\mathcal{I}_{1,nc} &= \mathrm{E}\left[\mathrm{E}\left[\nabla_\tau \log q(x;\tau_0)|x_{\mathrm{obs}}\right]^{\otimes 2}\right], \\ \mathcal{I}_{3,nc} &= \mathrm{E}\left[\nabla_\tau \log q(x;\tau_0)^{\otimes 2}\right].\end{aligned}$$

So, noting that $\mathcal{I}_{3,nc}$ is a positive definite matrix, and $\mathcal{I}_{3,nc}$ and $\mathcal{I}_{1,nc}$ are symmetric matrices, we can express $\mathcal{I}_{3,nc} = RR^\top$ and $\mathcal{I}_{1,nc} = R\Lambda R^\top$ using a nonsingular matrix R (Rao, 2008). Because $\mathcal{I}_{3,nc} - \mathcal{I}_{1,nc}$ is a positive matrix from Jensen's inequality, each element in Λ is less than 1. Then, we get

$$\mathcal{I}_{3,nc}^{-1}\mathcal{I}_{2,nc} = \mathcal{I}_{3,nc}^{-1}(\mathcal{I}_{3,nc} - \mathcal{I}_{1,nc}) = R^{-1}(I-\Lambda)R.$$

Finally,

$$(\mathcal{I}_{3,nc}^{-1}\mathcal{I}_{2,nc})^j = R^{-1}(I-\Lambda)^j R.$$

Therefore, $\{\mathcal{I}_{3,nc}^{-1}\mathcal{I}_{2,nc}\}^j$ converges to zero as j tends to infinity. ■

C Multiple missing patterns

We explain how to handle the case when the missing pattern is multiple. In the most general case, we have to introduce a missing pattern indicator δ that takes values in $0, 1, \ldots, 2^K - 1$, where K is the dimension of x, for each sample. This is based on the fact that there are 2^K possible missing patterns for x. In the main manuscript, we have defined δ to take a value of 0 or 1 because there are only two missing patterns, i.e., one value is missing or not. Then, we have to introduce an importance distribution $b(x_{mis})$ separately for each missing pattern in the general case. In practice, we can simply select the importance distribution for each coordinate and take their products. Thus, the proposed methods can be applied to the general missing case. For example, in Algorithm 2, we impute missing values of each sample using the importance distribution of corresponding missing pattern. Then, W-step and M-step are essentially the same. Here is a concrete example.

Example C.1 *Consider the case* $x_{1,obs} = [0, NA, 1]$, $x_{2,obs} = [NA, NA, 4]$, $x_{3,obs} = [1, 2, 3]$.

In this case, there are 3 *missing patterns:* $(*, *, *)$, $(*, NA, *)$, $(NA, NA, *)$*, where * means an observed value and NA means a missing value. Then, we introduce a missing indicator* δ *to take values in* $0, 1, 2$*. Namely,* $\delta = 0$ *corresponds to* $(*, *, *)$*,* $\delta = 1$ *corresponds to* $(*, NA, *)$*, and* $\delta = 2$ *corresponds to* $(NA, NA, *)$*. Then, for* $x_{1,obs} = [0, NA, 1]$*, we have* $\delta_1 = 1$*. For* $x_{2,\mathrm{obs}} = [NA, NA, 4]$*, we have* $\delta_2 = 2$*. Since* x_3 *is observed without missing, then* $\delta_3 = 0$*.* (6) *is rewritten:*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{2}\mathrm{I}(\delta_{\mathrm{i}}=\mathrm{k})\mathrm{E}[z_{nc}(x_i;\tau)\mid x_{i,\mathrm{obs}};\theta].$$

D COMPARISON BETWEEN FINCE and VNCE

Here, we compare FINCE and variational NCE (VNCE) (Rhodes and Gutmann, 2019). From (7), the difference between the estimator proposed in this paper and VNCE (Rhodes and Gutmann, 2019) is clearly shown. Mainly, there are two differences: (1) VNCE attempts to maximize the observed likelihood directly, whereas FINCE attempts to solve the observed estimating equation, (2) VNCE assumes that the dimension of $a(x)$ is the same as the dimension of x_{obs}, whereas FINCE assumes that the dimension of $a(x)$ is the same as the dimension of x.

More specifically, an ideal loss function in VNCE is

$$\begin{aligned}
&\arg\max_{s} J_{\mathrm{VNCE}}(\tau, s(x_{\mathrm{mis}})) = J_{\mathrm{VNCE}}(\tau, q(x_{\mathrm{mis}}|x_{\mathrm{obs}}))\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\log\left\{\frac{1}{1+\frac{q(x_{\mathrm{mis}}|x_{i,\mathrm{obs}})a(x_{i,\mathrm{obs}})}{q(x_{\mathrm{mis}},x_{i,\mathrm{obs}})}}\right\}|x_{i,\mathrm{obs}}\right]+\frac{1}{n}\sum_{j=1}^{n}\log\left\{\frac{a(y_j)}{a(y_j)+\mathrm{E}[q(y_j,y_{j,mis})|y_j]}\right\}\\
=&\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{q(x_{i,\mathrm{obs}})}{q(x_{i,\mathrm{obs}})+a(x_{i,\mathrm{obs}})}\right\}+\frac{1}{n}\sum_{j=1}^{n}\log\left\{\frac{a(y_j)}{a(y_j)+q(y_j)}\right\},
\end{aligned} \tag{15}$$

where $q(x_{\mathrm{obs}}) = \int q(x_{\mathrm{mis}}, x_{\mathrm{obs}})\mu(\mathrm{d}x_{\mathrm{mis}})$. On the other hand, the loss function of our proposed estimator is (7). In general, the efficiencies of the two loss function are not directly comparable. The following points highlight the comparison between two methods.

- In terms of inferences, our proposed methods (FINCE, FISCORE) are superior to VNCE because it is difficult to achieve the upper bound in (15) in VNCE.
- Unless the family of variational distribution includes the true posterior, VNCE does not have consistency. On the other hand, FINCE has consistency and also asymptotic normality without requiring such conditions.
- In terms of the scalability, VNCE is superior to the proposed methods because VNCE does not require any sampling methods.
- FINCE can be applied even if the missing data mechanism is MAR or MNAR. However, VNCE cannot be directly applied when the missing mechanism is MAR or NMAR.

E INFERENCE OF FISCORE WHEN m IS FIXED

We consider an asymptotic result of FISCORE when m is fixed. Actually, the estimating equation $Z_{sc,m}$ is not unbiased estimator for $\bar{Z}_{sc}$ because a self normalizing importance sampling is used rather than importance sampling (Owen, 2013). This means that the derived estimator is theoretically not consistent; however, practically, a self normalized importance sampling is preferable to importance sampling because of its robustness. Here, we consider the case when the weight is defined as $w(x|x_{\rm obs}) = p(x_{\rm mis}|x_{\rm obs};\theta_0)/b(x)$.

As in the proof of Theorem 3, we have

$$\hat{\theta}_{sc,m} - \theta_0 = -\mathcal{I}_{3,sc}^{-1} Z_{sc,m}(\theta_0|\hat{\theta}_p) + \mathrm{o}_p(n^{-1/2}) \tag{16}$$

This term is decomposed into two terms: $-\mathcal{I}_{3,sc}^{-1}\bar{Z}_{sc}(\theta_0|\hat{\theta}_p)$ and $-\mathcal{I}_{3,sc}^{-1}\{Z_{sc,m}(\theta_0|\hat{\theta}_p) - \bar{Z}_{sc}(\theta_0|\hat{\theta}_p)\}$. These two terms in (16) are independent. The first term is equal to $\hat{\theta}_{sc,\infty} - \theta_0$, of which the asymptotic property is shown in Theorem 3. The second term converges to the normal distribution with mean 0 and variance $\mathcal{I}_{3,sc}^{-1}\mathrm{E}[\mathrm{var}_b\{Z_{sc,m}(\theta_0|\hat{\theta}_p)\}]\mathcal{I}_{3,sc}^{\top -1}$.

Theorem 4 *When $\hat{\theta}_p = \hat{\theta}_{sc}$, the asymptotic variance of $\hat{\theta}_{sc,m}$ is equal to*

$$\mathcal{I}_{1,sc}^{-1}\mathcal{J}_{1,sc}\mathcal{I}_{1,sc}^{\top -1} + m^{-1}\mathcal{I}_{3,sc}^{-1}\mathcal{J}_{2,sc}\mathcal{I}_{3,sc}^{\top -1},$$

where $w(x|x_{\rm obs}) = p(x_{\rm mis}|x_{\rm obs};\theta_0)/b(x)$ and

$$\mathcal{J}_{2,sc} = n^{-1}\mathrm{E}[\mathrm{E}_{b(x_{\rm mis})}[w^2(x)\{z_{sc}(\theta_0)^{\otimes 2}\}|x_{\rm obs}]] - n^{-1}\mathrm{E}[\mathrm{E}[z_{sc}(\theta_0)|x_{\rm obs}]^{\otimes 2}].$$

F EXTENSION TO MULTIPLE IMPUTATION: MISCORE AND MINCE

MI was originally developed with Bayesian flavor (Rubin, 1987; Meng, 1994). In this paper, we consider frequentist MI rather than Bayesian MI (Tsiatis, 2006) to avoid the additional computation. In addition, it is shown that frequentist MI is asymptotically more efficient than Bayesian MI (Wang and Robins, 1998; Robins and Wang, 2000).

In MI, the crucial assumption is that the sample can be obtained from $p(x_{\rm mis}|x_{\rm obs};\theta)$. When the missing data mechanism is MAR, it is easy to sample from $p(x_{\rm mis}|x_{\rm obs};\theta)$ using the MCMC based on (5). The algorithm is described as in Algorithm 4. In this paper, this approach is referred to as MISCORE. MINCE is also defined similarly. Nevertheless, we do not recommend Algorithm 4 for the practical reason of its instability and computational burden.

Algorithm 4: MISCORE

repeat
  W-step: Take a set of m samples from $x_{\rm mis}^{*k} \sim p(x_{\rm mis}|x_{\rm obs};\hat{\theta}_t)$ using MCMC for each i
  M-step: Update the solution to the following function with respect to θ as $\hat{\theta}_{t+1}$:

$$\frac{1}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m} m_{sc}(x_i^{*k};\theta).$$

until $\hat{\tau}_t$ *converges*;

Dues to the challenges associated with Algorithm 4, we recommend the following algorithm. This algorithm is similar to the one in Levine and Casella (2001). In the original MISCORE, a set of samples is generated at every step. This requires tremendous computational cost and causes instability. In Algorithm 5, by constructing a $\sqrt{n}$–consistent estimator based on FISCORE at each step and updating by MISCORE one time, this limitation is overcome.

Algorithm 5: One step MISCORE

repeat
  Do W-step and M-step in Algorithm 2 (FISCORE)
until $\hat{\tau}_t$ *converges*;
Do W-step and M-step in Algorithm 4 (MISCORE)

Table 6: Monte Carlo median square error and bias

Setting 1			
n		MINCE	MISCORE
500	(bias)	0.15	0.10
	(mse)	0.037	0.024
1000	(bias)	0.13	0.12
	(mse)	0.030	0.020

Table 6 illustrates the experimental result. We generated a set of 50 samples for each i using MCMC in the last step. Compared with FINCE and FISCORE, the performance of one step MISCORE is worse. Perhaps, more step is needed.

The asymptotic property is obtained as follows.

Corollary 4 *When $\hat{\theta}_p = \hat{\theta}_{sc}$ and m is fixed, the asymptotic variance of $\hat{\theta}_{sc,\infty}$ is equal to*

$$\mathcal{I}_{1,sc}^{-1}\mathcal{J}_{1,sc}\mathcal{I}_{1,sc}^{\top -1} + m^{-1}\mathcal{I}_{3,sc}^{-1}\mathcal{J}_{2,sc}\mathcal{I}_{3,sc}^{\top -1},$$

where

$$\mathcal{J}_{2,sc} = n^{-1}\{\mathrm{E}[z_{sc}(\theta_0)^{\otimes 2}] - \mathrm{E}[\mathrm{E}[z_{sc}(\theta_0)|x_{\mathrm{obs}}]^{\otimes 2}]\},$$

and other terms are the same as in Theorem 3.

Proof of Corollary 4. We just replace $b(x_{\mathrm{mis}})$ with $p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta_0)$ in Theorem 4. ∎

Finally, there are two things to note about MISCORE and MINCE. When the missing data mechanism is MNAR, we have to sample from $\tilde{p}(x_{\mathrm{mis}}|x_{\mathrm{obs}},\delta;\eta) \propto \tilde{p}(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta)\pi(\delta|x_{\mathrm{mis}},x_{\mathrm{obs}};\phi)$. In this case, the distribution becomes a doubly-intractable distribution (Mller et al., 2006; Murray et al., 2006), and it is generally difficult to sample. Secondly, when we use a Bayesian multiple imputation assuming the prior distribution $\rho(\theta)$, even if the missing mechanism is MAR, we have to sample from $\tilde{p}(x_{\mathrm{mis}},\theta|x_{\mathrm{obs}}) \propto \tilde{p}(x_{\mathrm{mis}},x_{\mathrm{obs}};\theta)\rho(\theta)$. Often, data augmentation is utilized for this purpose (Tanner and Wong, 1987). However, even if the data augmentation is applied, we still have to deal with doubly–intractable distributions to calculate $\Pr(\theta|x) \propto \rho(\theta)p(x;\theta)$.

G EXTENSION TO CONTRASTIVE DIVERGENCE METHODS

Although there are several variations of contrastive divergence methods (Younes, 1989; Tieleman, 2008), the basic idea is that θ is updated by adding the gradient of log-likelihood $\log p(\mathbf{x};\theta)$ with respect to θ:

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \log\tilde{p}(x_i;\theta) - \mathrm{E}_{p(x;\theta)}[\nabla_\theta \log\tilde{p}(x;\theta)],$$

multiplying some learning rate. When some data is not observed, the expected gradient becomes

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[\nabla_\theta \log\tilde{p}(x_i;\theta)|x_{i,\mathrm{obs}};\theta] - \mathrm{E}[\nabla_\theta \log\tilde{p}(x;\theta)].$$

The expectation of the first term is taken under $p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta)$. It is possible to sample from MCMC like (5) without involving doubly-intractable distributions (Mller et al., 2006). Therefore, the gradient is approximated as

$$\frac{1}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m}\nabla_\theta \log\tilde{p}(x_i^{*k};\theta) - \frac{1}{n}\sum_{j=1}^{n}\nabla_\theta \log\tilde{p}(y_j;\theta),$$

where $x_i^{*k} \sim p(x_{\mathrm{mis}}|x_{i,\mathrm{obs}};\theta)$ and $y_j \sim p(y;\theta)$. We refer the updating method using the above gradient as MICD.

We can still use a FI approach for the approximation. By introducing an auxiliary distribution with a density $b(x)$, the gradient is approximated as

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m}w_{ik}\nabla_\theta \log\tilde{p}(x_i^{*k};\theta) - \frac{1}{n}\sum_{j=1}^{n}\nabla_\theta \log\tilde{p}(y_j;\theta).$$

where $x_i^{*k} \sim b(x), w_{ik} \propto \tilde{p}(x_i^{*k};\theta)/b(x_i^{*k}), y_j \sim p(y;\theta)$. We refer this approach to FICD.

Furthermore, by introducing a noise distribution with a density $a(y)$ to prevent using MCMC totally, the gradient is approximated as

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} w_{ik}\nabla_\theta \log \tilde{p}(x_i^{*k};\theta) - \frac{1}{n}\sum_{j=1}^{n} r_j \nabla_\theta \log \tilde{p}(y_j;\theta),$$

where $x_i^{*k} \sim b(x)$, $w_{ik} \propto \tilde{p}(x_i^{*k};\theta)/b(x_i^{*k})$, $y_j \sim a(y)$, and $r_j \propto \tilde{p}(y_j;\theta)/a(y_j)$. In this case, the gradient is essentially equivalent to the loss function of FINCE when $f(x) = x\log x$ by profiling-out c.

H DETAILED ALGORITHM OF FINCE WITH MNAR DATA

The algorithm is described as in Algorithm 6.

Algorithm 6: FINCE witn MNAR data

Initialize $t = 0$, $\hat{\zeta}_0 = (\hat{c}_0, \hat{\theta}_0, \hat{\phi}_0)$
Take n samples $\{y_j\}_{j=1}^n$ from $a(y)$.
For i with $\delta_i = 0$, take m samples $\{x_{i,\mathrm{mis}}^{*k}\}_{k=1}^m$ from $b(x)$.
For i with $\delta_i = 1$, set m samples $\{x_i^{*k}\}_{k=1}^m$ to $x_i^{*k} = x_i$
repeat
  W-Step:
  For i with $\delta_i = 0$; $w_{ik} \propto q(x_i^{*k};\hat{\tau}_t)\pi(\delta_i|x_i^{*k};\hat{\phi}_t)/b(x_{i,\mathrm{mis}}^{*k})$,
  For i with $\delta_i = 1$; $w_{ik} = 1/m$.
  M-step: Solve the following equation for $\hat{\zeta}_{t+1}$ w.r.t ζ:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} w_{ik} z_{nc1}(x_i^{*k};\tau) + Z_{nc2}(\mathbf{y};\zeta) = 0,$$
$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} \nabla_\phi w_{ik} \log \pi(\delta_i|x_i^{*k};\phi) = 0.$$

  $t = t + 1$
until $\hat{\tau}_t$ *converges*;

I VARIANCE ESTIMATORS OF FISCORE AND FINCE

I.1 FISCORE

The variance estimator of FISCORE in the case of Corollary is defined as follows: $\hat{\mathcal{I}}_{1,sc}^{-1}\hat{\mathcal{J}}_{1,sc}\hat{\mathcal{I}}_{1,sc}^{\top -1}|_{\hat{\theta}}$, where

$$
\begin{aligned}
\hat{\mathcal{I}}_{1,sc} &= \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mathcal{I}}_{1,sc1}(x_{i,\mathrm{obs}}) + \hat{\mathcal{I}}_{1,sc2}(x_{i,\mathrm{obs}}) - \hat{\mathcal{I}}_{1,sc3}(x_{i,\mathrm{obs}})\},\\
\hat{\mathcal{I}}_{1,sc1}(x_{i,\mathrm{obs}}) &= \sum_{k=1}^{m} w(x_i^{*k};\theta)\left(\sum_{s=1}^{d}\nabla_\theta c_s(x_i^{*k})\right)^{\otimes 2},\\
\hat{\mathcal{I}}_{1,sc2}(x_{i,\mathrm{obs}}) &= \sum_{k=1}^{m} w(x_i^{*k};\theta) z_{sc}(x_i^{*k})\nabla_{\theta^\top}\log\tilde{p}(x_i^{*k};\theta),\\
\hat{\mathcal{I}}_{1,sc3}(x_{i,\mathrm{obs}}) &= \left(\sum_{k=1}^{m} w(x_i^{*k};\theta) z_{sc}(x_i^{*k})\right)\left(\sum_{k=1}^{m} w(x_i^{*k};\theta)\nabla_{\theta^\top}\log\tilde{p}(x_i^{*k};\theta)\right),\\
\hat{\mathcal{J}}_{1,sc} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\sum_{k=1}^{m} w(x_i^{*k};\theta) z_{sc}(x_i^{*k};\theta) - \bar{z}\right)^{\otimes 2},\\
\bar{z} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} w(x_i^{*k};\theta) z_{sc}(x_i^{*k};\theta),\\
z_{sc}(\theta) &= \sum_{s=1}^{d}\left\{c_s(x)\nabla_\theta(c_s(x)) + \nabla_{x^s}(\nabla_\theta c_s(x))\right\},\ c_s(x;\theta) = \nabla_{x^s}\log\tilde{p}(x;\theta).
\end{aligned}
$$

Next, consider an loss function and a variance estimator in truncated exponential family cases (Hyvärinen, 2007). Assume that $\tilde{p}(x;\theta)$ is given by

$$
\log\tilde{p}(x;\theta) = \sum_{k=1}^{d}\theta_k F_k(x).
$$

Let us denote two matrices: $d_\theta \times d$ matrix $K_1(x)$ with elements $\nabla_{x^b}F_a$ $(1 \le a \le d_\theta, 1 \le b \le d)$ and $d_\theta \times 1$ matrix, $K_{i,2}(x)$ with elements $\nabla\nabla_{x^i}F_a$ $(1 \le a \le d_x)$. The loss function is written as $n^{-1}\sum_{i=1}^{n} z_{sc,t}(x_i;\theta)$.

$$
z_{sc,t}(x;\theta) = 0.5\theta^\top K_1(x)K_1(x)^\top\theta + \theta^\top\sum_{i=1}^{d_x}K_{i,2}(x).
$$

The variance estimator is obtained almost in the same by replacing $z_{sc}(x)$ with $z_{sc,t}(x)$. The only modification is

$$
\hat{\mathcal{I}}_{1,sc1}(x_{i,\mathrm{obs}}) = \sum_{k=1}^{m} w(x_i^{*k};\theta)K_1(x_i^{*k})K_1(x_i^{*k})^\top.
$$

I.2 FINCE

The variance estimator of FINCE in the case of Corollary 2 is defined as follows: $\hat{\mathcal{I}}_{1,nc}^{-1}\hat{\mathcal{J}}_{1,nc}\hat{\mathcal{I}}_{1,nc}^{\top -1}|_{\hat{\tau}}$, where

$$
\begin{aligned}
\hat{\mathcal{I}}_{1,nc} &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{m} w(x_i^{*k};\tau)\frac{\nabla_\tau \log q(x_i^{*k};\tau)}{1+q(x_i^{*k})/a(x_i^{*k})}\right)\left(\sum_{k=1}^{m} w(x_i^{*k};\tau)\nabla_{\tau^\top}\log q(x_i^{*k};\tau)\right),\\
\hat{\mathcal{J}}_{1,nc} &= \hat{\mathcal{J}}_{1,nc1} + \hat{\mathcal{J}}_{1,nc2},\\
\hat{\mathcal{J}}_{1,nc1} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\sum_{k=1}^{m} w(x_i^{*k};\tau) z_{nc1}(x_i^{*k};\tau) - \bar{z}_{nc1}\right)^{\otimes 2},\\
\hat{\mathcal{J}}_{1,nc2} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(z_{nc2}(y_i;\tau) - \bar{z}_{nc2}\right)^{\otimes 2},\\
\bar{z}_{nc1} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} w(x_i^{*k};\tau) z_{nc1}(x_i^{*k};\tau),\ \bar{z}_{nc2} = \frac{1}{n}\sum_{j=1}^{n} z_{nc2}(y_j;\tau),\\
z_{nc1}(\tau) &= -\frac{\nabla_\tau \log q(x;\tau)}{1+r},\ z_{nc2}(\tau) = \frac{r\nabla_\tau \log q(x;\tau)}{1+r}.
\end{aligned}
$$