3-12-2019

Imputation estimators for unnormalized models with missing data

Masatoshi Uehara
Harvard University

Takeru Matsuda
The University of Tokyo

Jae Kwang Kim
Iowa State University, jkim@iastate.edu

Imputation estimators for unnormalized models with missing data

Abstract
We propose estimation methods for unnormalized models with missing data. The key concept is to combine a modern imputation technique with estimators for unnormalized models including noise contrastive estimation and score matching. Further, we derive asymptotic distributions of the proposed estimators and construct the confidence intervals. The application to truncated Gaussian graphical models with missing data shows the validity of the proposed methods.

Keywords
Unnormalized Models, Noise Contrastive Estimation, Score Matching, Missing Data, Graphical Models, Missing Not at Random

Disciplines
Statistical Methodology | Statistical Models | Theory and Algorithms

Comments
This pre-print is made available through arxiv: https://arxiv.org/abs/1903.03630.

Imputation estimators for unnormalized models with missing data

Masatoshi Uehara *[1], Takeru Matsuda[2], and Jae Kwang Kim[3]

[1]Harvard University
[2]The University of Tokyo
[3]Iowa State University

March 12, 2019

Abstract

We propose estimation methods for unnormalized models with missing data. The key concept is to combine a modern imputation technique with estimators for unnormalized models including noise contrastive estimation and score matching. Further, we derive asymptotic distributions of the proposed estimators and construct the confidence intervals. The application to truncated Gaussian graphical models with missing data shows the validity of the proposed methods.

Key Words: Unnormalized Models; Noise Contrastive Estimation ; Score Matching; Missing Data; Graphical Models; Missing Not at Random

*uehra_m@g.harvard.edu

1 Introduction

Several statistical models are presented in the form of unnormalized densities and the calculation of the normalization constant (or the partition function) is intractable. Namely,

$$p(x;\theta) = \frac{1}{Z(\theta)}\tilde{p}(x;\theta), \tag{1}$$

where $Z(\theta) = \int \tilde{p}(x;\theta)\mu(\mathrm{d}x)$, μ is a baseline measure such as Lebesgue measure or counting measure, and we only have access to $\tilde{p}(x;\theta)$. Such unnormalized models are widely used in many settings: Markov random fields (Besag, 1975), Boltzmann machines (Hinton, 2002), overcomplete independent component analysis models (Hyvärinen et al., 2001) and graphical models (Lin et al., 2016; Yu et al., 2016). Several methods for estimating θ have been developed such as noise contrastive estimation (NCE) (Gutmann and Hyvärinen, 2010) and score matching (Hyvärinen, 2005).

In this study, we investigate the estimation methods of unnormalized models with missing data. Missing data is frequently encountered and may cause nonresponse bias (Little and Rubin, 2002). Thus, how to handle missing data is an important problem.

Our problem setting is as follows. Let x be sampled from the unnormalized model (1) and suppose that we observe only part of x, which is denote by x_{obs}. The objective of this study is to estimate θ based on the observed data x_{obs}. The existing estimation methods for unnormalized models are not applicable here since all these methods assume that the complete data is fully observed.

To solve this issue, we develop estimation methods that are developed through combination of NCE and score matching with fractional imputation (Kim, 2011), which is a computationally efficient technique for the missing data free from Markov Chain Monte Carlo (MCMC). Note that Rhodes and Gutmann (2019) proposed a variational NCE for unnormalized latent variable models corresponding to a special case of the current problem (missing at random, MAR). Though variational inference is fast and useful for a large–scale problem, it is challenging to conduct statistical inference (Blei et al., 2017). On the other hand, the proposed methods enable the construction of confidence intervals based on the asymptotic theory. In addition, the proposed methods are valid under general missing mechanisms, including missing not at random (MNAR) case. Our main contributions are as follows.

- We propose imputation estimators for unnormalized models with missing data. These estimators are consistent under the general missing mechanism, including an MNAR case, and are computationally efficient.

- We derive the asymptotic distributions of the proposed estimators and construct confidence intervals.
- We confirm the validity of the proposed methods in a simulation with truncated Gaussian graphical models with missing data.

2 Preliminary

2.1 Notations

The parameters with a zero in the subscript such as θ_0 and τ_0, denote the true parameters. The notation ∇_θ denotes a differentiation with respect to θ, and $t(x)^{\otimes 2} = t(x)t(x)^\top$. The expectation and variance of $f(x)$ under the density $g(x)$ is denoted as $\mathrm{E}_g[f(x)]$ and $\mathrm{var}_g[f(x)]$, respectively. We often omit the subscript when it is obvious from the context. We present a summary of the notation in the Supplementary materials.

2.2 Missing data and imputation methods

We briefly review the framework of the missing data and the imputation methods. For more details, see Kim and Shao (2013).

Suppose that $\{x_i\}_{i=1}^n$ are independently and identically distributed (i.i.d.) samples from a distribution with density $p(x;\theta)$. We consider the situation where some part of x_i may be missing. Let $\{\delta_i\}_{i=1}^n$ be the missing indicators. Accordingly, $x_i = (x_{i,\mathrm{obs}}, x_{i,\mathrm{mis}})$ is fully observed when $\delta_i = 1$, while only $x_{i,\mathrm{obs}}$ is observed and $x_{i,\mathrm{mis}}$ is missing when $\delta_i = 0$. We assume that δ_i follows the Bernoulli distribution with probability $\Pr(\delta_i = 1 \mid x_i)$. The case with several missing patterns (the dimension of $x_{i,\mathrm{obs}}$ may differ with i) can be easily considered by extending this notation (Seaman et al., 2013).

The missing mechanism is called missing at random (MAR) if $\Pr(\delta = 1 \mid x) = \Pr(\delta = 1 \mid x_{\mathrm{obs}})$ holds. Importantly, the selection mechanism can be ignored for estimation of θ in the MAR cases (Little and Rubin, 2002), because

$$\begin{aligned} p(x_{\mathrm{obs}};\theta) &= \int p(x_{\mathrm{obs}}, x_{\mathrm{mis}};\theta)\Pr(\delta \mid x)\mu(\mathrm{d}x_{\mathrm{mis}}) \\ &\propto \int p(x_{\mathrm{obs}}, x_{\mathrm{mis}};\theta)\mu(\mathrm{d}x_{\mathrm{mis}}). \end{aligned}$$

As a special case of MAR, a missing mechanism is referred to as missing completely at random (MCAR) if $\Pr(\delta = 1 \mid x)$ does not depend on x at all. When the MAR does not hold, the missing mechanism is referred to as missing not at random (MNAR).

For estimating θ from observations, the fundamental algorithm is the Expectation Maximization (EM) algorithm (Dempster et al., 1977; Meng and Van Dyk, 1997), which maximizes the observed likelihood $p(x_{\mathrm{obs}}; \theta)$. Equivalently, the EM algorithm solves the following observed (mean) score equation with respect to θ (Louis, 1982; Elashoff and Ryan, 2004):

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\nabla_\theta \log p(x_i;\theta) \mid x_{i,\mathrm{obs}}; \theta\right] = 0. \tag{2}$$

However, the EM algorithm requires a closed-form expression of the conditional expectation in (2), which is often intractable. To solve this problem, Fractional Imputation (FI) has been proposed (Kim, 2011; Yang and Kim, 2016), which is closely connected with the Monte Carlo EM algorithm (Wei and Tanner, 1990). FI is fast because it uses only importance sampling as an approximation procedure, and does not rely on MCMC. However, it is difficult to approximate the conditional expectation using only importance sampling for large-scale problems. In such cases, Multiple Imputation (MI) is commonly used, which utilizes MCMC for approximation (Rubin, 1987; Murray, 2018).

2.3 Estimation methods for unnormalized models

Several methods have been developed for estimating unnormalized models such as score matching (Hyvärinen, 2005), noise contrastive estimation (NCE) (Gutmann and Hyvärinen, 2010), Monte Carlo maximum likelihood estimation (Monte Carlo MLE) (Geyer, 1994), and contrastive divergence (CD) (Hinton, 2002). We briefly review NCE and score matching in the following. Note that both methods take the form of Z-estimators or M-estimators (van der Vaart, 1998).

2.3.1 Generalized NCE

We review the generalized NCE from the divergence perspective (Pihlaja et al., 2010; Gutmann and Hirayama, 2011). Suppose we have $\mathbf{x} = \{x_i\}_{i=1}^n$ from the true distribution with density $g(x)$, and $\mathbf{y} = \{y_i\}_{i=1}^n$ from a noise distribution with density $a(y)$. Note that all the algorithms below can be easily extended to the case where the noise sample size is different from the original sample size.

In the NCE, we introduce a one-parameter extended model $q(x;\tau) = \exp(-c)\tilde{p}(x;\theta)$, where $\tau = (c, \theta^\top)^\top$ and c is an unknown nuisance parameter to approximate the normalizing constant. Note that it is different from the normalized model $p(x;\theta)$. For a twice differentiable strictly convex function $f(\cdot)$, a noise contrastive divergence is

defined as

$$D_{NC}(g, q(x;\tau)) = \int \mathrm{Br}_f \left(\frac{g(x)}{a(x)}, \frac{q(x;\tau)}{a(x)} \right) a(x)\mu(\mathrm{d}x), \tag{3}$$

where $\mathrm{Br}_f(o_1, o_2)$ is given by $f(o_1) - f(o_2) - f'(o_2)(o_1 - o_2)$, and $f(\cdot)$ is the divergence function. By subtracting a term not associated with θ from $D_{NC}(g, q(x;\tau))$, the cross entropy between $g(x)$ and $q(x;\tau)$ is given by

$$d_{NC}(g, p) = \mathrm{E}_{g(x)} \left[M_{nc1}(\mathbf{x}) \right] + \mathrm{E}_{a(y)} \left[M_{nc2}(\mathbf{y}) \right],$$

where

$$M_{nc1}(\mathbf{x}) = -\frac{1}{n} \sum_{i=1}^{n} f'((r(x_i)),$$
$$M_{nc2}(\mathbf{y}) = \frac{1}{n} \sum_{j=1}^{n} \left\{ f' \left(r(y_j) \right) r(y_j) - f(r(y_j)) \right\},$$

and $q(x;\tau)/a(x) = r(x;\tau)$. The objective function is defined as $M_{nc1}(\mathbf{x}) + M_{nc2}(\mathbf{y})$ because $\mathrm{D}_{NC}(g, q(x;\tau))$ takes the maximum when τ is equal to τ_0. NCE is defined as the minimizer of this objective function regarding τ. By differentiating the above $d_{NC}(g, q(x;\tau))$ regarding τ, the following moment condition is obtained:

$$\mathrm{E}_{g(x)} \left[Z_{nc1}(\mathbf{x};\tau) \right] + \mathrm{E}_{a(y)} \left[Z_{nc2}(\mathbf{y};\tau) \right] |_{\tau_0} = 0, \tag{4}$$

where $Z_{nc1}(\mathbf{x}) = 1/n \sum_{i=1}^{n} z_{nc1}(x_i;\tau)$, $Z_{nc2}(\mathbf{y}) = 1/n \sum_{j=1}^{n} z_{nc2}(y_j;\tau)$,

$$z_{nc1}(x) = -\nabla_\tau \log q(x;\tau) f'' \left(r(x;\tau) \right) r(x;\tau),$$
$$z_{nc2}(y) = \nabla_\tau \log q(y;\tau) f'' \left(r(y;\tau) \right) r(y;\tau)^2.$$

The estimator is also regarded as the solution to $Z_{nc}(\mathbf{x}, \mathbf{y};\tau) = 0$ where $Z_{nc}(\mathbf{x}, \mathbf{y};\tau) = Z_{nc1}(\mathbf{x};\tau) + Z_{nc2}(\mathbf{y};\tau)$. Specific examples of an objective function are as follows.

Example 2.1 (Monte Carlo MLE) *When $f(x) = x \log x$, the generalized NCE is defined as the minimizer of the following function with respect to τ:*

$$-\frac{1}{n} \sum_{i=1}^{n} \log q(x_i;\tau) + \left(\frac{1}{n} \sum_{j=1}^{n} r(y_j;\tau) \right).$$

The objective function is essentially the same as the Monte Carlo MLE by profiling-out c (Geyer, 1994).

Example 2.2 (Original NCE) *When $f(x) = x \log x - (1+x)\log(1+x)$, the generalized NCE is defined as the minimizer of the following function with respect to τ:*

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{r(x_i;\tau)}{r(x_i;\tau)+1} - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{r(y_j;\tau)+1}.$$

In this case, the objective function is the same as the original NCE (Gutmann and Hyvärinen, 2010). The function $f(x)$ is optimal from the perspective of asymptotic variance (Uehara et al., 2018).

2.3.2 Generalized score matching

Next, we review the score matching approach. The original score matching is introduced as a tool for minimizing the distance between the score function of the model and the data score function (Hyvärinen, 2005). It has been generalized to many settings: for truncated distributions (Hyvärinen, 2007; Lin et al., 2016), the cases involving high-order score functions (Lyu, 2009; Dawid et al., 2012; Parry et al., 2012). Here, we introduce score matching from the divergence perspective.

The divergence between $\tilde{p}(x;\theta)$ and $\tilde{p}(x;\theta')$ of the score matching, $D_{SC}(\tilde{p}(x;\theta), \tilde{p}(x;\theta'))$, is given by

$$\int \sum_{s=1}^{d_x} \mathrm{Br}_f\left(-c_s(x;\theta), -c_s(x;\theta')\right) g(x)\mu(\mathrm{d}x), \tag{5}$$

where $c_s(x;\theta) = \nabla_{x^s} \log \tilde{p}(x;\theta)$, x^s is the s–th coordinate of x, d_x is the dimension of x, and $f(\cdot)$ is the divergence function. Here, note that $c_s(x;\theta)$ is different from the score function $\nabla_\theta \log p(x;\theta)$ in the usual sense. The cross entropy is defined as $\mathrm{E}_{g(x)}[M_{sc}(\mathbf{x};\theta)]$, where $M_{sc}(\mathbf{x};\theta) = n^{-1}\sum_{i=1}^{n} m_{sc}(x_i)$, and $m_{sc}(x)$ is

$$\sum_{s=1}^{d_x}\left\{-f\left(c_s(x)\right) + \nabla_{x^s}\left(f'(c_s(x))\right) + f'(c_s(x))c_s(x)\right\}. \tag{6}$$

The estimator is defined as the minimizer of the objective function $M_{sc}(\mathbf{x};\theta)$ with respect to θ.

Example 2.3 (Score matching) *Consider the case when $f(x) = 0.5x^2$. The objective function becomes*

$$M_{sc}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n}\sum_{s=1}^{d_x}\left\{0.5c_s^2(x_i) + \nabla_{x^s}(c_s(x_i))\right\},$$

which reduces to the original score matching (Hyvärinen, 2005). It can be extended to the case where the data is on positive orthant (Hyvärinen, 2007). The objective function becomes $M_{sc}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n}\sum_{s=1}^{d_x}\{2x_{si}c_s(x_i) + x_{si}^2(0.5c_s(x_i)^2 + \nabla_{x^s}(c_s(x_i)))\}$, where x_{si} is a s–th component of x_i.

3 FINCE and FISCORE

We propose estimation methods for unnormalized models with missing data: FINCE (fractional imputation noise contrastive estimation) and FISCORE (fractional imputation score matching). For methods using MI, see Supplementary materials.

In this section, we focus on the MAR case, that is, $\Pr(\delta = 1 \mid x) = \Pr(\delta = 1 \mid x_{\mathrm{obs}})$. In Section 5, we discuss an extension to the case of missing not at random (MNAR).

3.1 NCE with EM algorithm

We incorporate the EM algorithm to NCE. Though the score equation cannot be used as in (2), an estimating equation such as the one in (4) can be used. The estimator for θ is defined based on the solution to the following equation with respect to τ:

$$\mathrm{E}[\nabla_\tau Z_{nc1}(\mathbf{x};\tau)|\mathbf{x}_{\mathrm{obs}};\theta] + \nabla_\tau Z_{nc2}(\mathbf{y};\tau) = 0, \tag{7}$$

where the expectation is taken with respect to the posterior predictive model $p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta)$:

$$\frac{p(x;\theta)}{\int p(x;\theta)\mu(\mathrm{d}x_{\mathrm{mis}})} = \frac{\tilde{p}(x;\theta)}{\int \tilde{p}(x;\theta)\mu(\mathrm{d}x_{\mathrm{mis}})}. \tag{8}$$

More specifically, the estimator is defined as the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[z_{nc1}(x;\tau)|x_{i,\mathrm{obs}};\theta] + \frac{1}{n}\sum_{j=1}^{n} z_{nc2}(y_j;\tau) = 0. \tag{9}$$

Note that the conditional expectation in (9) formally means

$$\frac{1}{n}\sum_{i=1}^{n} \left\{\delta_i z_{nc1}(x_i;\tau) + (1-\delta_i)\mathrm{E}[z_{nc1}(X;\tau)|x_{i,\mathrm{obs}}]\right\}. \tag{10}$$

This is because the dimension of x_{obs} is different for each sample. Throughout this paper, we implicitly assume this conversion following the convention in the literature of missing data (Seaman et al., 2013).

Generally, it is difficult to analytically calculate the conditional expectation under $p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta)$ in (9). In subsequent sections, we discuss how this problem can be resolved. Here, assuming that the conditional expectation in (9) can be calculated analytically, EM algorithm is described in Algorithm 1 to solve the equation (9):

Note that the third line of Algorithm 1 can be replaced with M-estimators. For example, when $f(x) = x\log x$, $\hat{\tau}_{t+1}$ is the solution to the minimizer of the following function:

$$-\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[\log q(x;\tau)|x_{i,\mathrm{obs}};\hat{\theta}_t] + \left(\frac{1}{n}\sum_{j=1}^{n} r(y_j;\tau)\right).$$

Algorithm 1: NCE with EM algorithm

1 Take a set of n samples $\{y_i\}_{i=1}^n$ from $a(y)$ and initialize $\hat{\tau}_0$

2 **repeat**

3 Solve the following equation and update the solution as $\hat{\tau}_{t+1}$:

$$\mathrm{E}[Z_{nc1}(\mathbf{x};\tau)|\mathbf{x}_{\mathrm{obs}};\hat{\theta}_t] + Z_{nc2}(\mathbf{y};\tau) = 0.$$

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

Moreover, when $f(x) = x\log x - (1+x)\log(1+x)$, $\hat{\tau}_{t+1}$ is the solution to the minimizer of the following function:

$$-\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\frac{r(x)}{r(x)+1}|x_{i,\mathrm{obs}};\hat{\theta}_t\right] - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{r(y_j)+1}. \tag{11}$$

The form (11) clearly explains the difference between Algorithm 1 and VNCE (Rhodes and Gutmann, 2019). For the details, refer to the Supplementary materials.

3.2 NCE with fractional imputation (FINCE)

The challenge in using the EM algorithm is it is often infeasible to calculate the conditional expectation analytically. Therefore, in the same spirit of FI (Kim, 2011), it is natural to incorporate an importance sampling using a random variable with a density $b(x)$. The idea is

$$\begin{aligned}&\int u(x)p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta)\mu(\mathrm{d}x_{\mathrm{mis}})\\ &= \int u(x)\frac{\frac{\tilde{p}(x_{\mathrm{mis}},x_{\mathrm{obs}};\theta)}{b(x_{\mathrm{mis}})}b(x_{\mathrm{mis}})\mu(\mathrm{d}x_{\mathrm{mis}})}{\int\frac{\tilde{p}(x_{\mathrm{mis}},x_{\mathrm{obs}};\theta)}{b(x_{\mathrm{mis}})}b(x_{\mathrm{mis}})\mu(\mathrm{d}x_{\mathrm{mis}})}.\end{aligned}$$

for any function $u(x)$. Using the above technique, we estimate $\mathrm{E}[Z_1(\mathbf{x};\tau)|\mathbf{x}_{\mathrm{obs}};\theta]$ by the importance sampling in (9). The estimator is defined as in Algorithm 2. Here, $\propto$ in the second step indicates a normalization so that the summation over k is equal to 1.

Generally, it is difficult to solve (12) directly. We can solve it with an EM approach as shown in Algorithm 3. In the EM-style algorithm, the weights are fixed at every step.

Note that Z–estimators in M–step can be replaced with M–estimators. For example, when $f(x) = x\log x$, M-step is the minimization of the following function with respect

Algorithm 2: FINCE

1 Take a set of m samples $x_{i,\mathrm{mis}}^{*k} \sim b(x)$ for each i with $\delta_i = 0$ and take a set of n samples from $y_j \sim a(y)$ $(1 \leq k \leq m, 1 \leq j \leq n)$.

2 Calculate the normalized weight: $w_{ik} \propto q(x_i^{*k}; \tau)/b(x_{\mathrm{mis}}^{*k})$, where $x_i^{*k} = (x_{i,\mathrm{obs}}, x_{\mathrm{mis}}^{*k})$. This means

$$w_{ik}(x;\tau) = \frac{q(x_i^{*k};\tau)/b(x_{\mathrm{mis}}^{*k})}{\sum_{k=1}^{m} q(x_i^{*k};\tau)/b(x_{\mathrm{mis}}^{*k})}.$$

3 Solve the following equation with respect to τ:

$$0 = \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} z_{nc1}(x_i^{*k};\tau) w_{ik}(x;\tau) \right) + Z_{nc2}(\mathbf{y};\tau) \tag{12}$$

Algorithm 3: FINCE with EM algorithm

1 Take the same first step as before in Algorithm 2

2 **repeat**

3 W-Step: $w_{ik} \propto q(x_i^{*k}; \hat{\tau}_t)/b(x_{\mathrm{mis}}^{*k})$

4 M-step Update the solution to the following function with respect to τ as $\hat{\tau}_{t+1}$:

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} z_{nc1}(x_i^{*k};\tau) w_{ik} + Z_{nc2}(\mathbf{y};\tau) = 0.$$

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

to τ:

$$-\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} w_{ik} \log q(x_i^{*k};\tau) + \left(\frac{1}{n} \sum_{j=1}^{n} r(y_j;\tau) \right). \tag{13}$$

The choice of the noise and the auxiliary distribution is important. The noise distribution $a(x)$ should be generally close to $p(x_{\mathrm{mis}}, x_{\mathrm{obs}}; \theta_0)$, the auxiliary distribution $b(x)$ should be closer to $p(x_{\mathrm{mis}}; \theta_0)$ in terms of statistical efficiency. When there are complete data for some set of samples as in Section 6, moment matching can be used to determine $a(x)$ and $b(x)$.

3.3 Score matching with fractional imputation (FISCORE)

Score matching is defined in the form of M-estimators. Thus, the idea in Section 3.2 can similarly be incorporated when there are missing data. The estimator is defined as the solution to the following equation with respect to θ:

$$\mathrm{E}[Z_{sc}(\mathbf{x};\theta)|\mathbf{x}_{\mathrm{obs}};\theta] = 0,\ Z_{sc}(\mathbf{x};\theta) = \nabla_\theta M_{sc}(\mathbf{x};\theta). \tag{14}$$

However, the calculation of the conditional expectation can be challenging. By introducing the auxiliary density $b(x)$, the above equation can be solved by an EM approach as shown in Algorithm 4.

Algorithm 4: FISCORE with EM algorithm

1 Take a set of m samples $x_{i,\mathrm{mis}}^{*k} \sim b(x)$ for each i with $\delta_i = 0$

2 **repeat**

3 W-Step: $w_{ik} \propto \tilde{p}(x_i^{*k};\hat{\theta}_t)/b(x_{\mathrm{mis}}^{*k})$.

4 M-step: Update the solution to the minimizer of the following term with respect to θ as $\hat{\theta}_{t+1}$:

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

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

4 Asymptotics and confidence intervals

We derive the asymptotic distributions of FINCE and FISCORE by extending results of Wang and Robins (1998) and Kim (2011). Based on the asymptotic distributions, we also construct confidence intervals, which enable hypothesis testing. This is an advantage of the proposed methods compared with variational NCE (Rhodes and Gutmann, 2019).

4.1 FISCORE

First, we consider the case of FISCORE. Given an initial $\sqrt{n}$-consistent estimator $\hat{\theta}_p$ for θ, we obtain the imputed equation:

$$Z_{sc,m}(\theta|\hat{\theta}_p) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} z_{sc}(\theta;x_i^{*k}) w(x_i^{*k};\hat{\theta}_p),$$

where $x_i^{*k} = (x_{i,\mathrm{obs}}, x_{i,\mathrm{mis}}^{*k})$, $x_{i,\mathrm{mis}}^{*k} \sim b(x)$,

$$w(x_i^{*k};\theta) = \frac{\tilde{p}(x_i^{*k};\theta)/b(x_{i,\mathrm{mis}}^{*k})}{\sum_{k=1}^{m}\tilde{p}(x_i^{*k};\theta)/b(x_{i,\mathrm{mis}}^{*k})}.$$

As an initial step, we consider the case $m \to \infty$ irrespective of the size of n. This result is easily applied to the case when $m \to \infty$ as $n \to \infty$. Refer to Supplementary materials when m is finite. When m is infinity, the above imputed equation $Z_{sc,m}(\theta|\hat{\theta}^p)$ converges to

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

We define the solution to $\bar{Z}_{sc}(\theta|\hat{\theta}_p) = 0$ as $\hat{\theta}_{sc,\infty}$. Ideally, when the EM algorithm is solved analytically, the estimator is defined as the solution to $Z_{sc,\mathrm{obs}}(\mathbf{x}_{\mathrm{obs}};\theta) = 0$, where

$$Z_{sc,\mathrm{obs}}(\mathbf{x}_{\mathrm{obs}};\theta) = \mathrm{E}[Z_{sc}(\theta)|\mathbf{x}_{\mathrm{obs}};\theta].$$

We define this solution as $\hat{\theta}_{s,f}$. Based on the theory of Z-estimators (van der Vaart, 1998), $\hat{\theta}_{s,f}$ has the following asymptotic property:

Theorem 1 *The term $\hat{\theta}_{s,f} - \theta_0$ is equal to*

$$-\mathrm{E}[\nabla_{\theta^\top} Z_{sc,\mathrm{obs}}(\theta_0)]^{-1} Z_{sc,\mathrm{obs}}(\theta_0) + \mathrm{o}_p(n^{-1/2}).$$

The term $\hat{\theta}_{s,f} - \theta_0$ asymptotically converges to the normal distribution with mean 0 and variance $\mathcal{I}_{1,sc}^{-1}\mathcal{J}_{1,sc}\mathcal{I}_{1,sc}^{\top -1}$, where

$$\mathcal{I}_{1,sc} = \mathrm{E}[\nabla_{\theta^\top} Z_{sc,\mathrm{obs}}(\theta_0)],\ \mathcal{J}_{1,sc} = \mathrm{Var}[Z_{sc,\mathrm{obs}}(\theta_0)].$$

Next, consider the asymptotic variance of $\hat{\theta}_{sc,\infty}$, and the corresponding result when $f(x) = 0.5x^2$. In the case of $f(x) = 0.5x^2$, each term is specified more explicitly because some terms cancel out using integration by parts.

Theorem 2 *The term $\hat{\theta}_{sc,\infty} - \theta_0$ is equal to*

$$(\hat{\theta}_{sc,f} - \theta_0) + \mathcal{I}_{3,sc}^{-1}\mathcal{I}_{2,sc}(\hat{\theta}_p - \hat{\theta}_{sc,f}) + \mathrm{o}_p(n^{-1/2}),$$

where

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

Corollary 4.1 *When $f = 0.5x^2$ and the missing data mechanism is MAR, each term becomes*

$$
\begin{aligned}
z_{sc}(\theta) &= \sum_{s=1}^{d_x} \left\{ c_s(x) \nabla_\theta (c_s(x)) + \nabla_{x^s} (\nabla_\theta c_s(x)) \right\}, \\
\mathcal{I}_{2,sc} &= \mathrm{E}\left[-\mathrm{cov}[z_{sc}(\theta), \log \tilde{p}(x;\theta) | x_{\mathrm{obs}}] \right] |_{\theta_0}, \\
\mathcal{I}_{3,sc} &= \mathrm{E}\left[\sum_{s=1}^{d_x} \left\{ \nabla_\theta c_s(x)^{\otimes 2} \right\} \right] |_{\theta_0}, \\
\mathcal{I}_{1,sc} &= \mathcal{I}_{3,sc} - \mathcal{I}_{2,sc}, \\
\mathcal{J}_{1,sc} &= n^{-1} \mathrm{Var}[\mathrm{E}[z_{sc}(\theta_0) | x_{\mathrm{obs}}]].
\end{aligned}
$$

In the proof of Theorem 2, we used the relation: $\mathcal{I}_{3,sc} = \mathcal{I}_{1,sc} + \mathcal{I}_{2,sc}$. This relation corresponds to the missing information principle or Louis' formula (Kim and Shao, 2013; Orchard and Woodbury, 1972; Louis, 1982) when the normalized model is used. Specifically, when $Z_{sc}(\theta)$ is a true score equation: $S_{sc}(\mathbf{x};\theta) = \nabla_\theta \log\{p(\mathbf{x};\theta)\}$, the result is reduced to the one in Wang and Robins (1998). In this case, $\mathcal{I}_{3,sc}, \mathcal{I}_{1,sc}$ and $\mathcal{I}_{2,sc}$ become

$$
\begin{aligned}
\mathcal{I}_{com} &= \mathrm{E}[\nabla_{\theta^\top} S_{sc}(\theta_0)], \ \mathcal{I}_{\mathrm{obs}} = \mathrm{E}[\nabla_{\theta^\top} S_{\mathrm{obs}}(\theta_0)], \\
\mathcal{I}_{\mathrm{mis}} &= \mathrm{E}[S_{\mathrm{mis}}(\theta_0)^{\otimes 2}], \\
S_{\mathrm{mis}}(\theta) &= S_{sc}(\theta) - \mathrm{E}[S_{sc}(\theta) | \mathbf{x}_{\mathrm{obs}}; \theta], \\
S_{\mathrm{obs}}(\theta) &= \int S_{sc}(\theta) \mu(\mathrm{d}\mathbf{x}_{\mathrm{mis}}),
\end{aligned}
$$

respectively, and the relation $\mathcal{I}_{com} = \mathcal{I}_{\mathrm{obs}} + \mathcal{I}_{\mathrm{mis}}$ holds.

The term $\mathcal{I}_{com}^{-1} \mathcal{I}_{\mathrm{mis}}$ is often called the fraction of missing information (Kim and Shao, 2013). For the current problem, $\mathcal{I}_{3,sc}^{-1} \mathcal{I}_{2,sc}$ can be considered as an analog.

Writing $\hat{\theta}^{(t)}$ to be the t–th EM update of θ that is computed by solving $\bar{Z}_{sc}(\theta | \hat{\theta}^{(t-1)}) = 0$, we obtain the following Corollary.

Corollary 4.2 *We have*

$$
\hat{\theta}^{(t)} = \hat{\theta}^{(t-1)} + \{ \mathcal{I}_{3,sc}^{-1} \mathcal{I}_{2,sc} \}^{t-1} (\hat{\theta}^{(0)} - \hat{\theta}_{sc,f}).
$$

When the spectral radius of $\mathcal{I}_{3,sc}^{-1} \mathcal{I}_{2,sc}$ is less than 1, $\hat{\theta}^{(t)}$ converges to $\hat{\theta}_{sc,f}$.

Generally, it is difficult to prove that the spectral radius of $\mathcal{I}_{3,sc}^{-1} \mathcal{I}_{2,sc}$ is less than 1. However, experimental results in Section 6 show that this algorithm converges.

4.2 FINCE

Next, we consider the case of FINCE. Given an initial $\sqrt{n}$–consistent estimator $\hat{\tau}_p$, we can obtain an imputed equation $Z_{nc,m}(\tau|\hat{\tau}_p)$:

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} z_{nc1}(x_i^{*k};\tau)w(x_i^{*k};\hat{\tau}_p)\right\} + Z_{nc2}(\mathbf{y};\tau) = 0,$$

where $w(x;\tau) = q(x;\tau)/b(x)$.

For the case where m is infinity. Then, $Z_{nc,m}(\tau|\hat{\tau}_p)$ converges to $\bar{Z}_{nc}(\tau|\hat{\tau}_p)$;

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[z_{nc1}(X;\tau)|x_{i,\mathrm{obs}};\hat{\tau}_p]\right\} + \frac{1}{n}\sum_{j=1}^{n} z_{nc2}(y_j;\tau).$$

Furthermore, when the EM algorithm can be solved analytically, the estimator is defined as the solution to the following equation with respect to τ:

$$\begin{aligned} 0 &= Z_{nc,obs}(\mathbf{x}_{\mathrm{obs}},\mathbf{y};\tau),\\ Z_{nc,obs}(\tau) &= \mathrm{E}[Z_{nc1}(\tau)|\mathbf{x}_{\mathrm{obs}};\tau] + Z_{nc2}(\mathbf{y};\tau). \end{aligned}$$

Here, we refer this solution to $\hat{\tau}_{nc,f}$. Similar to Theorem 1, we have the following asymptotic property.

Theorem 3 *The term $\hat{\tau}_{nc,f} - \tau_0$ converges to the normal distribution with mean 0 and variance $\mathcal{I}_{1,nc}^{-1}\mathcal{J}_{1,nc}\mathcal{I}_{1,nc}^{\top -1}$,*

$$\mathcal{I}_{1,nc} = \mathrm{E}[\nabla_{\tau^\top} Z_{nc,obs}(\tau_0)],\ \mathcal{J}_{1,nc} = \mathrm{Var}[Z_{nc,obs}(\tau_0)].$$

Especially, in the case of the original NCE, each term is specified more explicitly as follows because some terms cancel out. Refer to the Supplementary materials for variance estimators based on this result.

Corollary 4.3 *When the missing data mechanism is MAR and $f(x) = x\log x - (1+x)\log(1+x)$, all of the terms become as follows, where $\nabla_\tau \log q(x;\tau) = v(x;\tau)$ and*

$$\begin{aligned} \mathcal{I}_{1,nc} &= \mathrm{E}\left[\mathrm{E}\left[\frac{v(x;\tau_0)}{1+r_0}|x_{\mathrm{obs}}\right]\mathrm{E}\left[v(x;\tau_0)^\top|x_{\mathrm{obs}}\right]\right],\\ \mathcal{I}_{3,nc} &= \mathrm{E}\left[\frac{v(x;\tau_0)^{\otimes 2}}{1+r_0}\right],\ r_0 = q(x;\tau_0)/a(x),\\ \mathcal{J}_{1,nc} &= n^{-1}(\mathrm{var}_q[\mathrm{E}[z_{nc1}(x;\tau_0)|x_{\mathrm{obs}}]]\\ &\quad + \mathrm{var}_a[z_{nc2}(y;\tau_0)]),\\ z_{nc1}(\tau) &= -\frac{v(x;\tau_0)}{1+r_0},\ z_{nc2}(\tau) = \frac{rv(x;\tau_0)}{1+r_0}. \end{aligned}$$

Actually, when $f(x) = x \log x$, we can prove that $\{\mathcal{I}_{3,nc}^{-1}\mathcal{I}_{2,nc}\}^j$ tends to zero as j tends to infinity.

Corollary 4.4 *When $f(x) = x \log x$, $\mathcal{I}_{1,nc}$ and $\mathcal{I}_{3,nc}$ become as follows:*

$$\mathcal{I}_{1,nc} = \mathrm{E}\left[\mathrm{E}\left[v(x;\tau_0)|x_{\mathrm{obs}}\right]^{\otimes 2}\right],$$
$$\mathcal{I}_{3,nc} = \mathrm{E}\left[v(x;\tau_0)^{\otimes 2}\right].$$

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

Note when there is no missing data, NCE is more efficient than Monte Carlo MLE (Uehara et al., 2018). On the other hand, when there is missing data, this statement does not hold. However, the efficiency of the methods depends on the underlying generating mechanism.

5 Some extensions

5.1 Extension to MNAR case

In general, the nonparametric identification condition does not hold in the MNAR case (Robins and Ritov, 1997). However, assuming the existence of nonresponse instrument and parametric models, the parameter can be identified in some cases (Kim and Kim, 2012; Wang et al., 2014). We hereafter assume the existence of nonresponse instrument so that the parameter can be idenfitied.

To estimate the parameter under MNAR data, FISCORE and FINCE can be still applied. First, we specify a propensity score model $\pi(\delta|x;\phi)$ for $\Pr(\delta|x)$. For the case of FISCORE, we want to solve the equation with respect to η:

$$\mathrm{E}\left[\begin{pmatrix} Z_{sc}(\mathbf{x};\theta) \\ \nabla_\phi \log \pi(\boldsymbol{\delta}|\mathbf{x};\phi) \end{pmatrix} |\mathbf{x}_{\mathrm{obs}}, \boldsymbol{\delta}; \eta\right] = 0, \tag{15}$$

where the expectation is taken under $t(\mathbf{x}_{\mathrm{mis}}|\mathbf{x}_{\mathrm{obs}}, \boldsymbol{\delta}; \eta) \propto p(\mathbf{x};\theta)\pi(\boldsymbol{\delta}|\mathbf{x};\phi)$, and $\eta = (\theta, \phi)$. Importantly, we can take care of the selection mechanism unlike in the MAR and MCAR cases, because $p(x_{\mathrm{mis}}|x_{\mathrm{obs}}) = p(x_{\mathrm{mis}}|\delta, x_{\mathrm{obs}})$ does not hold. The difference is evident when we compare (15) with (14). Owing to MNAR, the first modification is such that the selection mechanism $\pi(\delta|x)$ appears when calculating the fractional weight: $w_{ik} \propto \tilde{p}(x_i^{*k};\hat{\theta}_t)\pi(\delta_i|x_i^{*k};\hat{\phi}_t)/b(x_{\mathrm{mis}}^{*k})$. The second modification is the score of the propensity score model which is shown in (15).

In the case of FINCE, let $\zeta = (\tau^\top, \phi^\top)^\top$ and $Z_{nc}(\boldsymbol{\delta}, \mathbf{x}, \mathbf{y}; \zeta)$ be defined as an augmented estimating equation:

$$\begin{pmatrix} Z_{nc}(\boldsymbol{\delta}, \mathbf{x}, \mathbf{y}; \tau) \\ \nabla_\phi \log \pi(\boldsymbol{\delta}|\mathbf{x}; \phi) \end{pmatrix}.$$

The algorithm is modified to solve the following equation with respect to ζ:

$$\mathrm{E}\left[\begin{pmatrix} Z_{nc}(\boldsymbol{\delta}, \mathbf{x}, \mathbf{y}; \tau) \\ \nabla_\phi \log \pi(\boldsymbol{\delta}|\mathbf{x}; \phi) \end{pmatrix} |\mathbf{x}_{\mathrm{obs}}, \boldsymbol{\delta}; \zeta \right] = 0.$$

5.2 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 (8) 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 objective function of FINCE when $f(x) = x\log x$ by profiling-out c.

6 Simulation results

We present some simulation results to show the performance of FINCE and FISCORE under the following two settings: (1) truncated normal distribution with missing data including MNAR case and (2) truncated Gaussian graphical models with missing data.

6.1 Truncated normal distribution

Consider a truncated normal distribution: $\phi(x;\Sigma^{-1}) = \exp(-0.5x^\top\Sigma^{-1}x)\mathrm{I}(x>0)$ where Σ is a 2 by 2 matrix parameter and $x = (x_1, x_2)$ is a two-dimensional vector. Assume x_1 is fully observed; however, x_2 is subject to missingness. The random variable δ is binary; if $\delta = 1$, x_2 is not missing, and if $\delta = 0$, x_2 is subject to missingness. We performed simulations under two settings using a R-package developed by Genz et al. (2018). In both cases, the parameter values under the missing data models are chosen so that the overall missing rates are about 30%.

- MAR : $\Pr(\delta = 1|x) = 1/[1+\exp\{-(x_1-0.9)/0.3\}]$ and

$$\Sigma = \begin{pmatrix} 2 & 1.3 \\ 1.3 & 2.0 \end{pmatrix}.$$

- MNAR : $\Pr(\delta = 1|x) = 1/[1+\exp\{-(x_2-\mu)/\sigma\}]$ where $\mu = 0.9$ and $\sigma = 0.2$, and the same Σ as in the first setting.

We compared the following estimators:

- **COMP**: This estimator uses an NCE based on complete data only. We used a truncated distribution as an auxiliary distribution and noise distribution.
- **FINCE**: This estimator uses an FINCE with $m = 100$.

Table 1: Monte Carlo median square error and bias

		MAR		
n		COMP	FINCE	FISCORE
500	(bias)	0.29	0.03	0.03
	(mse)	0.040	0.024	0.021
1000	(bias)	0.25	0.02	0.02
	(mse)	0.032	0.015	0.011
		NMAR		
n		Comp	FINCE	FISCORE
500	(bias)	0.33	0.18	0.12
	(mse)	0.041	0.027	0.021
1000	(bias)	0.24	0.14	0.14
	(mse)	0.039	0.020	0.012

- **FISCORE**: This estimator uses an FISCORE with $m = 100$. In this case, we used a score matching for a truncated tensity (Hyvärinen, 2007). See Supplementary materials for details.

We do not compare them with variational NCE because it does not take into account a MNAR case and does not give a confidence interval.

Table 6.1 shows the results of Monte Carlo median of absolute bias and square errors. The results revealed that **COMP** leads to the significant bias. This outcome is expected because using only complete cases leads to the bias in the case of MAR, although it it not in the case of MCAR (Little and Rubin, 2002). On the other hand, it is shown that **FINCE** and **FISCORE** are consistent estimators. Though the performance of **FISCORE** is better than that of **FINCE** in this experiment, by increasing the number of auxiliary samples, it is expected that the efficiency of **FINCE** will be improved.

We also constructed a 95%confidence interval based on the variance estimators in Supplementary materials. Table 2 shows the result of the coverage rate.

Table 2: Coverage rate under Setting 1

n	FINCE	FISCORE
500	94%	89%
1000	94%	92%

6.2 Truncated Gaussian graphical model

Next, we consider the estimation of the truncated Gaussian graphical model (GGM) considered in Lin et al. (2016) with missing data.

Let $G = (V, E)$ be an undirected graph where $V = \{1, \cdots, d\}$. Then, the truncated GGM with graph G is defined as $p(x \mid \Sigma) \propto \exp\left(-0.5 x^\top \Sigma^{-1} x\right) \quad (x \in \mathbb{R}^d_+)$, where $\Sigma \in \mathbb{R}^{d \times d}$ is a positive definite matrix satisfying $(\Sigma^{-1})_{ij} = 0$ for $(i, j) \notin E$. Similar to the original GGM (Lauritzen, 1996), X_i and X_j are conditionally independent on the other variables X_k $(k \neq i, j)$ if $(i, j) \notin E$. Here, we estimate G by using the confidence intervals of the entries of Σ^{-1}.

We generated $n = 1000$ independent samples $\{x_i\}_{i=1}^n$ from a truncated GGM (6.2) with $d = 10$ and the G given in the top panel of Figure 1. Namely, there are three clusters $(x_1, x_2, x_3), (x_4, x_5, x_6)$, and (x_7, x_8, x_9) of three variables and one isolated variable x_{10}. We set all the diagonal entries of Σ^{-1} to 1 and all the nonzero off-diagonal entries of Σ^{-1} to 0.5. We introduced missing values on x_3, x_6 and x_9 by using the following MAR mechanism: for $k = 1, 2, 3$, random vector $c_k \in \mathbb{R}^{10}$ was generated by $(c_k)_3 = (c_k)_6 = (c_k)_9 = 0$ and $(c_k)_j \sim \mathrm{N}(0, 1)$ $(j \neq 3, 6, 9)$ and then x_{3k} was missed with the probability $1/(3 + \exp(c_k^\top x))$. The proportion of complete data was about 40%.

Then, we fitted the truncated GGM (6.2) to $\{x_i\}_{i=1}^n$ by using FINCE and FISCORE with 100 imputations. We used $\mathrm{N}(0, 2)$ truncated to the positive orthant as the proposal distribution for missing entries. In FINCE, we generated $n = 1000$ noise samples $\{y_i\}_{i=1}^n$ from the product of the coordinate-wise exponential distributions with the same mean as $\{x_i\}_{i=1}^n$.

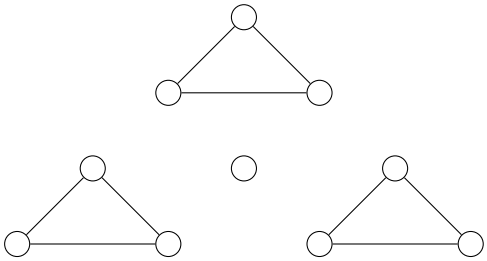
We determined the graph G by collecting all edges (i, j) such that the 95 % confidence interval of $(\Sigma^{-1})_{ij}$ did not include zero. Figure 1 shows the result of one realization.

We calculated the proportions of falsely selected edges (false positive) and falsely unselected edges (false negative) in 100 realizations. The results are given in Table 3. It shows that the coverage probabilities of the confidence intervals are approximately equal to 95% in both FINCE and FISCORE.

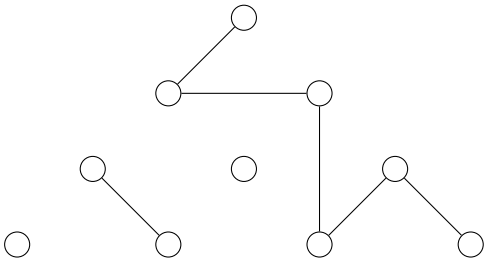
Table 3: Proportions of false positive and false negative

	FINCE	FISCORE
FP	10.5%	6.4%
FN	12.6%	23.3%

truth



FINCE



FISCORE

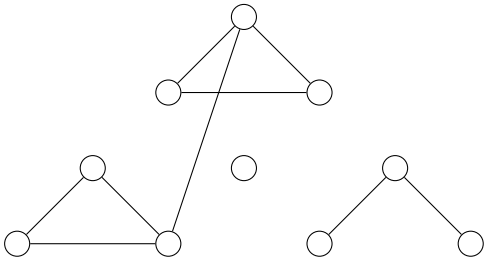


Figure 1: Selected graphs

7 Conclusion

We have proposed estimation methods for unnormalized models with missing data: FINCE and FISCORE. The proposed methods are computationally efficient, valid under generel missing mechanisms, and enable statistical inference using the confidence intervals.

In this study, we focus on NCE and score matching. It is an interesting future work to investigate the theory of FICD (fractional imputation with contrastive divergence) and its application to large scale problems. An extension of the recently developed statistically efficient estimators for unnormalized models (Uehara et al., 2019) to missing data setting is another interesting future problem.

References

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician) 24*, 179–195.

Blei, D. M., A. Kucukelbir, and J. D. Mcauliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association 112*, 859–877.

Dawid, A. P., S. Lauritzen, and M. Parry (2012). Proper local scoring rules on discrete sample spaces. *The Annals of Statistics 40*, 593–608.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–37.

Elashoff, M. and L. Ryan (2004). An em algorithm for estimating equations. *Journal of Computational and Graphical Statistics 13*, 48–65.

Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2018). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-8.

Geyer, C. (1994). On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B 56*, 261–274.

Gutmann, M. and J. Hirayama (2011). Bregman divergence as general framework to estimate unnormalized statistical models. *In Proccedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2011)*.

Gutmann, M. and A. Hyvärinen (2010). Noise contrastive estimation: A new estimation principle for unnormalized statistical models. *In Procceedings of the International Conference on Artificial Intelli-gence and Statistics (AISTATS 2010)*.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation 14*, 1771–1800.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research 6*, 695–709.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics and Data Analysis 51*, 2499–2512.

Hyvärinen, A., J. Karhunen, and E. Oja (2001). *Independent component analysis*. New York: J. Wiley.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika 98*, 119–132.

Kim, J. K. and J. Shao (2013). *Statistical Methods for Handling Incomplete Data*. Chapman & Hall / CRC.

Kim, J. Y. and J. K. Kim (2012). Parametric fractional imputation for nonignorable missing data. *Journal of the Korean Statistical Society 41*, 291–303.

Lauritzen, S. L. (1996). *Graphical models*. Oxford statistical science series ; 17. Oxford: Clarendon Press.

Levine, R. A. and G. Casella (2001). Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics 10*, 422–439.

Lin, L., M. Drton, and A. Shojaie (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electronic journal of statistics 10*, 806–854.

Little, R. J. A. and D. B. Rubin (2002). *Statistical analysis with missing data* (2nd ed. ed.). Hoboken, N.J.: Wiley.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B 44*, 226–233.

Lyu, S. (2009). Interpretation and generalization of score matching. *In Proccedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2009)*.

Meng, X. and D. Van Dyk (1997). The em algorithman old folksong sung to a fast new tune. *Journal of the Royal Statistical Society 59*, 511–567.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science 9*, 538–573.

Mller, J., A. N. Pettitt, R. Reeves, and K. K. Berthelsen (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika 93*, 451–458.

Murray, I., Z. Ghahramani, and D. MacKay (2006). Mcmc for doubly-intractable distributions. *In Proccedings of the Conference on Uncertainty in Artificial Intellegence (UAI 2006)*.

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science 33*, 142–159.

Orchard, T. and M. Woodbury (1972). A missing information principle: theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, Berkeley, California, pp. 695–715. University of California Press.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples*.

Parry, M. F., A. P. Dawid, and S. L. Lauritzen (2012). Proper local scoring rules. *Annals of Statistics 40*, 561–592.

Pihlaja, M., M. Gutmann, and A. Hyvärinen (2010). A family of computationally efficient and simple estimators for unnormalized statistical models. *In Procceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2010)*.

Rao, C. R. (2008). *Linear Statistical Inference and its Applications: Second Editon*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Rhodes, B. and M. Gutmann (2019). Variational noise-contrastive estimation. *In Proceedings of the International Workshop on Artificial Intelligence and Statistics 2019 (AISTATS 2019)*.

Robins, J. M. and Y. Ritov (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine 16*, 285–319.

Robins, J. M. and N. Wang (2000). Inference for imputation estimators. *Biometrika 87*, 113–124.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Seaman, S., J. Galati, D. Jackson, and J. Carlin (2013). What is meant by "missing at random"? *Statistical Science 28*, 257–268.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distribution by data augmentation. *Journal of the American Statistical Association 82*, 528–540.

Tieleman, T. (2008). Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the international conference on machine learning (ICML 2008)*.

Tsiatis, A. (2006). *Semiparametric theory and missing data*. Springer series in statistics. New York: Springer.

Uehara, M., T. Kanamori, T. Takenouchi, and T. Matsuda (2019). Unified estimation framework for unnormalized models with statistical efficiency. *arXiv preprint arXiv:1901.07710*.

Uehara, M., T. Matsuda, and F. Komaki (2018). Analysis of noise contrastive estimation from the perspective of asymptotic variance. *arXiv preprint arXiv:1808.07983*.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, UK ; New York, NY, USA: Cambridge University Press.

Wang, N. and J. M. Robins (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika 85*, 935–948.

Wang, S., J. Shao, and J. K. Kim (2014). Identifiability and estimation in problems with nonignorable nonresponse. *Statistical Science 24*, 1097 – 1116.

Wei, G. C. G. and M. A. Tanner (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85*, 699–704.

Yang, S. and J. K. Kim (2016). Fractional imputation in survey sampling: A comparative review. *Statistical Science 31*, 415–432.

Younes, L. (1989). Maximum likelihood estimation for Gaussian fields. *Probability Theory and Related Fields 82*, 625–645.

Yu, M., M. Kolar, and V. Gupta (2016). Statistical inference for pairwise graphical models using score matching. *In Advances in Neural Information Processing Systems (NIPS 2016)*.

A Summary of notations

Table 4: 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
$\hat{\eta}_{sc,f}$	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. First, we have

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

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

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

Therefore,

$$\begin{aligned}\bar{Z}_{sc}(\theta_0|\hat{\theta}_p) &= Z_{sc,\text{obs}}(\theta_0) - \mathcal{I}_{2,sc}(\hat{\theta}_p - \theta_0) + \mathrm{o}_p(n^{-1/2}) \\ &= -\mathcal{I}_{1,sc}(\hat{\theta}_{s,f} - \theta_0) - \mathcal{I}_{2,sc}(\hat{\theta}_p - \theta_0) + \mathrm{o}_p(n^{-1/2}) \qquad (16) \\ &= (-\mathcal{I}_{1,sc} - \mathcal{I}_{2,sc})(\hat{\theta}_{s,f} - \theta_0) - \mathcal{I}_{2,sc}(\hat{\theta}_p - \hat{\theta}_{s,f}) + \mathrm{o}_p(n^{-1/2}),\end{aligned}$$

where

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

From the first line to the second line (16), we used $\mathrm{E}[Z_{sc,\text{mis}}(\theta_0)|\mathbf{x}_{\text{obs}};\theta_0] = 0$ and Theorem 1.

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

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

where

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

Therefore, we get

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

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

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

■

Proof of Corollary 4.1. 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

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

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

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

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

In addition,

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

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 4.3.

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 $\mathcal{I}_{1,nc}$:

$$\begin{aligned}\mathcal{I}_{1,nc} &= \mathrm{E}[\nabla_{\tau^\top} Z_{nc,obs}(\tau)]|_{\tau_0} = \mathrm{E}[\nabla_{\tau^\top} z_{nc,obs}(\tau)]|_{\tau_0} = \mathrm{E}[\nabla_{\tau^\top}\{\mathrm{E}[z_{nc}(x,y;\tau)|x_{\mathrm{obs}};\tau]\}]|_{\tau_0} \\ &= \mathrm{E}[\nabla_{\tau^\top} z_{nc}(x,y;\tau)]|_{\tau_0} + \mathrm{E}[z_{nc}(x,y;\tau)\{\nabla_{\tau^\top}\log\bar{q}(x_{\mathrm{mis}}|x_{\mathrm{obs}};\tau)\}]|_{\tau_0} \quad (17) \\ &= \mathcal{I}_{3,nc} - \mathcal{I}_{2,nc}. \quad (18)\end{aligned}$$

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 (18) 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 (18) 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 (18), 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 4.4. By some algebra, as in the proof of Corollary 4.3, 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 Comparison between FINCE and VNCE

Here, we compare FINCE and variational NCE (VNCE) (Rhodes and Gutmann, 2019). The form (11) clearly shows the difference between the estimator proposed in this paper and VNCE (Rhodes and Gutmann, 2019). 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 objective function in VNCE is

$$\begin{aligned}
&\arg\max_s J_{\text{VNCE}}(\tau, s(x_{\mathrm{mis}})) = J_{\text{VNCE}}(\tau, q(x_{\mathrm{mis}}|x_{\mathrm{obs}})) \\
=&\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\log\left\{\frac{1}{1+\frac{q(x_{\mathrm{mis}}|x_{i,\mathrm{obs}})a(x_{i,\mathrm{obs}})}{q(x_{\mathrm{mis}},x_{i,\mathrm{obs}})}}\right\}|x_{i,\mathrm{obs}}\right] + \frac{1}{n}\sum_{j=1}^{n}\log\left\{\frac{a(y_j)}{a(y_j)+\mathrm{E}[q(y_j,y_{j,mis})|y_j]}\right\} \\
=&\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{q(x_{i,\mathrm{obs}})}{q(x_{i,\mathrm{obs}})+a(x_{i,\mathrm{obs}})}\right\} + \frac{1}{n}\sum_{j=1}^{n}\log\left\{\frac{a(y_j)}{a(y_j)+q(y_j)}\right\}, \qquad (19)
\end{aligned}$$

where $q(x_{\mathrm{obs}}) = \int q(x_{\mathrm{mis}}, x_{\mathrm{obs}})\mu(\mathrm{d}x_{\mathrm{mis}})$. On the other hand, the objective function of our proposed estimator is (11). In general, the efficiencies of the two objective function are not directly comparable. In terms of inferences, our proposed methods (FINCE, FISCORE) are more effective than VNCE because in VNCE, it is difficult to achieve the upper bound in (19). In terms of the scalability, VNCE is more scalable than the proposed methods because VNCE does not require any sampling methods.

D 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 $w(x|x_{\mathrm{obs}}) = p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta_0)/b(x)$.

As in the proof of Theorem 2, we have

$$\hat{\theta}_{sc,m} - \theta_0 = -\mathcal{I}_{3,sc}^{-1} Z_{sc,m}(\theta_0|\hat{\theta}_p) + \mathrm{o}_p(n^{-1/2}) \qquad (20)$$

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 (20) are independent. The first term is equal to

$\hat{\theta}_{sc,\infty} - \theta_0$, of which the asymptotic property is shown in Theorem 2. 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 $\hat{\theta}_p = \hat{\theta}_{sc,f}$, 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 $w(x|x_{\mathrm{obs}}) = p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta_0)/b(x)$ and

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

E Extension to multiple imputation: MISCORE and MINCE

Multiple imputation 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 $p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta)$. When the missing data mechanism is MAR, it is easy to sample from $p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\theta)$ using the MCMC based on (8). The algorithm is described as in Algorithm 5. In this paper, this approach is referred to as MISCORE. MINCE is also defined similarly. Nevertheless, we do not recommend Algorithm 5 for the practical reason of its instability and computational burden.

Algorithm 5: MISCORE

1 **repeat**

2 W-step: Take a set of m samples from $x_{\mathrm{mis}}^{*k} \sim p(x_{\mathrm{mis}}|x_{\mathrm{obs}};\hat{\theta}_t)$ using MCMC for each i

3 M-step: Update the solution to the minimizer of the following term with respect to θ as $\hat{\theta}_{t+1}$:

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

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

Dues to the challenges associated with Algorithm 5, we recommend the following algorithm. This algorithm is similar to the one in Levine and Casella (2001). In the

Table 5: 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

original MISCORE, a set of samples is generated at every step. This requires tremendous computational cost and causes instability. In Algorithm 6, by constructing a $\sqrt{n}$–consistent estimator based on FISCORE at each step and updating by MISCORE one time, this limitation is overcome.

Algorithm 6: One step MISCORE

1 **repeat**
2 | Do W-step and M-step in Algorithm 2 (FISCORE)
3 **until** $\hat{\tau}_t$ *converges*;
4 Do W-step and M-step in Algorithm 5 (MISCORE)

Table 4 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 E.1 *When $\hat{\theta}_p = \hat{\theta}_{sc,f}$ 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 2.

Proof of Corollary E.1. 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 $\Pr(\theta|x) \propto \rho(\theta)p(x;\theta)$.

F Variance estimators of FISCORE and FINCE

F.1 FISCORE

The variance estimator of FISCORE in the case of Corollary 4.1 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

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

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

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

Let us denote two matrices: $d_\theta \times d_x$ matrix $K_1(x)$ with elements $\nabla_{x^b}F_a\,(1 \le a \le d_\theta, 1 \le b \le d_x)$ 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 objective 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

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

F.2 FINCE

The variance estimator of FINCE in the case of Corollary 4.3 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

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