Scandinavian
Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics
doi: 10.1111/sjos.12184

Likelihood-based Inference with Missing Data Under Missing-at-Random

SHU YANG
Department of Biostatistics, Harvard T. H. Chan School of Public Health
JAE KWANG KIM
Department of Statistics, Iowa State University

ABSTRACT. Likelihood-based inference with missing data is challenging because the observed log likelihood is often an (intractable) integration over the missing data distribution, which also depends on the unknown parameter. Approximating the integral by Monte Carlo sampling does not necessarily lead to a valid likelihood over the entire parameter space because the Monte Carlo samples are generated from a distribution with a fixed parameter value.
We consider approximating the observed log likelihood based on importance sampling. In the proposed method, the dependency of the integral on the parameter is properly reflected through fractional weights. We discuss constructing a confidence interval using the profile likelihood ratio test. A Newton–Raphson algorithm is employed to find the interval end points. Two limited simulation studies show the advantage of the Wilks inference over the Wald inference in terms of power, parameter space conformity and computational efficiency. A real data example on salamander mating shows that our method also works well with high-dimensional missing data.

Key words: confidence interval, fractional imputation, likelihood ratio, nonresponse, profile likelihood ratio

1. Introduction

Missing data are frequently encountered in practice. Little & Rubin (2002), Molenberghs & Kenward (2007) and Kim & Shao (2013) provide comprehensive overviews of statistical methods handling missing data. In particular, likelihood-based inference plays a central role. Fisher (1925) is the first to give a formula for the maximum likelihood estimator (MLE) of parameters with incomplete data. Thereafter, most of the literature has focused on calculating the MLE and standard errors (Hartley, 1958; Dempster *et al.*, 1977; Louis, 1982; Ibrahim, 1990; Allison, 2001; McLachlan & Krishnan, 2007; Sung & Geyer, 2007). Then, the Wald inference of $\boldsymbol{\theta}$ can be conducted based on

$$\hat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}, \hat{I}_{obs}^{-1}\right), \tag{1}$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ and $\hat{I}_{obs}$ is the observed information matrix evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Since the pioneering work of Wilks (1938), likelihood ratio statistics have been used to obtain confidence regions (Pawitan, 2001; Severini, 2001; Owen, 2001). However, the literature on the likelihood ratio test with missing data is somewhat sparse because the observed likelihood, the marginal density of the observed part of the data, is an integral expression. Exceptions include Meng & Rubin(1992), Rao & Wang(2002), Nielsen(2003) and Qin *et al.*(2009), which investigate asymptotic properties of the Monte Carlo approximation to the likelihood. According to Rao & Wang (2002), the likelihood ratio statistic based on imputed data follows a scaled chi-squared distribution in the limit; nonetheless, finding the quantiles of the limiting distribution is difficult.

We are interested in developing inference methods based on the Wilks theorem:

$$-2\left\{l_{obs}(\boldsymbol{\theta}) - l_{obs}(\hat{\boldsymbol{\theta}})\right\} \sim \chi^2_p \tag{2}$$

where $l_{obs}(\boldsymbol{\theta})$ is the observed log likelihood and p is the dimension of $\boldsymbol{\theta}$. Two attractive features of the Wilks inference are (i) the resulting confidence interval respects the parameter space and (ii) a likelihood-ratio interval is invariant with respect to parameter transformation. For example, if $\boldsymbol{\theta}_L, \boldsymbol{\theta}_U$ is the 95% confidence interval (CI) for $\boldsymbol{\theta}$, then $g(\boldsymbol{\theta}_L), g(\boldsymbol{\theta}_U)$ is the 95% CI for a monotone increasing function $g(\boldsymbol{\theta})$. To our knowledge, the likelihood ratio property (2) has not been shown in the missing data literature. Sung & Geyer (2007) established the asymptotic normality of the Monte Carlo MLE of $\boldsymbol{\theta}$, but they did not discuss the likelihood ratio property (2).

In this paper, we propose an approximation for the observed log likelihood, establish a version of the Wilks theorem and give the asymptotic theory for the profile log likelihood. Using the importance sampling idea, the dependency of the log likelihood on the parameter is reflected properly through fractional weights. Furthermore, the proposed method based on fractional imputation (FI) is computationally attractive compared with iterative methods based on Markov Chain Monte Carlo.

The paper is organized as follows: Section 2 presents the basic setup. Section 3 develops a novel way of constructing an observed log likelihood and establishes two theoretical results for the proposed method. Section 4 describes the Newton–Raphson algorithm for finding the profile likelihood CI endpoints. Section 5 presents two limited simulation studies of constructing CIs and likelihood ratio tests for continuous data. Section 6 reports a real data example using the salamander mating data (McCullagh & Nelder, 1989). Discussion follows in Section 7.

2. Basic setup

Suppose that $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are n independent realizations of a random variable $\mathbf{Y}$ with a parametric distribution function $F(\mathbf{y}) \in \{F_{\boldsymbol{\theta}}(\mathbf{y}); \boldsymbol{\theta} \in \Theta\}$, where Θ is in a p-dimensional Euclidean space. Suppose $\mathbf{y}_i$ has missing values for some i. Let $\mathbf{y}_{i,obs}$ and $\mathbf{y}_{i,mis}$ denote the observed part and the missing part of $\mathbf{y}_i$, respectively. We assume the missing mechanism is ignorable or missing at random (MAR) in the sense of Rubin (1976). Thus, the MLE of $\boldsymbol{\theta}$ can be obtained by maximizing the observed log likelihood function

$$l_{obs}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\{f_{obs,i}(\mathbf{y}_{i,obs}; \boldsymbol{\theta})\} = \sum_{i=1}^{n} \log\left\{\int f(\mathbf{y}_i; \boldsymbol{\theta}) d\mathbf{y}_{i,mis}\right\}, \tag{3}$$

where $f(\mathbf{y}_i; \boldsymbol{\theta})$ is the joint distribution of $\mathbf{y}_i = (\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis})$ and $f_{obs,i}(\mathbf{y}_{i,obs}; \boldsymbol{\theta})$ is the marginal distribution of $\mathbf{y}_{i,obs}$. To simplify the discussion, we assume the observed log likelihood in (3) has a unique maximum almost surely. Maximizing (3) often requires numerical methods because the integral is often intractable. As an alternative, one can consider maximizing

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} E\left\{\log f(\mathbf{y}_i; \boldsymbol{\theta}) | \mathbf{y}_{i,obs}; \boldsymbol{\theta}\right\}, \tag{4}$$

with respect to $\boldsymbol{\theta}$, as suggested by Dempster *et al.* (1977). Louis (1982) showed that, under some regularity conditions, the solution can also be obtained by solving the mean score equation

$$\bar{S}(\boldsymbol{\theta}) = \sum_{i=1}^{n} E\left\{S(\boldsymbol{\theta}; \mathbf{y}_i) | \mathbf{y}_{i,obs}; \boldsymbol{\theta}\right\} = 0, \tag{5}$$

where $S(\boldsymbol{\theta}; \mathbf{y}_i) = \partial \log f(\mathbf{y}_i; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$.

Monte Carlo methods can be used to compute the conditional expectation in (5). The Monte Carlo imputation methods for approximating the mean score function with missing data include the Monte Carlo EM (MCEM) algorithm (Wei & Tanner, 1990; Ibrahim *et al.*, 1999) and multiple imputation (Glynn *et al.*, 1993). In the MCEM algorithm, imputed values are generated by either ordinary Monte Carlo or Markov Chain Monte Carlo for each EM iteration, with heavy computation for both. Moreover, the MCEM sequence of the parameter estimates resulting from each M-step is not guaranteed to converge for a finite imputation size (Booth & Hobert, 1999). Multiple imputation takes the Bayesian approach in which the imputed values are generated from the posterior predictive distribution. However, the convergence to a stable posterior predictive distribution is hard to check and often requires tremendous computation time (Tanner & Wong, 1987).

Once the solution to (5) is obtained, an approximate $(1-\alpha)$ Wald CI can be constructed for each element in $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}}_j \pm z_{1-\alpha}\left\{I^*_{obs}(\hat{\boldsymbol{\theta}})\right\}^{-1/2}_{jj},$$

where $z_{1-\alpha}$ is the upper $(1-\alpha)$-th quantile of the standard normal distribution and $I^*_{obs}(\boldsymbol{\theta})$ is the approximate information matrix of $\boldsymbol{\theta}$ (Louis, 1982; Oakes, 1999; Robins & Wang, 2000). The Wald CI is based on the asymptotic normality of the MLE, which often has poor coverage when the sampling distribution of the MLE is skewed or if the standard error is a poor estimate of the standard deviation of the estimator. The Wilks CI, derived from the asymptotic chi-squared distribution of the likelihood ratio test, appears to be more stable for small samples. If the observed log likelihood function $l_{obs}(\boldsymbol{\theta})$ is known, an approximate $(1-\alpha)$ confidence region for $\boldsymbol{\theta}$ can be constructed as

$$\left\{\boldsymbol{\theta}\in\Theta : -2\{l_{obs}(\boldsymbol{\theta}) - l_{obs}(\hat{\boldsymbol{\theta}})\} \le \chi^2_{p,1-\alpha}\right\},$$

where $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)$-th quantile of the chi-squared distribution with p degrees of freedom. However, computing the observed log likelihood $l_{obs}(\boldsymbol{\theta})$ in (3) is challenging because the integration over the random variable $\mathbf{y}_{i,mis}$ is often intractable, and implicitly depends on $\boldsymbol{\theta}$.

3. Proposed method

To approximate the observed log likelihood in (3), write

$$\begin{aligned} f_{obs,i}(\mathbf{y}_{i,obs};\boldsymbol{\theta}) &= \int \frac{f(\mathbf{y}_i;\boldsymbol{\theta})}{f(\mathbf{y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta})} f(\mathbf{y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta}) d\mathbf{y}_{i,mis} \\ &= E\left\{\frac{f(\mathbf{y}_i;\boldsymbol{\theta})}{f(\mathbf{y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta})}|\mathbf{y}_{i,obs};\boldsymbol{\theta}\right\}. \end{aligned} \tag{6}$$

The typical Monte Carlo approximation for the observed log likelihood is

$$\tilde{l}_{obs}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left\{\frac{1}{m}\sum_{j=1}^{m} \frac{f\left(\mathbf{y}_i^{*(j)};\boldsymbol{\theta}\right)}{f\left(\mathbf{y}_{i,mis}^{*(j)}|\mathbf{y}_{i,obs};\boldsymbol{\theta}\right)}\right\}, \tag{7}$$

where $\mathbf{y}_{mis}^{*(j)}$ $(j = 1,\ldots,m)$ are the Monte Carlo samples generated from $\mathbf{y}_{mis}^{*(j)} \sim f(\mathbf{y}_{mis}|\mathbf{y}_{obs};\hat{\boldsymbol{\theta}})$. The resulting MLE, by maximizing $\tilde{l}_{obs}(\boldsymbol{\theta})$, is consistent to the true parameter value because $\tilde{l}_{obs}(\boldsymbol{\theta})$ is a good approximation for (3) in the neighbourhood of the true parameter value. However, the Wilks theorem does *not* hold for likelihood ratio calculated with (7) because the Monte Carlo samples are generated from a distribution with a fixed parameter value.

To correctly account for the dependency of the observed log likelihood on $\boldsymbol{\theta}$, we use the importance sampling idea to remove the dependency of the imputed values of $\mathbf{y}_{i,mis}$ on $\boldsymbol{\theta}$. Let $\mathbf{y}_{ij}^* = (\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis}^{*(j)})$ be the j-th imputed value of $\mathbf{y}_i$, where $\mathbf{y}_{i,mis}^{*(j)}$ is generated from a proposal distribution $h(\cdot)$, which does not depend on $\boldsymbol{\theta}$. Based on these imputed values, the Monte Carlo approximation to the mean score equation in (5) is

$$\bar{S}^*(\boldsymbol{\theta}) \equiv \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^*(\boldsymbol{\theta}) S(\boldsymbol{\theta}; \mathbf{y}_{ij}^*) = 0, \tag{8}$$

where

$$w_{ij}^*(\boldsymbol{\theta}) = \frac{f(\mathbf{y}_{ij}^*; \boldsymbol{\theta}) / h\left(\mathbf{y}_{i,mis}^{*(j)}\right)}{\sum_{k=1}^{m} \left\{ f\left(\mathbf{y}_{ik}^*; \boldsymbol{\theta}\right) / h\left(\mathbf{y}_{i,mis}^{*(k)}\right) \right\}}$$

is the fractional weight assigned to $\mathbf{y}_{ij}^*$.

The solution to (8) can be obtained by the EM algorithm. In the EM algorithm, the M-step updates the parameter value from $\hat{\boldsymbol{\theta}}^{(t)}$ to $\hat{\boldsymbol{\theta}}^{(t+1)}$ by solving

$$\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^* \left(\hat{\boldsymbol{\theta}}^{(t)}\right) S(\boldsymbol{\theta}; \mathbf{y}_{ij}^*) = 0$$

for $\boldsymbol{\theta}$. Note that the imputed values are not regenerated, only the fractional weights are updated for each EM iteration, and the EM sequence $\{\hat{\boldsymbol{\theta}}^{(0)}, \hat{\boldsymbol{\theta}}^{(1)}, \ldots\}$ converges to a stationary point $\hat{\boldsymbol{\theta}}^*$, which converges to the MLE $\hat{\boldsymbol{\theta}}$ as $m \to \infty$ (Kim, 2011).

Now, write the observed density of $\mathbf{y}_{i,obs}$ in (6) as

$$f_{obs,i}(\mathbf{y}_{i,obs}; \boldsymbol{\theta}) = \int \frac{f(\mathbf{y}_i; \boldsymbol{\theta})}{h(\mathbf{y}_{i,mis})} h(\mathbf{y}_{i,mis}) d\mathbf{y}_{i,mis} = E_h \left\{ \frac{f(\mathbf{y}_i; \boldsymbol{\theta})}{h(\mathbf{y}_{i,mis})} \right\}.$$

In this expression, the expectation is taken over the density h, which does not depend on $\boldsymbol{\theta}$. We can approximate the marginal density of $\mathbf{y}_{i,obs}$ as

$$f_{obs,i}(\mathbf{y}_{i,obs}; \boldsymbol{\theta}) \cong \frac{m^{-1} \sum_{j=1}^{m} \{ f(\mathbf{y}_{ij}^*; \boldsymbol{\theta}) / h(\mathbf{y}_{i,mis}^{*(j)}) \}}{m^{-1} \sum_{j=1}^{m} \{ 1 / h(\mathbf{y}_{i,mis}^{*(j)}) \}} = \frac{1}{\sum_{j=1}^{m} \{ w_{ij}^*(\boldsymbol{\theta}) / f(\mathbf{y}_{ij}^*; \boldsymbol{\theta}) \}}, \tag{9}$$

where we use $m^{-1} \sum_{j=1}^{m} \{ f(\mathbf{y}_{ij}^*; \boldsymbol{\theta}) / h(\mathbf{y}_{i,mis}^{*(j)}) \}$ to approximate $E_h \{ f(\mathbf{y}_i; \boldsymbol{\theta}) / h(\mathbf{y}_{i,mis}) \} = f_{obs,i}(\mathbf{y}_{i,obs}; \boldsymbol{\theta})$ and $m^{-1} \sum_{j=1}^{m} \{ 1 / h(\mathbf{y}_{i,mis}^{*(j)}) \}$ to approximate 1 in the denominator. By such an approximation, we can express the marginal density function of $\mathbf{y}_{i,obs}$ based on the fractional weight function and the joint density function of $\mathbf{y}_i$. Thus, the observed log likelihood function $l_{obs}(\boldsymbol{\theta})$ can be approximated by

$$l_{obs}^*(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log \left[\sum_{j=1}^{m} \{ w_{ij}^*(\boldsymbol{\theta}) / f(\mathbf{y}_{ij}^*; \boldsymbol{\theta}) \} \right]. \tag{10}$$

Now, the log likelihood (10) depends on $\boldsymbol{\theta}$ in two places, in $f(\mathbf{y}_{ij}^*; \boldsymbol{\theta})$ and in the fractional weights $w_{ij}^*(\boldsymbol{\theta})$, which takes full account for the dependency of the log likelihood on $\boldsymbol{\theta}$.

Now, we establish two theoretical results. One is the limiting distribution of the likelihood ratio statistic constructed from $l_{obs}^*(\boldsymbol{\theta})$ in (10). The other is the limiting distribution of the profile likelihood ratio statistic. Theorem 1 presents the limiting distribution of

$W_1 = -2\{l^*_{obs}(\boldsymbol{\theta}_0) - l^*_{obs}(\hat{\boldsymbol{\theta}})\}$ under the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is the true parameter value. We show that the Wilks theorem holds for $l^*_{obs}(\boldsymbol{\theta})$ under certain regularity conditions, and consequently, the likelihood ratio test (LRT) can be constructed from $l^*_{obs}(\boldsymbol{\theta})$.

Theorem 1. *Under the regularity conditions stated in the online Supporting Information S1,*

$$W_1 = -2\left\{l^*_{obs}(\boldsymbol{\theta}_0) - l^*_{obs}(\hat{\boldsymbol{\theta}})\right\} \xrightarrow{d} \chi^2(p),$$

as $m, n \to \infty$.

The proof of Theorem 1 is presented in the online Supporting Information S1. By Theorem 1, we reject the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ if $W_1 > \chi^2_{p,1-\alpha}$ with α being the significance level. Computing the statistic W_1 is easy because $l^*_{obs}(\boldsymbol{\theta})$ is readily computable from the fractionally imputed data. Given the imputed values $\mathbf{y}^*_{ij}$, only $f(\mathbf{y}^*_{ij}; \boldsymbol{\theta})$ and $h(\mathbf{y}^*_{ij})$ are needed to compute $l^*_{obs}(\boldsymbol{\theta})$, which leads to W_1.

We now consider asymptotic properties of profile likelihood ratio statistics for testing $H_0 : \boldsymbol{\theta} = g(\boldsymbol{\nu})$, where $\boldsymbol{\nu}$ is a $(p-q)$-vector of unknown parameters and $g : \mathbb{R}^{p-q} \mapsto \mathbb{R}^p$ is a continuously differentiable function and satisfies $\partial g(\boldsymbol{\nu})/\partial \boldsymbol{\nu}$ is of rank $(p-q)$. For example, if $\boldsymbol{\theta} = (\theta_1, \theta_2)^T \in \mathbb{R}^2$ and $H_0 : \theta_1 = 0$, then $\nu = \theta_2$, $g(\nu) = (g_1(\nu), g_2(\nu)) = (0, \theta_2)^T$ and $\partial g(\nu)/\partial \nu = (0, 1)^T$ with rank 1. In this case, we can use $W_2 = -2\{l^*_{obs}(g(\hat{\boldsymbol{\nu}})) - l^*_{obs}(\hat{\boldsymbol{\theta}})\}$, where $\hat{\boldsymbol{\nu}} = \arg\max_{H_0} l^*_{obs}(g(\boldsymbol{\nu}))$. Theorem 2 presents the limiting distribution of W_2 under H_0.

Theorem 2. *Let* $H_0 : \boldsymbol{\theta} = g(\boldsymbol{\nu})$, *where* $\boldsymbol{\nu}$ *is a* $(p-q)$*-vector of unknown parameters and* $g : \mathbb{R}^{p-q} \mapsto \mathbb{R}^p$ *is a continuously differentiable function, which satisfies* $\partial g(\boldsymbol{\nu})/\partial \boldsymbol{\nu}$ *is of rank* $(p-q)$. *Under the regularity conditions of Theorem 1 and* H_0,

$$W_2 = -2\left\{l^*_{obs}(g(\hat{\boldsymbol{\nu}})) - l^*_{obs}(\hat{\boldsymbol{\theta}})\right\} \to \chi^2(q),$$

as $m, n \to \infty$, *where* $\hat{\boldsymbol{\nu}} = \arg\max_{H_0} l^*_{obs}(g(\boldsymbol{\nu}))$.

The proof of Theorem 2 is presented in the online Supporting Information S2. The MLE $\hat{\boldsymbol{\nu}}$ under H_0 can be obtained by solving $\bar{S}^*(g(\boldsymbol{\nu})) = 0$ for $\boldsymbol{\nu}$, where

$$\bar{S}^*(g(\boldsymbol{\nu})) = \sum_{i=1}^{n}\sum_{j=1}^{m} \tilde{w}^*_{ij}(g(\boldsymbol{\nu}))S(g(\boldsymbol{\nu}); \mathbf{y}^*_{ij}),$$

and

$$\tilde{w}^*_{ij}(g(\boldsymbol{\nu})) = \frac{f\left(\mathbf{y}^*_{ij}; g(\boldsymbol{\nu})\right)/h\left(\mathbf{y}^{*(j)}_{i,mis}\right)}{\sum_{k=1}^{m}\left\{f\left(\mathbf{y}^*_{ik}; g(\boldsymbol{\nu})\right)/h\left(\mathbf{y}^{*(k)}_{i,mis}\right)\right\}}.$$

Remark 1. (Choices of the proposal distribution and the imputation size) The importance sampling requires the proposal distribution h to satisfy the condition that $h(\mathbf{y}_{mis}) > 0$ whenever the target distribution $f(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta}) > 0$ for $\boldsymbol{\theta}$ in the neighbourhood of the true parameter value $\boldsymbol{\theta}_0$. In practice with finite samples and imputations, a well-specified proposal distribution may improve the performance of the imputation estimator. For estimating the population mean of $\mathbf{y}$, Kim (2011) showed that the optimal h^* that minimizes the Monte Carlo approximation variance of $\bar{\mathbf{y}}^*_i \equiv \sum_{j=1}^{M} w^*_{ij}\mathbf{y}^*_{ij}$ is

$$h^*(\mathbf{y}_{i,mis} \mid \mathbf{y}_{i,obs}) = f(\mathbf{y}_{i,mis}|\mathbf{y}_{i,obs}; \hat{\boldsymbol{\theta}})\frac{|\mathbf{y}_i - E(\mathbf{y}_i|\mathbf{y}_{i,obs}; \hat{\boldsymbol{\theta}})|}{E\{|\mathbf{y}_i - E(\mathbf{y}_i|\mathbf{y}_{i,obs}; \hat{\boldsymbol{\theta}})| \mid \mathbf{y}_{i,obs}; \hat{\boldsymbol{\theta}}\}}.$$

Therefore, $h(\mathbf{y}_{i,mis}) = f(\mathbf{y}_{i,mis}|\mathbf{y}_{i,obs};\hat{\boldsymbol{\theta}})$ with a consistent estimator $\hat{\boldsymbol{\theta}}$ is a reasonable choice in terms of statistical efficiency. To remove the dependency of h on $\boldsymbol{\theta}$, one may consider a mixture distribution $h(\mathbf{y}_{i,mis}) = \int f(\mathbf{y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ for some density function $\pi(\boldsymbol{\theta})$ satisfying $\int \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$. Because the t-distribution is the marginal posterior distribution for the mean in a normal distribution with unknown variance (with a flat prior distribution), we can also specify the proposal distribution h as a t-distribution when f is a normal distribution. For example, we can use a t-distribution with small number of degrees of freedom, whose mean and variance match with that of $f(\mathbf{y}_{i,mis}|\mathbf{y}_{i,obs};\hat{\boldsymbol{\theta}})$.

The choice of the imputation size m is a matter of trade-off between statistical efficiency and computation efficiency: small m may lead to large variability in Monte Carlo approximation, whereas large m may increase computational cost. The magnitude of the imputation error is $O(1/\sqrt{m})$, which can be reduced for large m. Thus, if computational power allows, the larger m, the better. The simulation study presented in Section 5.1 showed that the results are not sensitive to reasonable choices of the proposal distribution h and the imputation size m.

Example 1. To illustrate the proposed method, consider bivariate data (x_i, y_i). Assume that $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, and we are interested in testing $H_0 : \beta_1 = 0$. Under MAR, the MLE of $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$ is computed from the complete observations. That is, assuming that the first r units are observed in (x, y) and the remaining $n-r$ units are missing y, we have

$$\hat{\beta}_0 = \bar{y}_r - \hat{\beta}_1 \bar{x}_r,\ \hat{\beta}_1 = S_{xy,r}/S_{xx,r},\ \hat{\sigma}^2 = r^{-1}\sum_{i=1}^{r}\left(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1\right)^2,$$

where $(\bar{x}_r, \bar{y}_r) = r^{-1}\sum_{i=1}^{r}(x_i, y_i)$ and $(S_{xx,r}, S_{xy,r}) = r^{-1}\sum_{i=1}^{r}(x_i - \bar{x}_r)(x_i, y_i)$. Once the MLE $\hat{\boldsymbol{\theta}}$ is obtained, m imputed values of missing y_i, denoted by $y_i^{*(1)}, \ldots, y_i^{*(m)}$, can be generated from $f(y \mid x_i; \hat{\boldsymbol{\theta}})$ and the imputed values are assigned fraction weights $w_{ij}^*(\hat{\boldsymbol{\theta}}) = 1/m$. Then, the maximum of the observed log likelihood under the full model can be approximated by

$$l_{obs}^*(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{m} w_{ij}^*(\hat{\boldsymbol{\theta}})/f(y_i^{*(j)}|x_i, \hat{\boldsymbol{\theta}})\right\} = -\sum_{i=1}^{n}\log\left(\sum_{j=1}^{m}[1/\{mf(y_i^{*(j)}|x_i, \hat{\boldsymbol{\theta}})\}]\right).$$

Under the null hypothesis, $H_0 : \beta_1 = 0$, the MLE of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_{(0)} = (\hat{\beta}_{0(0)}, \hat{\beta}_{1(0)}, \hat{\sigma}^2_{(0)})$, where $\hat{\beta}_{0(0)} = \bar{y}_r$, $\hat{\beta}_{1(0)} = 0$ and $\hat{\sigma}^2_{(0)} = r^{-1}\sum_{i=1}^{r}(y_i - \bar{y}_r)^2$. Thus, the maximum of the observed log likelihood under the null hypothesis can be approximated by

$$l_{obs}^*(\hat{\boldsymbol{\theta}}_{(0)}) = -\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{m} w_{ij}^*\left(\hat{\boldsymbol{\theta}}_{(0)}\right)/f\left(y_i^{*(j)}|x_i, \hat{\boldsymbol{\theta}}_{(0)}\right)\right\},$$

where

$$w_{ij}^*(\hat{\boldsymbol{\theta}}_{(0)}) = \frac{f\left(y_i^{*(j)} \mid x_i; \hat{\boldsymbol{\theta}}_{(0)}\right)/f\left(y_i^{*(j)} \mid x_i; \hat{\boldsymbol{\theta}}\right)}{\sum_{k=1}^{m}\left\{f\left(y_i^{*(k)} \mid x_i; \hat{\boldsymbol{\theta}}_{(0)}\right)/f(y_i^{*(k)} \mid x_i; \hat{\boldsymbol{\theta}})\right\}}.$$

The test statistic for testing $H_0 : \beta_1 = 0$ is computed as $W_2 = -2\left\{l_{obs}^*(\hat{\boldsymbol{\theta}}_{(0)}) - l_{obs}^*(\hat{\boldsymbol{\theta}})\right\}$. If $W_2 > \chi^2_{1,1-\alpha}$, we reject the null hypothesis.

Example 1 does not involve the EM algorithm. The next two examples illustrate the use of the EM algorithm for inference.

Example 2. Consider the following bivariate normal distribution

$$\mathbf{y}_i = \begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right]$$

for $i = 1, \ldots, n$. We are interested in testing $H_0 : \mu_1 = \mu_2$. The joint distribution and the score function are given by

$$f(\mathbf{y_i}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left\{ -\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) \right\},$$

and

$$S(\boldsymbol{\theta}; \mathbf{y}_i) = \begin{pmatrix} \Sigma^{-1}(\mathbf{y_i} - \boldsymbol{\mu})_{2\times 1} \\ -\frac{1}{2}[tr(\Sigma^{-1}\Sigma_1) - (\mathbf{y_i} - \boldsymbol{\mu})^T \Sigma^{-1}\Sigma_1\Sigma^{-1}(\mathbf{y_i} - \boldsymbol{\mu})] \\ -\frac{1}{2}[tr(\Sigma^{-1}\Sigma_2) - (\mathbf{y_i} - \boldsymbol{\mu})^T \Sigma^{-1}\Sigma_2\Sigma^{-1}(\mathbf{y_i} - \boldsymbol{\mu})] \\ -\frac{1}{2}[tr(\Sigma^{-1}\Sigma_3) - (\mathbf{y_i} - \boldsymbol{\mu})^T \Sigma^{-1}\Sigma_3\Sigma^{-1}(\mathbf{y_i} - \boldsymbol{\mu})] \end{pmatrix}_{5\times 1},$$

respectively, where tr is the trace operator, $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)^T$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$, $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, $\Sigma_1 = \begin{pmatrix} 2\sigma_1 & \rho\sigma_2 \\ \rho\sigma_2 & 0 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & 0 \end{pmatrix}$ and $\Sigma_3 = \begin{pmatrix} 0 & \rho\sigma_1 \\ \rho\sigma_1 & 2\sigma_2 \end{pmatrix}$. Assume that there are missing values in y_{1i} and y_{2i}, and the original sample is partitioned into three sets:

H = both y_1 and y_2 are observed K = only y_1 is observed

L = only y_2 is observed

Let n_H, n_K, and n_L denote the size of H, K, and L. Under MAR, from set H, we can obtain an initial value of the parameter $\boldsymbol{\theta}^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)}, \rho^{(0)})^T$. For $i \in K$, m imputed values of missing y_{2i}, denoted by $y_{2i}^{*(1)}, \ldots, y_{2i}^{*(m)}$, can be generated from $h_2(y_{2i}|y_{1i})$, where $h_2(y_{2i}|y_{1i})$ is the distribution of $N(\mu_2^{(0)} + (\sigma_2^{(0)}/\sigma_1^{(0)})\rho^{(0)}(y_{1i} - \mu_1^{(0)}), (1-\rho^{(0)2})\sigma_1^{(0)2})$, and the fractional weights in the t-th EM iteration are given by

$$w_{ij}^*(\boldsymbol{\theta}^{(t)}) = \frac{f(\mathbf{y}_i^{*(j)}; \boldsymbol{\theta}^{(t)})/h_2(y_{2i}^{*(j)}|y_{1i})}{\sum_{k=1}^m \{f(\mathbf{y}_i^{*(k)}; \boldsymbol{\theta}^{(t)})/h_2(y_{2i}^{*(k)}|y_{1i})\}}.$$

For $i \in L$, m imputed values of missing y_{1i}, denoted by $y_{1i}^{*(1)}, \ldots, y_{1i}^{*(m)}$, can be generated from $h_1(y_{1i}|y_{2i})$, where $h_1(y_{1i}|y_{2i})$ is the distribution of $N(\mu_1^{(0)} + (\sigma_1^{(0)}/\sigma_2^{(0)})\rho^{(0)}(y_{2i} - \mu_2^{(0)}), (1-\rho^{(0)2})\sigma_2^{(0)2})$, and the fractional weights in the t-th EM iteration are given by

$$w_{ij}^*(\boldsymbol{\theta}^{(t)}) = \frac{f(\mathbf{y}_i^{*(j)}; \boldsymbol{\theta}^{(t)})/h_1(y_{1i}^{*(j)}|y_{2i})}{\sum_{k=1}^m \{f(\mathbf{y}_i^{*(k)}; \boldsymbol{\theta}^{(t)})/h_1(y_{1i}^{*(k)}|y_{2i})\}}.$$

The parameter is updated from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$ by solving

$$\sum_{i=1}^n w_{ij}^*(\boldsymbol{\theta}^{(t)}) S(\boldsymbol{\theta}; \mathbf{y}_i^{*(j)}) = 0.$$

Then, the maximum of the observed log likelihood under the full model can be approximated by

$$l^*_{obs}(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} w^*_{ij}(\hat{\boldsymbol{\theta}}) / f(\mathbf{y}_i^{*(j)}; \hat{\boldsymbol{\theta}}) \right\},$$

where $\hat{\boldsymbol{\theta}}$ is the convergent point of the EM sequence $\{\boldsymbol{\theta}^{(t)}\}$.

Under $H_0: \mu_1 = \mu_2$, the score function of $\boldsymbol{\theta}_2 = (\mu_2, \sigma_1, \sigma_2, \rho)$ becomes

$$\tilde{S}(\boldsymbol{\theta}_2; \mathbf{y}_i) = \begin{pmatrix} \mathbf{1}^T \Sigma^{-1} (\mathbf{y_i} - \mu_2 \mathbf{1}) \\ -\frac{1}{2}[tr(\Sigma^{-1}\Sigma_1) - (\mathbf{y_i} - \mu_2\mathbf{1})^T \Sigma^{-1}\Sigma_1\Sigma^{-1} (\mathbf{y_i} - \mu_2\mathbf{1})] \\ -\frac{1}{2}[tr(\Sigma^{-1}\Sigma_2) - (\mathbf{y_i} - \mu_2\mathbf{1})^T \Sigma^{-1}\Sigma_2\Sigma^{-1} (\mathbf{y_i} - \mu_2\mathbf{1})] \\ -\frac{1}{2}[tr(\Sigma^{-1}\Sigma_3) - (\mathbf{y_i} - \mu_2\mathbf{1})^T \Sigma^{-1}\Sigma_3\Sigma^{-1} (\mathbf{y_i} - \mu_2\mathbf{1})] \end{pmatrix}_{4\times 1},$$

For $i \in K$, the fractional weights in the t-th EM iteration are given by

$$\tilde{w}^*_{ij}(\boldsymbol{\theta}_2^{(t)}) = \frac{f(\mathbf{y}_i^{*(j)}; \boldsymbol{\theta}_2^{(t)}) / h_2(y_{2i}^{*(j)} | y_{1i})}{\sum_{k=1}^{m} \{ f(\mathbf{y}_i^{*(k)}; \boldsymbol{\theta}_2^{(t)}) / h_2(y_{2i}^{*(k)} | y_{1i}) \}}.$$

For $i \in L$, the fractional weights in the t-th EM iteration are given by

$$\tilde{w}^*_{ij}(\boldsymbol{\theta}_2^{(t)}) = \frac{f(\mathbf{y}_i^{*(j)}; \boldsymbol{\theta}_2^{(t)}) / h_1(y_{1i}^{*(j)} | y_{2i})}{\sum_{k=1}^{m} \{ f(\mathbf{y}_i^{*(k)}; \boldsymbol{\theta}_2^{(t)}) / h_1(y_{1i}^{*(k)} | y_{2i}) \}}.$$

The parameter is updated from $\boldsymbol{\theta}_2^{(t)}$ to $\boldsymbol{\theta}_2^{(t+1)}$ by solving

$$\sum_{i=1}^{n} \tilde{w}^*_{ij}(\boldsymbol{\theta}_2^{(t)}) \tilde{S}(\boldsymbol{\theta}_2; \mathbf{y}_i^{*(j)}) = 0.$$

The MLE of $\boldsymbol{\theta}$ under H_0 is $\hat{\boldsymbol{\theta}}_{(0)} = (\hat{\mu}_{1(0)}, \hat{\boldsymbol{\theta}}_{2(0)})$, where $\hat{\mu}_{1(0)} = \hat{\mu}_{2(0)}$ and $\hat{\boldsymbol{\theta}}_{2(0)}$ is the convergent point of the EM sequence $\{\boldsymbol{\theta}_2^{(t)}\}$. Then, the maximum of the observed log likelihood under the null model can be approximated by

$$l^*_{obs}(\hat{\boldsymbol{\theta}}_{(0)}) = -\sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \tilde{w}^*_{ij}(\hat{\boldsymbol{\theta}}_{2(0)}) / f(\mathbf{y}_i^{*(j)}; \hat{\boldsymbol{\theta}}_{2(0)}) \right\}.$$

The test statistic for testing $H_0: \mu_1 = \mu_2$ is computed as $W_2 = -2\left\{ l^*_{obs}(\hat{\boldsymbol{\theta}}_{(0)}) - l^*_{obs}(\hat{\boldsymbol{\theta}}) \right\}$. If $W_2 > \chi^2_{1,1-\alpha}$, then we reject the null hypothesis.

Example 3. For a categorical data example, we consider a 2×2 table with supplemental margins for both the binary variables Y_1 and Y_2, presented in Table 1. Let $\pi_{ij} = Pr(Y_1 = i, Y_2 = j)$. For orthogonal parametrization, we use $\boldsymbol{\theta} = (\pi_{1|1}, \pi_{1|2}, \pi_{+1})^T$ where $\pi_{1|1} = P(y_2 = 1 | y_1 = 1)$, $\pi_{1|2} = P(y_2 = 1 | y_1 = 2)$, $\pi_{+1} = P(y_1 = 1)$. We are interested in testing $H_0: Y_1$ and Y_2 are independent, which is the same as $H_0: \pi_{1|1} = \pi_{1|2}$. Therefore, this is a full versus reduced model problem. In FI, given the parameter values $\boldsymbol{\theta}^{(t)}$, we take all the possible values as the imputed values for $n_{11,K}$ and compute the conditional probability of $n^*_{11,K} = l$ as the fractional weight $w^{*l}_{11,K} = P(n^*_{11,K} = l | n_{1+,K}, \boldsymbol{\theta}^{(t)})$, which is the probability mass function of a Bernoulli distribution with size $n_{1+,K}$ and probability $\pi^{(t)}_{11}/\pi^{(t)}_{1+}$, where $\pi^{(t)}_{1+} = \sum_{j=1}^{2} \pi^{(t)}_{1j}$. Similarly, we can impute the missing values in other cells. Update

Table 1. *A* 2×2 *table with supplemental margins for both variables* y_1 *and* y_2

Set	y_1	y_2	Count
	1	1	$n_{11,H} = 100$
H	1	2	$n_{12,H} = 50$
	2	1	$n_{21,H} = 75$
	2	2	$n_{22,H} = 75$
K	1		$n_{1+,K} = 30$
	2		$n_{2+,K} = 60$
L		1	$n_{+1,L} = 28$
		2	$n_{+2,L} = 60$

the parameters by $\pi_{i|j}^{(t+1)} = n_{ij}^{*(t)}/n_{+j}^{*(t)}$, where $n_{ij}^{*(t)} = n_{ij,H} + \sum_{i\in K} w_{ij,K}^{*l(t)} + \sum_{i\in L} w_{ij,L}^{*l(t)}$ and $n_{+j}^{*(t)} = \sum_{i=1}^{2} n_{ij}^{*(t)}$. Then, the maximum of the observed log likelihood under the full model can be approximated as

$$l_{obs}^{*}(\hat{\boldsymbol{\theta}}) = \sum_{H} n_{ij,H} \log \hat{\pi}_{ij} + \sum_{K} n_{i+,K} \log \hat{\pi}_{i+} + \sum_{L} n_{+j,L} \log \hat{\pi}_{+j},$$

where $\hat{\boldsymbol{\theta}}$ is the convergent point of the EM sequence $\{\boldsymbol{\theta}^{(t)}\}$.

Under the null model where $H_0 : \pi_{1|1} = \pi_{1|2}$, the MLE of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_{(0)} = (\hat{\pi}_{1|1(0)}, \hat{\pi}_{1|2(0)}, \hat{\pi}_{+1(0)})^T$ where $\hat{\pi}_{1|1(0)} = \hat{\pi}_{1|2(0)} = (n_{11,H} + n_{12,H})/n_{++,H}$, and $\hat{\pi}_{+1(0)} = (n_{+1,H} + n_{+1,L})/(n_{++,H} + n_{++,L})$. Then, the maximum of the observed log likelihood under the full model can be approximated as

$$l_{obs}^{*}(\hat{\boldsymbol{\theta}}_{(0)}) = \sum_{H} n_{ij,H} \log \hat{\pi}_{ij(0)} + \sum_{K} n_{i+,K} \log \hat{\pi}_{i+(0)} + \sum_{L} n_{+j,L} \log \hat{\pi}_{+j(0)}.$$

The test statistic for testing $H_0 : \pi_{1|1} = \pi_{1|2}$ is computed as $W_2 = -2\{l_{obs}^{*}(\hat{\boldsymbol{\theta}}_{(0)}) - l_{obs}^{*}(\hat{\boldsymbol{\theta}})\}$. If $W_2 > \chi_{1,1-\alpha}^2$, we reject the null model.

For the data in Table 1, under the full model, $\hat{\pi}_{11} = 0.279$, $\hat{\pi}_{12} = 0.174$, $\hat{\pi}_{21} = 0.238$ and $\hat{\pi}_{22} = 0.308$. The maximum of the observed log likelihood under the full model is $l_{obs}^{*}(\hat{\boldsymbol{\theta}}) = -532$. Under the null model, $\hat{\pi}_{11(0)} = 0.262$, $\hat{\pi}_{12(0)} = 0.238$, $\hat{\pi}_{21(0)} = 0.262$ and $\hat{\pi}_{22(0)} = 0.238$. The maximum of the observed log likelihood under the null model is $l_{obs}^{*}(\hat{\boldsymbol{\theta}}_{(0)}) = -538$. Thus, $W_2 = 12 > \chi_{1,0.95}^2$ and the null hypothesis is rejected.

4. Computation details

We now discuss computing the Wilks interval based on the profile likelihood ratio test in the presence of nuisance parameters using the results in Section 3. Specifically, we consider the case of $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2)$ when θ_1 is a scalar and $\boldsymbol{\theta}_2$ is a vector of parameters, and discuss constructing a Wilks confidence interval for θ_1 based on the profile log likelihood. Under a complete data setting, the profile log likelihood for θ_1 is defined as $l_p(\theta_1) = l\left((\theta_1, \hat{\boldsymbol{\theta}}_2(\theta_1))\right)$, where $l(\boldsymbol{\theta})$ is the log likelihood function of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_2(\theta_1)$ maximizes $l(\boldsymbol{\theta})$ for each fixed θ_1. An approximate $(1-\alpha)$ Wilks CI for θ_1 is $\{\theta_1 : 2\{l(\hat{\boldsymbol{\theta}}) - l_p(\theta_1)\} \le \chi_{1,1-\alpha}^2\}$.

Obtaining a profile CI often requires repeated computation of $l_p(\theta_1)$ over a grid of values of θ_1 or a systematic search procedure such as the bisection of the interval. Both approaches are cumbersome. We now assume that the observed log likelihood surface is asymptotically quadratic and discuss a more efficient way of computing the endpoints of the CI by finding the roots of the equation

$$l_p(\theta_1) - h = 0, \tag{11}$$

where $h = l(\hat{\boldsymbol{\theta}}) - \chi^2_{1,1-\alpha}/2$. Venzon & Moolgavkar (1988) showed that the solution to equation (11) can be obtained by solving the following system of equations:

$$\left(l_p(\theta_1) - h,\ \frac{\partial l}{\partial \boldsymbol{\theta}_2}(\boldsymbol{\theta}) \right)^T = 0. \tag{12}$$

To use a Newton–Raphson algorithm to solve the system of equations, we need to calculate the derivative of each element in (12). Note that (12) specifies

$$l_p(\theta_1) = l(\theta_1, \boldsymbol{\theta}_2(\theta_1)), \tag{13}$$

and

$$\frac{\partial l}{\partial \boldsymbol{\theta}_2}(\theta_1, \boldsymbol{\theta}_2(\theta_1)) = 0. \tag{14}$$

By the chain rule, from (13), we have

$$\frac{d l_p(\theta_1)}{d\theta_1} = \frac{\partial l}{\partial \theta_1} + \frac{\partial l}{\partial \boldsymbol{\theta}_2^T} \frac{d\boldsymbol{\theta}_2}{d\theta_1}, \tag{15}$$

and from (14), we have

$$\frac{\partial^2 l}{\partial \theta_1 \partial \boldsymbol{\theta}_2} + \frac{\partial^2 l}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T} \frac{d\boldsymbol{\theta}_2}{d\theta_1} = 0. \tag{16}$$

Solving $d\boldsymbol{\theta}_2/d\theta_1$ from (16) and substituting in (15), we have

$$\frac{d l_p(\theta_1)}{d\theta_1} = \frac{\partial l}{\partial \theta_1} - \frac{\partial l}{\partial \boldsymbol{\theta}_2^T} \left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T} \right)^{-1} \frac{\partial^2 l}{\partial \theta_1 \partial \boldsymbol{\theta}_2}.$$

Then, applying the Newton–Raphson algorithm, the solution to the system of equations (12) can be obtained as follows:

$$\begin{pmatrix} \theta_1^{(t+1)} \\ \boldsymbol{\theta}_2^{(t+1)} \end{pmatrix} = \begin{pmatrix} \theta_1^{(t)} \\ \boldsymbol{\theta}_2^{(t)} \end{pmatrix} - \begin{pmatrix} \frac{\partial l}{\partial \theta_1} - \frac{\partial l}{\partial \boldsymbol{\theta}_2^T} \left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T} \right)^{-1} \frac{\partial^2 l}{\partial \theta_1 \partial \boldsymbol{\theta}_2} & 0 \\ \frac{\partial^2 l}{\partial \theta_1 \partial \boldsymbol{\theta}_2^T} & \frac{\partial^2 l}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T} \end{pmatrix}^{-1} \begin{pmatrix} l - h \\ \frac{\partial l}{\partial \boldsymbol{\theta}_2} \end{pmatrix} \Big|_{(\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})}, \tag{17}$$

where the superscripts (t) on $\theta_1, \boldsymbol{\theta}_2$ and $\boldsymbol{\theta}$ denote values at the t-th iteration. Under a missing data setting, by Theorem 1, $l^*_{obs}(\boldsymbol{\theta})$ well-approximates $l_{obs}(\boldsymbol{\theta})$ and a similar result holds for the corresponding profile log likelihood for θ_1. Thus, (17) can be computed as

$$\begin{aligned} \begin{pmatrix} \theta_1^{(t+1)} \\ \boldsymbol{\theta}_2^{(t+1)} \end{pmatrix} &= \begin{pmatrix} \theta_1^{(t)} \\ \boldsymbol{\theta}_2^{(t)} \end{pmatrix} - \begin{pmatrix} \frac{\partial l^*_{obs}}{\partial \theta_1} - \frac{\partial l^*_{obs}}{\partial \boldsymbol{\theta}_2^T} \left(\frac{\partial^2 l^*_{obs}}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T} \right)^{-1} \frac{\partial^2 l^*_{obs}}{\partial \theta_1 \partial \boldsymbol{\theta}_2} & 0 \\ \frac{\partial^2 l^*_{obs}}{\partial \theta_1 \partial \boldsymbol{\theta}_2^T} & \frac{\partial^2 l^*_{obs}}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T} \end{pmatrix}^{-1} \begin{pmatrix} l^*_{obs} - h^* \\ \frac{\partial l^*_{obs}}{\partial \boldsymbol{\theta}_2} \end{pmatrix} \Big|_{(\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})} \\ &= \begin{pmatrix} \theta_1^{(t)} \\ \boldsymbol{\theta}_2^{(t)} \end{pmatrix} - \begin{pmatrix} \bar{S}^*_1 - \bar{S}^*_2 \left(\frac{\partial^2 l^*_{obs}}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T} \right)^{-1} \frac{\partial^2 l^*_{obs}}{\partial \theta_1 \partial \boldsymbol{\theta}_2} & 0 \\ \frac{\partial^2 l^*_{obs}}{\partial \theta_1 \partial \boldsymbol{\theta}_2^T} & \frac{\partial^2 l^*_{obs}}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T} \end{pmatrix}^{-1} \begin{pmatrix} l^*_{obs} - h^* \\ \bar{S}^*_2 \end{pmatrix} \Big|_{(\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})}, \end{aligned}$$

where $h^* = l^*_{obs}(\hat{\boldsymbol{\theta}}) - \chi^2_{1,1-\alpha}/2$, $\partial^2 l^*_{obs}/(\partial \theta_1 \partial \boldsymbol{\theta}_2^T)$ and $\partial^2 l^*_{obs}/(\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T)$ are the $(1,2)$-th and $(2,2)$-th partitions in $I^*_{obs}(\boldsymbol{\theta})$, respectively.

5. Simulation study

To test our theory, we performed two simulation studies, one on constructing CIs and the other one on testing hypotheses.

5.1. Profile likelihood confidence interval

In the first simulation study, we investigated the performance of the Wilks CI based on the profile likelihood ratio test. $B = 2,000$ simulation samples were generated from $y_i = 2 + x_i + e_i$, $x_i \sim N(1,1)$, $e_i \sim N(0,1)$, x_i and e_i are independent, with x_i fully observed and y_i subject to missing. In addition, we generated δ_i, the response indicator variable of y_i, from Bernoulli(p_i) with $logit(p_i) = 0.45 + 0.1x_i$. Under this response model, the missing mechanism is MAR and the average response rate is about 0.6. We constructed 95% CIs for β_1 and σ^2 using two different methods: the Wald method based on the asymptotic normality and the Wilks method based on the result of Theorem 2. In this setting, the observed log likelihood is analytically available; we denoted the procedure that uses the analytical expression by Method 'l_{obs}'. In the FI method that uses importance sampling, we considered two factors: the proposal distribution and the imputation size. The proposal distributions are h_1, the conditional distribution

Table 2. *Monte Carlo length and coverage of the Wald confidence intervals (Wald CI) and the Wilks confidence intervals (Wilks CI) for β_1 and σ^2 over 2,000 simulated datasets: (i) Method l_{obs} is using the analytical log likelihood function; (ii) Method h_1 is the proposed method with the proposal distribution $f(y|x;\hat{\boldsymbol{\theta}}^{(0)})$; and (iii) Method h_2 is the proposed method with a t proposal distribution with 3 degrees of freedom and its mean and variance match with that of $f(y_i|x_i;\hat{\boldsymbol{\theta}}^{(0)})$*

			β_1				σ^2			
			95% Wald CI		95% Wilks CI		95% Wald CI		95% Wilks CI	
Method	n	m	Length	Coverage ×100	Length	Coverage ×100	Length	Coverage ×100	Length	Coverage ×100
l_{obs}	20	—	1.26	95	1.21	92	1.58	96	1.65	88
	50	—	0.72	95	0.70	94	0.98	95	0.99	92
	100	—	0.50	96	0.49	94	0.69	95	0.69	94
h_1	20	100	1.11	88	1.17	91	1.33	76	1.62	90
	20	500	1.11	89	1.19	91	1.33	76	1.64	90
	20	1000	1.11	89	1.19	91	1.33	76	1.65	90
	50	100	0.69	94	0.70	94	0.92	87	0.99	93
	50	500	0.69	94	0.70	94	0.92	87	0.99	93
	50	1000	0.69	94	0.70	94	0.92	87	0.99	93
	100	100	0.49	94	0.49	94	0.67	94	0.69	95
	100	500	0.49	94	0.49	94	0.67	95	0.69	95
	100	1000	0.49	94	0.49	94	0.67	95	0.69	95
h_2	20	100	1.12	89	1.20	91	1.33	76	1.67	90
	20	500	1.12	89	1.20	91	1.33	76	1.67	90
	20	1000	1.12	89	1.19	91	1.33	76	1.66	90
	50	100	0.69	94	0.70	94	0.92	87	0.99	93
	50	500	0.69	94	0.70	94	0.92	87	1.00	93
	50	1000	0.69	94	0.70	94	0.92	87	1.00	93
	100	100	0.493	94	0.492	94	0.67	94	0.69	95
	100	500	0.494	94	0.492	95	0.67	95	0.69	94
	100	1000	0.494	94	0.492	94	0.67	95	0.69	95

$f(y_i|x_i;\hat{\boldsymbol{\theta}}^{(0)})$, where $\hat{\boldsymbol{\theta}}^{(0)}$ is an initial estimate of $\boldsymbol{\theta}$ from the respondent set, and h_2, a t distribution with 3 degrees of freedom and its mean and variance match with that of $f(y_i|x_i;\hat{\boldsymbol{\theta}}^{(0)})$; the imputation sizes are $m = 100, 500$ and1000.

Table 2 shows the Monte Carlo average length and coverage of the 95% CIs for β_1 and σ^2. Compared with the analytical procedure, the Wald inference resulting from FI performs worse than that from Method 'l_{obs}', especially for small sample sizes $n = 20$ and 50, but its performance improves as n becomes larger. On the other hand, the Wilks inference resulting from FI is comparable to that from Method 'l_{obs}' for all sample sizes, which indicates that the Wilks inference is more stable than the Wald inference, especially for small samples.

For sensitivity analysis, the results from FI with the proposal distributions h_1 , h_2 and the imputation size $m = 100$, 500 and 1000 are pretty similar (see Table 2 for detailed numerical results). Therefore, the FI method is not sensitive to reasonable choices of the proposal distribution and the imputation size. In the following discussion, we focus on the FI method with the proposal distribution h_1 and $m = 100$. If the sampling distribution of the estimator is approximately normal, the Wald CI and the Wilks CI are comparable. This is the case for $\hat{\beta}_1$. Figures 1–3 in the online Supporting Information show the sampling distributions of the FI estimator of β_1 with sample sizes $n = 20,\ 50$ and 100. The sampling distributions of $\hat{\beta}_1$ are quite symmetric across the sample sizes. In such cases, the Wald CIs and the Wilks CIs perform equally well. However, the Wilks CI shows advantage over the Wald CI if the distribution of the parameter estimates is skewed. Figure 4–6 in the online Supporting Information show the sampling distributions of $\hat{\sigma}^2$ with sample sizes $n = 20,\ 50$ and 100. The sampling distribution of the FI estimator of σ^2 is skewed to the left when $n = 20$, which explains the low coverage of the Wald CI with $n = 20$ in Table 2. As the sample size n increases, the sampling distribution of $\hat{\sigma}^2$ becomes more symmetric and the coverage becomes closer to the nominal coverage. In the simulation study, when $n = 20$, about 8% of Monte Carlo Wald CIs have negative values for σ^2. The Wilks CIs are generally free from the problems described ealier. Although the Wilks CIs are slightly wider than the Wald CIs, the Wilks CIs feature a closer proximity to the nominal coverage level.

5.2. *Likelihood ratio test*

Robustness of LRT against non-normality. In this simulation study, we investigated the robustness of the Wilks inference based on the result of Theorem 2. $B = 2,000$ samples of size $n = 100$ were generated from $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$, where

$$\mathbf{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right],$$

and $e_i \sim (\chi^2(2) - 2)/2$, with $\mathbf{x}_i$ and e_i being independent. In addition, we generated δ_i, the response indicator variable of y_i, from Bernoulli (p_i) with $logit(p_i) = 0.5 + 0.2x_{1i} + 0.2x_{2i}$. Under this response model, the missing mechanism is MAR and the average response rate is about 0.6. We were interested in testing the null hypothesis $H_0 : \beta_2 = 0$ using the Wilks inference based on the result of Theorem 2 and the Wald inference based on the asymptotic normality. The true parameter values were set to be $(\beta_0, \beta_1) = (-2, 1)$ and β_2 being either 0 or 0.5. The proposal distribution is a t-distribution with 3 degrees of freedom and its mean and variance match with that of $f(y_i|x_i;\hat{\boldsymbol{\theta}}^{(0)})$, where $\hat{\boldsymbol{\theta}}^{(0)}$ is the initial estimate of $\boldsymbol{\theta}$ from the respondent set. The imputation size m is 100.

Table 3. *Monte Carlo type I error and power of the Wilks inference and the Wald inference for continuous data with sample size* $n = 100$

	$\alpha = 0.05$		$\alpha = 0.1$	
Parameter value	Wilks	Wald	Wilks	Wald
$\beta_2 = 0$	0.061	0.025	0.108	0.052
$\beta_2 = 0.3$	0.520	0.398	0.645	0.507
$\beta_2 = 0.5$	0.901	0.750	0.930	0.844

Table 3 shows the Monte Carlo type I error and power of the Wilks inference and the Wald inference, which were calculated as the relative frequency of rejecting the null hypothesis over the simulated datasets at two significance levels, $\alpha = 0.05$ and $\alpha = 0.1$. In the case of $\beta_2 = 0$, the sizes of the Wilks inference are close to the significant levels; however, the sizes of the Wald inference are 0.025 and 0.052 for $\alpha = 0.05$ and $\alpha = 0.1$, respectively. Thus, the Wald inference does not provide correct coverage under departure from the normality assumption. Figure 1 shows the histograms of p-values from the Wilks inference and the Wald inference from the samples generated from the null model. As we can see from Fig. 1, p-values from the Wilks inference (right panel) are uniformly distributed; however, p-values from the Wald inference (left panel) are not, which demonstrates that the Wilks inference is more robust than the Wald inference against non-normality. The robustness of the Wilks inference has been discussed by Kent (1982). Moreover, in the cases of $\beta_2 \neq 0$, the Wilks inference shows more power in detecting $\beta_2 \neq 0$ than the Wald counterpart, and its power increases as β_2 increases.

Bivariate normal with missing values. In this simulation study, we compared the Wilks inference based on the result of Theorem 2 with existing likelihood-based inference methods.

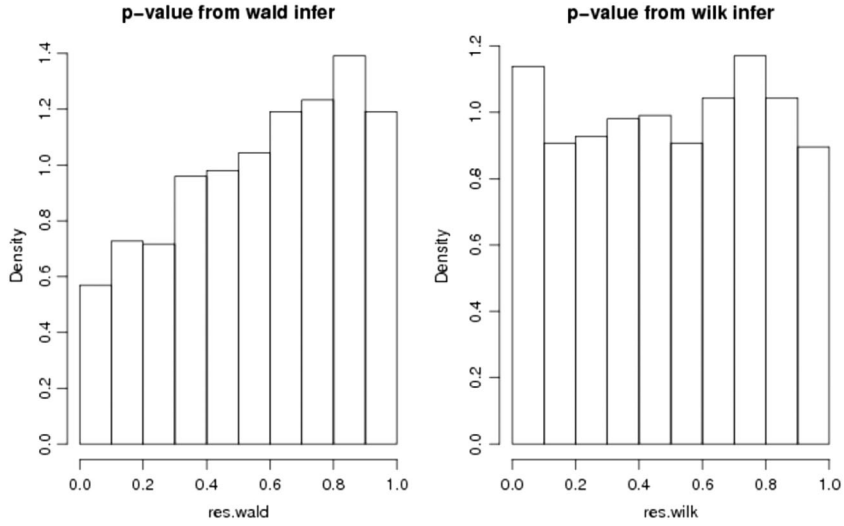

Fig. 1. Histograms of p-value from the Wald-inference and Wilks-inference.

$B = 2,000$ samples of size $n = 200$ were generated from a bivariate normal distribution

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}\right].$$

In addition, we generated δ_i, the response indicator variable of y_i, from Bernoulli(p_i). In this simulation, we considered two factors: the missing mechanism and the true parameter values. For the missing mechanism, we considered (i) Missingness completely at random (MCAR), where $p_i = 0.6$ for all i and (ii) MAR, where $logit(p_i) = 1 - 1.5x_i$. The true parameter values were set to be $(\sigma_{xx}, \sigma_{xy}, \sigma_{yy}) = (1, -0.5, 1)$ and $(\mu_x, \mu_y) \in \{(0,0), (-0.1, 0.1), (0, 0.2), (-0.2, 0), (-0.3, 0.3)\}$. We were interested in testing the null hypothesis $H_0 : \mu_x = \mu_y$ using LRT of the full sample (FULL), complete cases (CC), multiple imputation (MI) (Meng & Rubin,1992) with the imputation size $m = 1000$, and the proposed FI method using the importance sampling, where the proposal distribution is a t-distribution with 3 degrees of freedom and its mean and variance match with that of $f(y_i|x_i; \hat{\boldsymbol{\theta}}^{(0)})$, where $\hat{\boldsymbol{\theta}}^{(0)}$ is the initial estimate of $\boldsymbol{\theta}$ from the respondent set, and the imputation size $m = 1000$.

Table 4 shows the Monte Carlo type 1 error and power of LRTs of FULL, CC, MI and FI. Under MCAR, when $\mu_x = \mu_y = 0$, the sizes of all the tests are close to the significant levels; when $\mu_x \neq \mu_y$, CC loses power by discarding all the incomplete cases, compared with FI. Under MAR, when $\mu_x = \mu_y = 0$, FI controls type 1 error correctly, but CC does not provide correct coverages. The sizes of FI are close to the significant levels; however, the sizes of CC are 0.9977 for $\alpha = 0.05$ and 0.9995 for $\alpha = 0.1$. Because under the response model, units with larger x_i are less likely to report y_i, which leads to the LRT based on complete cases rejects the null hypothesis. Figure 2 shows the histograms of p-values from CC and FI under the null hypothesis. As we can see in Fig. 2, p-values from FI (right panel) are uniformly distributed; however, p-values from CC (left panel) are not. When $\mu_x \neq \mu_y$, FI is more powerful than MI consistently in all scenarios, and the power of FI is close to that of FULL, which suggests that FI is a powerful method. Moreover, as the difference between μ_x and μ_y increases, the power of FI also increases.

Table 4. *Monte Carlo type 1 error and power of likelihood ratio tests of $H_0 : \mu_x = \mu_y$ over 2,000 simulated datasets: FULL is based on the full sample; CC is based on the complete cases; MI is the multiple imputation method of Meng and Rubin (1992); FI is the proposed fractional imputation method using importance sampling with a t proposal distribution with 3 degrees of freedom and its mean and variance match with that of $f(y_i|x_i; \hat{\boldsymbol{\theta}}^{(0)})$*

		$\alpha = 0.05$				$\alpha = 0.1$			
	(μ_x, μ_y)	FULL	CC	MI	FI	FULL	CC	MI	FI
MCAR	**(0, 0)**	0.0468	0.0508	0.0494	0.0504	0.0988	0.1006	0.1041	0.1040
	(−0.1, 0.1)	0.3632	0.2466	0.2326	0.3464	0.4890	0.3508	0.3484	0.4574
	(−0.2, 0)	0.3632	0.2466	0.2326	0.3488	0.4890	0.3508	0.3484	0.4538
	(0, 0.2)	0.3632	0.2466	0.2326	0.3426	0.4890	0.3508	0.3484	0.4576
	(−0.3, 0.3)	0.9986	0.9652	0.6652	0.9964	0.9994	0.9852	0.8492	0.9992
MAR	**(0, 0)**	0.0468	0.9977	0.0483	0.0501	0.0988	0.9995	0.0977	0.0995
	(−0.1, 0.1)	0.3632	0.9998	0.2207	0.3210	0.4890	0.9998	0.3038	0.4490
	(−0.2, 0)	0.3632	0.9984	0.2183	0.3324	0.4890	0.9996	0.3222	0.4522
	(0, 0.2)	0.3632	0.9998	0.2128	0.3296	0.4890	0.9998	0.3164	0.4486
	(−0.3, 0.3)	0.9986	1	0.6509	0.9948	0.9994	1	0.8419	0.9984

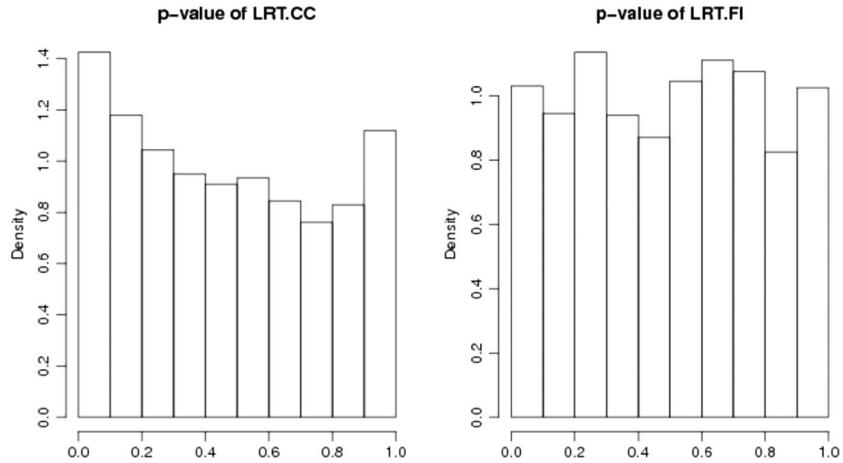

Fig. 2. Histogram of p-value from likelihood ratio test based on complete cases (LRT.CC) and that based on fractional imputation (LRT.FI).

6. Real data example

The salamander data of McCullagh & Nelder (1989) came from the experiment aimed to study the extent to which mountain dusky salamanders from different populations would interbreed. The data given here refer to two populations called Rough Butt (R) and Whiteside (W). Forty animals were used in each of three experiments, one conducted in the summer of 1986 and two in the Fall of the same year. The forty salamanders available in each of the three experiments were composed of 10 R males, 10 R females, 10 W males and 10 W females. Although there were 400 possible crosses between the females and males in each experiment, only 120 of these were permitted by the design. So totally, they observed 360 potential matings. The design of the experiment permits a comparison of the mating probabilities from the four possible crosses: RR, RW, WR and WW.

For the total 360 observations in the dataset, we consider models for the observed data conditionally on the actual animals used in the experiment. Denote y_{ij} to be a random variable representing the binary response indicator of a successful mating between the i-th female and the j-th male for $i,\ j = 1, 2, \ldots, 60$ where only 360 of the (i, j) pairs are relevant (each i corresponds to six j's). Let u_i^f denote the random effect that the i-th female salamander has cross matings in which she is involved, and define u_j^m similarly for the j-th male. Let $\mathbf{x}_{ij}$ denote a four-dimensional vector of covariates indicating the type of cross for the mating pair between female i and male j. We assume that the y_{ij}'s are all conditionally independent and follow a Binomial regression model

$$y_{ij}|u_i^f,\ u_j^m \sim \text{Bernoulli}(\pi_{ij}),$$

and

$$\eta_{ij} = g(\pi_{ij}) = logit(\pi_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i^f + u_j^m,$$

where $g(\cdot)$ is the logit link function, $\boldsymbol{\beta} = (\beta_{RR},\ \beta_{RW},\ \beta_{WR},\ \beta_{WW})^T$ is an unknown four-dimensional regression parameter vector, which are the fixed effects and u_i^f and u_j^m are the random effects. Assume $u_i^f \sim N(0,\ \sigma_f^2)$ and $u_j^m \sim N(0,\ \sigma_m^2)$.

Let $\mathbf{y}$ be the full data vector, and $\mathbf{u}^f$ and $\mathbf{u}^m$ be two 60-variate random variables with parametric densities $g_1(\mathbf{u}^f|\sigma_f^2)$ and $g_2(\mathbf{u}^m|\sigma_m^2)$, respectively. The joint distribution of

$(\mathbf{y}, \mathbf{u}^f, \mathbf{u}^m)$ is

$$\left\{ \prod_{i=1}^{60} \prod_{j=i1}^{i6} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{1-y_{ij}} \right\} g_1(\mathbf{u}^f|\sigma_f^2) g_2(\mathbf{u}^m|\sigma_m^2),$$

where $\pi_{ij} = \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\text{،}} + u_i^f + u_j^m) / \left\{1 + \exp(\mathbf{x}_{ij}^T{}_{\text{،}} \boldsymbol{\beta} + u_i^f + u_j^m)\right\}$. The likelihood function for $\boldsymbol{\gamma} = (\boldsymbol{\beta},\ \sigma_f^2,\ \sigma_m^2)$ is then

$$L(\boldsymbol{\gamma}|\mathbf{y}) = \int \int \left\{ \prod_{i=1}^{60} \prod_{j=i1}^{i6} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{y_{ij}} \right\} g_1(\mathbf{u}^f|\sigma_f^2) g_2(\mathbf{u}^m|\sigma_m^2) d\mathbf{u}^f d\mathbf{u}^m,$$

which involves intractable integrals whose dimension is 120.

To apply the proposed method, we consider a 120-dimensional multivariate t importance distribution, as suggested by Booth & Hobert (1999), with 3 degrees of freedom and its mean and variance match the mode and curvature of the target distribution $f(\mathbf{u}|\mathbf{y};\boldsymbol{\gamma}), \mathbf{u} = (\mathbf{u^f}, \mathbf{u^m})$. To specify such a t-distribution, write $f(\mathbf{u}|\mathbf{y};\boldsymbol{\gamma}) = a \exp\{l(\mathbf{u})\}$, where a is the normalizing constant and let $l^{(i)}(\mathbf{u})$ be the i-th derivative of $l(\mathbf{u})$. Let $\tilde{\mathbf{u}}$ denote the maximizer of $l(\mathbf{u})$, which solves the equation $l^{(1)}(\mathbf{u}) = 0$. The Laplace approximations of the mean and variance of $f(\mathbf{u}|\mathbf{y};\boldsymbol{\gamma})$ are $\tilde{\mathbf{u}}$ and $-l^{(2)}(\tilde{\mathbf{u}})$, respectively (Booth & Hobert,1999). Let $\mathbf{u}^{*(1)}, \ldots, \mathbf{u}^{*(M)}$ be a random sample from $h(\mathbf{u}|\mu, \Sigma)$, which is a multivariate t-distribution with 3 degrees of freedom, $\mu = \tilde{\mathbf{u}}$ and $\Sigma = -l^{(2)}(\tilde{\mathbf{u}})$, the fractional weight is given by

$$w^{*(k)}(\boldsymbol{\gamma}) = \frac{f(y, \mathbf{u}^{*(k)}|\boldsymbol{\gamma}) / h(\mathbf{u}^{*(k)}|\mu, \Sigma)}{\sum_{l=1}^{M} f(y, \mathbf{u}^{*(l)}|\boldsymbol{\gamma}) / h(\mathbf{u}^{*(l)}|\mu, \Sigma)},$$

and the FI estimator of γ is obtained by maximizing the approximate observed log likelihood

$$l^*(\boldsymbol{\gamma}) = -\log \left[\sum_{k=1}^{M} \left\{ w^{*(k)}(\boldsymbol{\gamma}) / f(\mathbf{y}, \mathbf{u}^{*(k)}|\boldsymbol{\gamma}) \right\} \right].$$

We first investigated the performance of the proposed FI method by simulation. The simulation design was based on the actual salamander dataset. In the data generating model, the true parameter values were set to be $(\beta_{RR}, \beta_{RW}, \beta_{WR}, \beta_{WW}) = (1.03, 0.32, -1.95, 0.99)$, $\sigma_f^2 = 1.18$ and $\sigma_m^2 = 1.12$. For comparison, we computed the approximated MLE where the integral in the observed log likelihood is approximated by the Laplace approximation, implemented by 'glmer' function in the software R. We did not consider other Monte Carlo methods because their computation are too heavy to carry out in simulations. Table 5 shows the Monte Carlo mean (variance) of fixed effects and variances of random effects over 1,000 simulated

Table 5. *Monte Carlo mean (variance) of fixed effects and variances of random effects over 1,000 simulated salamander datasets: Method 'Laplace' is the approximated MLE where the integral is approximated by the Laplace approximation and Method 'Proposed' is the proposed method using a multivariate t-distribution (Booth and Hobert, 1999), for imputation sizes m = 1000, 5000, and 10,000*

Method	m	β_{RR}	β_{RW}	β_{WR}	β_{WW}	σ_f^2	σ_m^2
True	—	1.03	0.32	−1.95	0.99	1.18	1.12
Laplace	—	1.07(0.163)	0.31(0.124)	−1.88(0.173)	0.96(0.128)	1.03(0.300)	0.88(0.249)
Proposed	1000	1.06(0.077)	0.34(0.055)	−1.92(0.086)	0.97(0.061)	1.16(0.040)	1.06(0.031)
	5000	1.06(0.078)	0.34(0.057)	−1.92(0.085)	0.97(0.060)	1.15(0.040)	1.06(0.032)
	10000	1.06(0.077)	0.34(0.056)	−1.92(0.085)	0.97(0.061)	1.15(0.041)	1.06(0.032)

Table 6. *Parameter estimates from salamander dataset (observations = 360)*

Method	β_{RR}	β_{RW}	β_{WR}	β_{WW}	σ_f^2	σ_m^2
MLE	1.03	0.32	−1.95	0.99	1.18	1.12
Gibbs	1.03	0.34	−1.98	1.07	1.49	1.37
MCMLE	1.00	0.53	−1.78	1.27	1.10	1.17
Proposed	1.00	0.34	−1.85	0.95	1.13	1.04

salamander datasets. The proposed method produces satisfactory results with smaller biases and variances for most of the parameters compared with the Laplace approximation method.

We applied the proposed method on the actual salamander dataset. Table 6 shows the results from the proposed estimator and various other estimation methods, including the MLE given by Booth & Hobert (1999), the MLE from a modified EM algorithm with Gibbs sampling method (Karim & Zeger (1992)), and the Monte Carlo MLE method (Sung & Geyer (2007)). The Gibbs sampling approach tends to produce larger estimates than the MLE. Moreover, Gibbs sampling involves heavy computation, which is not desirable in practice. Using the importance sampling significantly reduces the computation time. Based on Monte Carlo sample size 10^7, Monte Carlo MLE from Sung & Geyer (2007) is $\hat{\boldsymbol{\beta}} = (1.00, 0.53, -1.78, 1.27)$ and $(\hat{\sigma}_f^2, \hat{\sigma}_m^2) = (1.10, 1.17)$. Our proposed method improves the estimation of Sung & Geyer (2007) to $\hat{\boldsymbol{\beta}} = (1.00, 0.34, -1.85, 0.95)$ and $(\hat{\sigma}_f^2, \hat{\sigma}_m^2) = (1.13, 1.04)$, which is close to the MLE.

7. Discussion

We have developed an approximation for the observed log likelihood function over the entire parameter space with missing data using the importance sampling idea. The proposed observed log likelihood can be used to perform likelihood ratio tests and construct Wilks CIs. One attractive feature of the proposed Wilks method is that unlike the Wald method, we do not necessarily need the observed information matrix, but only the functional form of the fractional weight and the joint density of variables are needed.

The proposed method of computing the observed log likelihood can be directly applied to model selection or model comparison with missing data. For example, the Bayesian Information Criterion (BIC) of Schwarz (1978) under missing data can be computed by

$$\mathrm{BIC}(M) = 2l_{obs}(\hat{\boldsymbol{\theta}}|M) - (\log n)\dim(M),$$

for each candidate model M, where $l_{obs}(\hat{\boldsymbol{\theta}}|M)$ is the value of the log likelihood evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, the MLE of $\boldsymbol{\theta}$ under model M, $\dim(M)$ is the number of parameters estimated in model M and n is the sample size of the complete data. Thus, model selection using AIC or BIC is a straightforward extension of this research. Further investigation on this topic will be presented elsewhere.

Acknowledgements

We thank one anonymous referee and the Associate Editor for their constructive comments, which have helped to improve the quality of the paper. The research of the second author was partially supported by a grant from NSF (MMS-121339).

References

Allison, DA. (2001). *Missing data*, Sage Publication, New York.

Booth, JG & Hobert, JP. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol* **61**, 265–285.

Dempster, AP, Laird, NM & Rubin, DB. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **39**, 1–38.
Fisher, RA. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.
Glynn, RJ, Laird, NM & Rubin, DB. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association* **88**, 984–993.
Hartley, HO. (1958). Maximum likelihood estimation from incomplete data. *Biometics* **14**, 174–194.
Ibrahim, JG. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85**, 765–769.
Ibrahim, JG, Lipsitz, SR & Chen, MH. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 173–190.
Karim, MR & Zeger, SL. (1992). Generalized linear models with random effects: salamander mating revisited. *Biometrika* **48**, 631–644.
Kent, JT. (1982). Robust properties of likelihood ratio tests. *Biometrika* **69**, 19–27.
Kim, JK. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98**, 119–132.
Kim, JK & Shao, J. (2013). *Statistical methods for handling incomplete data*, Chapman & Hall / CRC, London.
Little, RJ & Rubin, DB. (2002). *Statistical analysis with missing data*, J Wiley & Sons, New York.
Louis, TA. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **44**, 226–233.
McCullagh, P & Nelder, JA. (1989). *Generalized linear models*, Chapman & Hall, London.
McLachlan, G & Krishnan, T. (2007). *The EM algorithm and extensions*, John Wiley & Sons, New York.
Meng, XL & Rubin, DB. (1992). Performing likelihood ratio tests with multiply-imputed datasets. *Biometrika* **79**, 103–111.
Molenberghs, G & Kenward, MG. (2007). Missing data in clinical studies. *Journal of Tropical Pediatrics* **53**, 294.
Nielsen, S.F. (2003). Proper and improper multiple imputation. *International Statistical Review* **71**, 593–607.
Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 479–482.
Owen, BA. (2001). *Empirical likelihood*, Chapman & Hall, New York.
Pawitan, Y. (2001). *In all likelihood: statistical modeling and inference using likelihood*, Oxford University Press, Oxford.
Qin, J, Zheng, B & Leung, D.H.Y. (2009). Empirical likelihood in missing data problems. *Journal of the American Statistical Association* **104**, 1492–1503.
Rao, JNK & Wang, Q. (2002). Empirical likelihood-based inference under imputation for missing response data. *Annals of Statistics* **30**, 896–924.
Rubin, DB. (1976). Inference and missing Data. *Biometrika* **63**, 581–592.
Robins, JM & Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.
Severini, TA. (2001). *Likelihood methods in Statistics*, Oxford Univ. Press, Oxford.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
Sung, YJ & Geyer, CJ. (2007). Monte Carlo likelihood inference for missing data models. *Annals of Statistics* **38**, 990–1011.
Tanner, MA & Wong, WH. (1987). The calculation of posterior distribution by data augmentation. *J. Amer. Statist. Assoc* **82**, 528–540.
Venzon, DJ & Moolgavkar, SH. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **37**, 87–94.
Wei, GC & Tanner, MA. (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.
Wilks, SS. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* **9**, 1–67.

Received October 2014, in final form August 2015

Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA, USA.
E-mail: jkim@iastate.edu

Supporting information

All supporting information may be found in the online version of this article. Additional information for this article is available online including: Appendix S1: regularity conditions and proof of Theorem 1; Appendix S2: proof of Theorem 2; Appendix S3: proof of Lemma C; and Appendix S4 figures in the simulation studies.