Ann Inst Stat Math (2016) 68:329–351
DOI 10.1007/s10463-014-0498-1

Expectation-robust algorithm and estimating equations for means and dispersion matrix with missing data

Ke-Hai Yuan · Wai Chan · Yubin Tian

Received: 21 February 2014 / Revised: 5 October 2014 / Published online: 25 November 2014

Abstract Means and covariance/dispersion matrix are the building blocks for many statistical analyses. By naturally extending the score functions based on a multivariate t-distribution to estimating equations, this article defines a class of M-estimators of means and dispersion matrix for samples with missing data. An expectation-robust (ER) algorithm solving the estimating equations is obtained. The obtained relationship between the ER algorithm and the corresponding estimating equations allows us to obtain consistent standard errors when robust means and dispersion matrix are further analyzed. Estimating equations corresponding to existing ER algorithms for computing M- and S-estimators are also identified. Monte Carlo results show that robust methods outperform the normal-distribution-based maximum likelihood when the population distribution has heavy tails or when data are contaminated. Applications of the results to robust analysis of linear regression and growth curve models are discussed.

Keywords Missing data · Monte Carlo · Robust means and dispersion matrix · Sandwich-type covariance matrix

K.-H. Yuan (✉)
Department of Psychology, University of Notre Dame,
Notre Dame, IN 46556, USA
e-mail: kyuan@nd.edu

W. Chan
Department of Psychology, The Chinese University of Hong Kong,
Shatin, New Territories, Hong Kong

Y. Tian
School of Mathematics, Beijing Institute of Technology,
Haidian District, Beijing 100081, China

1 Introduction

Means and covariance/dispersion matrix are among the most important concepts in statistics. They are essential in describing the distribution of a population or sample. They are also the building blocks in most widely used statistical methods (e.g., ANOVA, regression, correlations, factor analysis, principal component analysis, structural equation modeling, growth curves, etc.). Most topics in applied multivariate statistics can be regarded as the analyses of sample means and/or covatiance matrix (e.g., Johnson and Wichern 2002). However, real data tend to have heavy tails (Micceri 1989) and the sample means and covariance matrix can be very inefficient. In particular, with missing data that are even all missing at random (MAR) (Rubin 1976), biases in the normal-distribution-based maximum likelihood (NML) estimates (NMLEs) can be greater than the values of the population parameters, due to the interaction between heavy-tailed distribution and missing data (Yuan et al. 2012). In such a situation, robust estimates are desired. Robust procedures have been systematically introduced in textbooks (Hampel et al. 1986; Heritier et al. 2009; Huber 1981; Maronna et al. 2006; Wilcox 2012). Robust estimates of means and dispersion matrix with missing values have been developed using maximum likelihood (ML) based on multivariate t- or contaminated-normal distributions (Little 1988). However, either of the ML procedures might not be the best method when the underlying population distribution is unknown. Other M-estimators, S-estimator, and/or those obtained from certain hybrid-methods might be preferred (see e.g., Mehrotra 1995).

When robust estimates of means and dispersion matrix are subject to further analysis, we need to have a consistent estimator of their covariance matrix to obtain consistent standard errors (SEs) for the derived parameter estimates or proper test statistics for overall model evaluation. If the robust means and dispersion matrix satisfy a set of estimating equations, then a consistent sandwich-type covariance matrix of the robust estimates directly follows from the estimating equations (Godambe 1960; Huber 1967; Yuan and Jennrich 1998). Thus, it is important to relate robust estimates to estimating equations. With complete data, robust M-estimators of means and dispersion matrix are typically defined by estimating equations (Maronna 1976). With missing data, they have been presented as the output of expectation-robust (ER) algorithms in which certain weights are attached to cases with imputed data (Little and Smith 1987; Cheng and Victoria-Feser 2002). It is also necessary to identify their corresponding estimating equations if inference is needed in their applications.

The paper has four goals: (1) generalizing the maximum likelihood estimates with missing data based on a multivariate t-distribution to M-estimators using estimating equations; (2) providing an ER algorithm to solve the estimating equations; (3) identifying the estimating equations corresponding to existing algorithms for computing robust means and dispersion matrix with missing data; (4) comparing bias and efficiency of different robust estimators defined through estimating equations with missing values. We will review relevant literature for robust estimation with missing data in the development. But comparing all the existing robust methods theoretically or numerically is not our goal. Statistical theory suggests that it is impossible to identify the best method for a real data set whose population distribution is unknown.

In Sect. 2 we extend the estimating equations based on the multivariate t-distribution to those defining general M-estimators for samples with missing values. Special cases of the equations are also satisfied by S-estimators for samples with missing values. We then give the ER algorithm for solving the estimating equations. Estimating equations corresponding to algorithms for calculating robust means and dispersion matrix in the literature are also identified and discussed. Monte Carlo results concerning the efficiency of several robust estimators are presented in Sect. 3. Applications of the results to robust analysis of linear regression and growth curve models are considered in Sect. 4. We end the paper by discussing issues related to applications of robust estimation in practice.

2 Expectation-robust algorithm and estimating equations

Let $\mathbf{x}$ represent a population of p random variables. A sample $\mathbf{x}_i$, $i = 1, 2, \ldots, n$, from $\mathbf{x}$ is obtained. Due to missing values, $\mathbf{x}_i$ only contains p_i marginal realizations of $\mathbf{x}$. We are interested in estimating the means and dispersion matrix of $\mathbf{x}$ by a robust method. Let $\mathbf{x}_{im}$ be the vector containing the $p - p_i$ missing values. For notational convenience, we will use $\mathbf{x}_{ic} = (\mathbf{x}_i', \mathbf{x}_{im}')'$ to denote the complete data. Of course, the positions of missing values are not always at the end in practice. We can perform a permutation on each missing pattern so that all the algebraic operations in this article still hold. With the sample $\mathbf{x}_i$, $i = 1, 2, \ldots, n$, in mind, we will first present the EM algorithm based on a multivariate t-distribution and then extend it to an ER algorithm solving general estimating equations.

Let $Mt_p(\boldsymbol{\mu}, \Sigma, m)$ denote the p-variate t-distribution with m degrees of freedom, where $\boldsymbol{\mu}$ is the mean vector and Σ is the dispersion matrix. When $m > 2$, the maximum likelihood estimate (MLE) of $\mathrm{Cov}(\mathbf{x}) = m\Sigma/(m-2)$ can be obtained as $m\hat{\Sigma}/(m-2)$ with $\hat{\Sigma}$ being the MLE of Σ. Because the purpose of modeling with a multivariate t-distribution is mostly for robustness rather than regarding the data as truly coming from a t-distribution, many applications just directly work with $\hat{\Sigma}$ rather than $m\hat{\Sigma}/(m-2)$ in further analysis (e.g., Devlin et al. 1981). Actually, most statistical analyses based on $\hat{\Sigma}$ or a rescaling of it yield the same results.

To introduce the EM algorithm based on $\mathbf{x} \sim Mt_p(\boldsymbol{\mu}, \Sigma, m)$ with a given m, let $\boldsymbol{\mu}^{(j)}$ and $\Sigma^{(j)}$ be the values of $\boldsymbol{\mu}$ and Σ at the jth iteration, $\boldsymbol{\mu}_i^{(j)}$ and $\Sigma_i^{(j)}$ be the means and dispersion matrix corresponding to the observed $\mathbf{x}_i$. When $p_i < p$, we have

$$\boldsymbol{\mu}^{(j)} = \begin{pmatrix} \boldsymbol{\mu}_i^{(j)} \\ \boldsymbol{\mu}_{im}^{(j)} \end{pmatrix} \quad \text{and} \quad \Sigma^{(j)} = \begin{pmatrix} \Sigma_i^{(j)} & \Sigma_{iom}^{(j)} \\ \Sigma_{imo}^{(j)} & \Sigma_{imm}^{(j)} \end{pmatrix}, \tag{1}$$

where $\boldsymbol{\mu}_{im}^{(j)}$ corresponds to the means of $\mathbf{x}_{im}$; $\Sigma_{imm}^{(j)}$ and $\Sigma_{imo}^{(j)}$ correspond to the dispersion matrices of $\mathbf{x}_{im}$ with itself and with $\mathbf{x}_i$, respectively. Notice that the elements of $\boldsymbol{\mu}^{(j)}$ and $\Sigma^{(j)}$ in (1) are the same for all the observations, and the subscript i is used to indicate that cases may have different number of missing values. Let

$$d_i^2 = d^2(\mathbf{x}_i, \boldsymbol{\mu}_i, \Sigma_i) = (\mathbf{x}_i - \boldsymbol{\mu}_i)'\Sigma_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i)$$

be the Mahalanobis distance for the observed $\mathbf{x}_i$, and $(d_i^{(j)})^2 = d^2(\mathbf{x}_i, \boldsymbol{\mu}_i^{(j)}, \Sigma_i^{(j)})$. The E-step of the EM algorithm based on $\mathbf{x} \sim Mt_p(\boldsymbol{\mu}, \Sigma, m)$ in Little (1988) obtains the weight $w_i^{(j)} = (m + p_i)/\left[m + \left(d_i^{(j)}\right)^2\right]$, the conditional means

$$\hat{\mathbf{x}}_{ic}^{(j)} = E_j(\mathbf{x}_{ic}|\mathbf{x}_i) = \begin{pmatrix} \mathbf{x}_i \\ \hat{\mathbf{x}}_{im}^{(j)} \end{pmatrix}, \tag{2}$$

and the conditional covariance matrix[1]

$$\mathbf{C}_i^{(j)} = \mathrm{Cov}_j(\mathbf{x}_{ic}|\mathbf{x}_i) = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{imm}^{(j)} \end{pmatrix}, \tag{3}$$

where

$$\hat{\mathbf{x}}_{im}^{(j)} = \boldsymbol{\mu}_{im}^{(j)} + \Sigma_{imo}^{(j)}(\Sigma_i^{(j)})^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i^{(j)}) \text{ and } \mathbf{C}_{imm}^{(j)} = \Sigma_{imm}^{(j)} - \Sigma_{imo}^{(j)}(\Sigma_i^{(j)})^{-1}\Sigma_{iom}^{(j)}.$$

The M-step gives

$$\boldsymbol{\mu}^{(j+1)} = \frac{\sum_{i=1}^n w_i^{(j)} \hat{\mathbf{x}}_{ic}^{(j)}}{\sum_{i=1}^n w_i^{(j)}}, \tag{4}$$

$$\Sigma^{(j+1)} = \frac{\sum_{i=1}^n \left[w_i^{(j)} (\hat{\mathbf{x}}_{ic}^{(j)} - \boldsymbol{\mu}^{(j+1)})(\hat{\mathbf{x}}_{ic}^{(j)} - \boldsymbol{\mu}^{(j+1)})' + \mathbf{C}_i^{(j)} \right]}{n}. \tag{5}$$

The robustness of an M-estimator may depend on the starting values for the EM algorithm. We will discuss choices of starting values at the end of this section. At the convergence of the EM algorithm, we obtain the MLEs $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ based on the multivariate t-distribution with m degrees of freedom. Notice that the n in the denominator of (5) can be replaced by $\sum_{i=1}^n w_i^{(j)}$, which makes the EM algorithm converge faster (Kent et al. 1994; Meng and Dyk 1997).

Clearly, the t-distribution-based MLEs satisfy the estimating equations obtained by setting the score functions corresponding to $\mathbf{x}_i \sim Mt_{p_i}(\boldsymbol{\mu}_i, \Sigma_i, m)$ at zero. Let $w_i = (m + p_i)/(m + d_i^2)$,

$$\mathbf{V}_i = 2^{-1}(\Sigma_i^{-1} \otimes \Sigma_i^{-1}),$$

and $\boldsymbol{\sigma} = \mathrm{vech}(\Sigma)$ be the vector containing the elements in the low-triangular part of Σ. The estimating equations corresponding to $\mathbf{x}_i \sim Mt_{p_i}(\boldsymbol{\mu}_i, \Sigma_i, m), i = 1, 2, \ldots, n$, are given by

[1] The EM algorithm presented here is slightly different from that in Little (1988), where a conditional normal distribution is used.

$$\sum_{i=1}^{n} w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\mu}} \Sigma_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{6}$$

and

$$\sum_{i=1}^{n} \frac{\partial \mathrm{vec}'(\Sigma_i)}{\partial \boldsymbol{\sigma}} \mathbf{V}_i \mathrm{vec}[w_i(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_i - \boldsymbol{\mu}_i)' - \Sigma_i] = \mathbf{0}. \tag{7}$$

Notice that Eqs. (6), (7) and others that we call estimating equations in this article only involve the observed values $\mathbf{x}_i$, not the estimated component $\hat{\mathbf{x}}_{im} = E(\mathbf{x}_{im}|\mathbf{x}_i, \boldsymbol{\mu}, \Sigma)$, which are consistent with the estimating equation literature (e.g., Godambe 1991; Liang and Zeger 1986; Prentice and Zhao 1991).

As noted in the introduction, the MLEs corresponding to $\mathbf{x} \sim Mt_p(\boldsymbol{\mu}, \Sigma, m)$ may not be even asymptotically efficient unless the true underlying population follows the multivariate t-distribution. Many approaches have been proposed to obtain robust estimates of means and dispersion matrix with complete data (e.g., Maronna 1976; Maronna and Zamar 2002; Mehrotra 1995). In particular, both M-estimators and S-estimators[2] satisfy a set of estimating equations (Lopuhaä 1989; Rocke 1996). A natural generalization of (6) and (7) to accommodating different weights in estimating means and dispersion matrix of the observed data is given by

$$\sum_{i=1}^{n} w_{i1} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\mu}} \Sigma_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{8}$$

and

$$\sum_{i=1}^{n} \frac{\partial \mathrm{vec}'(\Sigma_i)}{\partial \boldsymbol{\sigma}} \mathbf{V}_i \mathrm{vec}[w_{i2}(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_i - \boldsymbol{\mu}_i)' - w_{i3}\Sigma_i] = \mathbf{0}, \tag{9}$$

where $w_{i1} = w_{i1}(d_i)$, $w_{i2} = w_{i2}(d_i)$ and $w_{i3} = w_{i3}(d_i)$ are typically nonincreasing functions of d_i.

Obviously, (6) and (7) are a special case of (8) and (9) when $w_{i1} = w_{i2} = (m + p_i)/(m + d_i^2)$ and $w_{i3} = 1$. Let $0 < \varphi < 1$ and r_i be the $(1 - \varphi)$th quantile corresponding to χ_{p_i}, the chi-distribution with p_i degrees of freedom. Equations (8) and (9) extend Huber-type M-estimators to samples with missing data when letting

$$w_{i1} = w_{i1}(d_i) = \begin{cases} 1, & \text{if } d_i \le r_i, \\ r_i/d_i, & \text{if } d_i > r_i, \end{cases} \tag{10}$$

$w_{i2} = w_{i1}^2/\tau_i$ and $w_{i3} = 1$, where τ_i is a constant such that $E[\chi_{p_i}^2 w_{i1}^2(\chi_{p_i})/\tau_i] = p_i$. They also extend the elliptical-distribution-based MLEs discussed in Kano et al. (1993) to samples with missing values when $w_{i3} = 1$, and $w_{i1} = w_{i2} = w_i(d_i)$ corresponds to the density function of the elliptical distribution.

[2] An S-estimator is not defined by estimating equations but by minimizing $|\Sigma|$ under a proper constraint.

Among studies of S-estimators, Tukey's biweight function

$$\rho_c(t) = \begin{cases} t^2/2 - t^4/(2c^2) + t^6/(6c^4), & |t| \le c, \\ c^2/6, & |t| > c \end{cases} \tag{11}$$

is most widely used (e.g., Lopuhaä 1989; Rocke 1996), where c is a tuning constant. With p_i variables being observed in $\mathbf{x}_i$, let $w_{i1} = \dot{\rho}_{c_i}(d_i)/d_i$, $w_{i2} = p_i w_{i1}$ and $w_{i3} = \rho_{c_i}(d_i)d_i - \rho_{c_i}(d_i) + b_i$, where $\dot{\rho}_c(d) = \partial\rho_c(d)/\partial d$ and $b_i = E[\rho_{c_i}(\chi_{p_i})]$. Then, Eqs. (8) and (9) are natural extension of equation (2.6) of Lopuhaä (1989) that S-estimators of $\boldsymbol{\mu}$ and Σ need to satisfy with complete data. In particular, for each observed pattern of the sample, the left sides of Eqs. (8 and (9) are mathematically equivalent to the left sides of the two equations in (2.6) of Lopuhaä if we let c_i be the same for all the observations within an observed pattern. Notice that the large sample breakdown point of the S-estimator for complete data is given by $6b/c^2$. We may choose c_i so that $6b_i/c_i^2$ is the same across all the observed patterns. Also notice that there might be multiple solutions to Eqs. (8) and (9), all of them are M-estimators but only one is an S-estimator (Tyler 1991). Multiple starting values might be needed to find the S-estimator (Ruppert 1992) that corresponds to the minimum value of $|\Sigma| > 0$.

The generality of (8) and (9) is that they are simply estimating equations. Unlike (6) and (7), the estimating equations may not correspond to the score functions of a particular log likelihood. Thus, the EM algorithm based on the multivariate t-distribution in (2) to (5) does not apply to (8) and (9). However, a slight modification of (4) and (5) yields solutions to (8) and (9). Specifically, the E-step is the same as in (2) and (3). Let $w_{i1}^{(j)}$, $w_{i2}^{(j)}$ and $w_{i3}^{(j)}$ be evaluated at $d_i^{(j)}$. The M-step is replaced by

$$\boldsymbol{\mu}^{(j+1)} = \frac{\sum_{i=1}^n w_{i1}^{(j)} \hat{\mathbf{x}}_{ic}^{(j)}}{\sum_{i=1}^n w_{i1}^{(j)}}, \tag{12}$$

$$\Sigma^{(j+1)} = \frac{\sum_{i=1}^n [w_{i2}^{(j)} (\hat{\mathbf{x}}_{ic}^{(j)} - \boldsymbol{\mu}^{(j+1)})(\hat{\mathbf{x}}_{ic}^{(j)} - \boldsymbol{\mu}^{(j+1)})' + w_{i3}^{(j)} \mathbf{C}_i^{(j)}]}{\sum_{i=1}^n w_{i3}^{(j)}}. \tag{13}$$

Following Little and Smith (1987), we will call (12) and (13) the robust (R) step. Notice that the ER algorithm in (2), (3), (12) and (13) is a special case of the iteratively reweighted least squares algorithm, whose convergence properties are studied by Green (1984). Denote $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ as the converged values of the ER algorithm. The proof for $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ to solve Eqs. (8) and (9) is given in Appendix A.

Equations (8) and (9) can also be solved using the Newton-Raphson algorithm (Kelley 2003), which involves the derivatives of each term in (8) and (9) with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. Notice that $w_{i1} = w_{i1}(d_i)$, $w_{i2} = w_{i2}(d_i)$, and $w_{i3} = w_{i3}(d_i)$ are functions of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, and their derivatives have to be computed at every iteration in addition to themselves. Also notice that $\boldsymbol{\mu}_i$ and Σ_i corresponding to different subsets of $\mathbf{x}$ contain distinct elements, and the derivatives need to be coded separately for each observed pattern in addition to accounting for different observed values of $\mathbf{x}_i$. Thus, although the Newton-Raphson algorithm can be used to solve (8) and (9), its coding is more

involved than that of the ER algorithm. It is also possible for the Newton-Raphson algorithm to take longer time than the ER algorithm to reach a convergence.

We now discuss the convergence properties of $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ as $n \to \infty$, which are different from the convergence properties of the ER or Newton-Raphson algorithm that yielded $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$. When $\mathbf{x}$ follows an elliptical distribution and without missing values, $\hat{\boldsymbol{\mu}}$ is consistent and $\hat{\Sigma}$ converges to $\kappa\Sigma$ for certain $\kappa > 0$, where Σ is the dispersion matrix of the elliptical distribution (Maronna 1976). With missing values that are MAR, the estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ are consistent and asymptotically most efficient when the left sides of Eqs. (8) and (9) are the score functions corresponding to the true population distribution of $\mathbf{x}_i$ by ignoring the missing values (Rubin 1976). For other scenarios, let $\mathbf{g}(\boldsymbol{\nu}) = (\mathbf{g}_1'(\boldsymbol{\nu}), \mathbf{g}_2'(\boldsymbol{\nu}))'$ with $\mathbf{g}_1(\boldsymbol{\nu})$ and $\mathbf{g}_2(\boldsymbol{\nu})$ being defined as the summation of functions on the left sides of (8) and (9), respectively, where $\boldsymbol{\nu} = (\boldsymbol{\mu}', \boldsymbol{\sigma}')'$. Then, under a set of regularity conditions (e.g., Yuan and Jennrich 1998), the estimate $\hat{\boldsymbol{\nu}} = (\hat{\boldsymbol{\mu}}', \hat{\boldsymbol{\sigma}}')'$ obtained by the ER algorithm converges to a vector $\boldsymbol{\nu}^*$ that satisfies $E[\mathbf{g}(\boldsymbol{\nu}^*)] = \mathbf{0}$, where the expectation is with respect to the true distribution[3] of each observed $\mathbf{x}_i$. Since the true population distribution of the observed sample is typically unknown in practice, nor is the missing data mechanism behind each missing value, it might be hard to know the exact properties of $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ in a specific application. We will use Monte Carlo simulation to evaluate the efficiency of different estimators in the next section.

Little and Smith (1987) give an ER algorithm for computing robust means and dispersion matrix but they do not provide the corresponding estimating equations. Their E-step is the same as in (2) and (3), and their R-step is

$$\boldsymbol{\mu}^{(j+1)} = \frac{\sum_{i=1}^n w_i^{(j)} \hat{\mathbf{x}}_{ic}^{(j)}}{\sum_{i=1}^n w_i^{(j)}}, \tag{14}$$

$$\Sigma^{(j+1)} = \frac{\sum_{i=1}^n \left[\left(w_i^{(j)}\right)^2 \left(\hat{\mathbf{x}}_{ic}^{(j)} - \boldsymbol{\mu}^{(j+1)}\right)\left(\hat{\mathbf{x}}_{ic}^{(j)} - \boldsymbol{\mu}^{(j+1)}\right)' + \mathbf{C}_i^{(j)} \right]}{\sum_{i=1}^n \left(w_i^{(j)}\right)^2 - 1}, \tag{15}$$

where the weight function $w_i^{(j)} = w_i(d_i^{(j)})$ or $w_i = w_i(d_i)$ is given in Little and Smith (1987). Using Eqs. (31) and (32) in the appendix of this article, it can be shown that (14) and (15) solve the Eq. (6) and

$$\begin{aligned} &\sum_{i=1}^n \frac{\partial \mathrm{vec}'(\Sigma_i)}{\partial \boldsymbol{\sigma}} \mathbf{V}_i \mathrm{vec}[w_i^2 (\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_i - \boldsymbol{\mu}_i)' - \Sigma_i] \\ &\quad + \sum_{i=1}^n (1 - w_i^2) \frac{\partial \mathrm{vec}'(\Sigma)}{\partial \boldsymbol{\sigma}} \mathrm{vec}(\Sigma^{-1}) + \frac{\partial \mathrm{vec}'(\Sigma)}{\partial \boldsymbol{\sigma}} \mathrm{vec}(\Sigma^{-1}) = \mathbf{0}. \end{aligned} \tag{16}$$

[3] The true distribution of the observed $\mathbf{x}_i$ will be different from the corresponding marginal distributions of $\mathbf{x}$ when the missing values are either missing at random or not at random.

Cheng and Victoria-Feser (2002) proposed an ER algorithm to compute S-estimators with missing data in their Eqs. (22) and (23), which can be written as

$$\sum_{i=1}^{n} w_{i1}\Sigma_i^{-1}(\hat{\mathbf{x}}_{ic} - \boldsymbol{\mu}) = \mathbf{0} \tag{17}$$

and

$$\sum_{i=1}^{n} \left\{ w_{i2}[(\hat{\mathbf{x}}_{ic} - \boldsymbol{\mu})(\hat{\mathbf{x}}_{ic} - \boldsymbol{\mu})' + \mathbf{C}_i] - w_{i3}\Sigma \right\} = \mathbf{0}. \tag{18}$$

Using the results in Appendix A, one can show that the solution to (17) satisfies (8), and the solution to (18) satisfies

$$\sum_{i=1}^{n} \left\{ \frac{\partial \text{vec}'(\Sigma_i)}{\partial \boldsymbol{\sigma}} \mathbf{V}_i \text{vec}\{w_{i2}[(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_i - \boldsymbol{\mu}_i)' - \Sigma_i]\} + (w_{i2} - w_{i3}) \frac{\partial \text{vec}'(\Sigma)}{\partial \boldsymbol{\sigma}} \text{vec}(\Sigma) \right\} = \mathbf{0}. \tag{19}$$

Cheng and Victoria-Feser (2002) also proposed a modification to (15) of Little and Smith's R-step, which is given by

$$\Sigma^{(j+1)} = \frac{\sum_{i=1}^{n} (w_i^{(j)})^2 \left[\left(\hat{\mathbf{x}}_{ic}^{(j)} - \boldsymbol{\mu}^{(j+1)} \right) \left(\hat{\mathbf{x}}_{ic}^{(j)} - \boldsymbol{\mu}^{(j+1)} \right)' + \mathbf{C}_i^{(j)} \right]}{\sum_{i=1}^{n} (w_i^{(j)})^2 - 1}. \tag{20}$$

It can be shown that (20) corresponds to the estimating equation

$$\sum_{i=1}^{n} \frac{\partial \text{vec}'(\Sigma_i)}{\partial \boldsymbol{\sigma}} \mathbf{V}_i \text{vec}\{w_i^2[(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_i - \boldsymbol{\mu}_i)' - \Sigma_i]\} + \frac{\partial \text{vec}'(\Sigma)}{\partial \boldsymbol{\sigma}} \text{vec}(\Sigma^{-1}) = \mathbf{0}. \tag{21}$$

Because a weight is attached to $(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_i - \boldsymbol{\mu}_i)'$ in (16), (19) and (21), solutions to each of the three equations might be robust. However, these three equations are not as natural as (9) when considered as generalizations of (7) or equations satisfied by M- and S-estimators as well as any elliptical-distribution-based MLEs for samples without missing values (Kano et al. 1993; Lopuhaä 1989; Maronna 1976; Rocke 1996). Cheng and Victoria-Feser (2002) called (17) and (18) estimating equations. Clearly, (17) and (18) involve the imputed/estimated data $\hat{\mathbf{x}}_{im} = E(\mathbf{x}_{im}|\mathbf{x}_i, \boldsymbol{\mu}, \Sigma)$ whereas (8) and (9) do not. Equations (8) and (9) are not only consistent with the literature but also easily generalizable. When structural models $\boldsymbol{\mu}(\boldsymbol{\theta}_1)$ and $\Sigma(\boldsymbol{\theta}_2)$ are of interest and there is no overlapping parameters between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the corresponding estimating equations are obtained after replacing the $\boldsymbol{\mu}$ in the denominator of (8) and the $\boldsymbol{\sigma}$ in (9) by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. When $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ share common parameters and let $\boldsymbol{\theta}$ be the vector of all parameters, the corresponding estimating equation is obtained after replacing both

the $\boldsymbol{\mu}$ in the denominator of (8) and the $\boldsymbol{\sigma}$ in (9) by $\boldsymbol{\theta}$, and adding the two equations. It is not clear how to generalize (17) and (18) to structural models.

The identification of the estimating equations for each ER algorithm allows us to obtain a consistent estimate of the covariance matrix of the resulting robust means and dispersion matrix. Let $\mathbf{g}_i(\boldsymbol{\nu}) = (\mathbf{g}'_{i1}(\boldsymbol{\nu}), \mathbf{g}'_{i2}(\boldsymbol{\nu}))'$ with

$$\mathbf{g}_{i1}(\boldsymbol{\nu}) = w_{i1}(d_i)\frac{\partial \boldsymbol{\mu}'_i}{\partial \boldsymbol{\mu}}\Sigma_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i)$$

and

$$\mathbf{g}_{i2}(\boldsymbol{\nu}) = \frac{\partial \mathrm{vec}'(\Sigma_i)}{\partial \boldsymbol{\sigma}}\mathbf{V}_i \mathrm{vec}[w_{i2}(d_i)(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_i - \boldsymbol{\mu}_i)' - w_{i3}(d_i)\Sigma_i].$$

According to the theory of estimating equations (Godambe 1960; Huber 1967), under standard regularity conditions (Yuan and Jennrich 1998), the asymptotic covariance matrix of $\hat{\boldsymbol{\nu}} = (\hat{\boldsymbol{\mu}}', \hat{\boldsymbol{\sigma}}')'$ obtained at the convergence of (2), (3), (12) and (13) is consistently estimated by

$$\hat{\Gamma} = \left[\sum_{i=1}^{n}\frac{\partial \mathbf{g}_i(\hat{\boldsymbol{\nu}})}{\partial \hat{\boldsymbol{\nu}}'}\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{g}_i(\hat{\boldsymbol{\nu}})\mathbf{g}'_i(\hat{\boldsymbol{\nu}})\right]\left[\sum_{i=1}^{n}\frac{\partial \mathbf{g}'_i(\hat{\boldsymbol{\nu}})}{\partial \hat{\boldsymbol{\nu}}}\right]^{-1}, \tag{22}$$

where w_{i1}, w_{i2} and w_{i3} are also functions of $\boldsymbol{\nu}$ when evaluating the derivatives. Consistent covariance matrices can also be obtained for the estimators satisfying (16), (19) or (21) by properly defining $\mathbf{g}_i(\boldsymbol{\nu})$ such that these equations can be writen as $\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\nu}) = \mathbf{0}$.

Now we turn to starting values for the EM or ER algorithm. Since Eqs. (6) and (7) or (8) and (9) may have multiple solutions, robust starting values might be needed for the EM algorithm in (2) to (5) or the ER algorithm in (2), (3), (12) and (13) to yield estimates that are least affected by data contamination or outliers. The minimum covariance determinant (MCD) estimator has been suggested to use as the starting value for Σ because its breakdown point is close to 50 % (e.g., Cheng and Victoria-Feser 2002). Instead, we prefer the estimates proposed by Mehrotra (1995), because no iteration is needed in their calculation. In Mehrotra's proposal, each mean is estimated by the marginal median, each dispersion σ_{jj} is obtained by rescaling the median absolute deviation (MAD) from the median, and σ_{jk} is obtained by combining MAD and the median of all pairwise slopes in the form of $(x_{i_2 j} - x_{i_1 j})/(x_{i_2 k} - x_{i_1 k})$, $1 \le i_1 < i_2 \le n$, excluding cases with $x_{i_2 k} = x_{i_1 k}$. As an estimator of the slope of the regression of x_j on x_k, the median of the pairwise slopes was originally proposed by Theil (1950) and Sen (1968), and has been shown by Wilcox (1998) to enjoy not only good robust properties but also good small sample efficiency. Thus, we will use Mehrotra's proposal to get starting values for the EM and ER algorithms in the simulation study in the next section, where marginal median and pairwise slopes are applied to the observed data. In particular, when the starting $\Sigma^{(0)}$ is not positive definite, an eigenvalue decomposition on $\Sigma^{(0)}$ is performed, and a new $\Sigma^{(0)}$ is obtained by replacing all the eigenvalues smaller than .01 with .01 in the decomposition. Such a process was referred to as

"filtering" by Mehrotra (1995), which may not be needed unless n is small or missing data proportion is high, and together with a near singular population Σ.

3 Monte Carlo results

Although the main purpose of the paper is to establish the relationship between estimating equations and ER algorithm for estimating means and dispersion matrix with missing data, it is informative to see how different estimators perform when the underlying population varies. A Monte Carlo study is conducted for such a purpose. Seven estimators are compared in the study: NMLEs; t-distribution-based MLEs with degrees of freedom 3 and 1, respectively; M-estimators satisfying Eq. (8) and (9) with $w_{i3} = 1$, w_{i1} and w_{i2} being determined by the Huber-type weights in (10) with $\varphi = 0.2$ and 0.1, respectively; and M/S-estimators satisfying Eqs. (8) and (9) with w_{i1}, w_{i2} and w_{i3} determined by the biweight function in (11) with large sample breakdown points $6b_i/c_i^2 = 0.1$ and 0.2, respectively. They are denoted respectively by Nm, $t(3)$, $t(1)$, H(0.1), H(0.2), B(0.1), and B(0.2) in our presentation.

Let $\mathbf{1}_p$ be a vector of p 1s, and $\mathbf{I}_p$ be the identity matrix of size p. We chose $p = 5$ with population mean vector $\boldsymbol{\mu}_0 = \mathbf{1}_5$ and covariance matrix $\Sigma_0 = 0.5(\mathbf{I}_5 + \mathbf{1}_5\mathbf{1}_5')$, which is also a correlation matrix with all the correlations equal to 0.5.

Five distribution conditions are used to generate samples with missing data. Let $\mathbf{A}$ be the lower triangular matrix satisfying $\mathbf{A}\mathbf{A}' = \Sigma_0$. The five conditions are respectively (C1) the normal distribution according to $\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}_0$, where $\mathbf{z} \sim N_5(\mathbf{0}, \mathbf{I}_5)$; (C2) an elliptical distribution according to $\mathbf{x} = r\mathbf{A}\mathbf{z} + \boldsymbol{\mu}_0$, where r follows the standardized exponential distribution and is independent with $\mathbf{z}$; (C3) a skew distribution according to $\mathbf{x} = r\mathbf{A}\mathbf{u} + \boldsymbol{\mu}_0$, where r follows the same distribution as in C2, $\mathbf{u} = (u_1, u_2, \ldots, u_5)'$ and the u_js are independent with each other and with r, and each u_j follows the standardized gamma distribution with shape parameter 3; (C4) a contaminated normal distribution with 10 % of the sample from C1 being multiplied by 3; (C5) a contaminated normal distribution with 20 % of the sample from C1 being multiplied by 3. It is easy to see that $E(\mathbf{x}) = \boldsymbol{\mu}_0$ and $\mathrm{Cov}(\mathbf{x}) = \Sigma_0$ for the population in C1. It is also straightforward to show that the population mean vector and covariance matrix in C2 and C3 are also given by $\boldsymbol{\mu}_0$ and Σ_0, respectively. In C4 and C5, the majority of the cases correspond to $N_5(\boldsymbol{\mu}_0, \Sigma_0)$; and the observed samples are skewed in distribution. Notice that, with $p = 5$, the breakdown point of an M-estimator is limited by $(p+1)^{-1} = 0.167$. C4 and C5 are chosen to examine how the robust estimators perform when the percent of contamination is below and above the breakdown point. Clearly, only NML is asymptotically optimal for the normal distribution in C1, no other method is known to work best for any of the five conditions. Such a design is motivated by the fact that we typically do not know which method works best for a given data set whose population distribution is unknown.

Three sample sizes are used: $n = 100$, 300, 500. For each sample, x_1 and x_2 are fully observed; x_3, x_4 and x_5 are missing when $x_1 + x_2$ is greater than certain threshold values. Thus, there are two observed patterns for each sample and the missing values are MAR. The threshold values are chosen for (x_3, x_4, x_5) to miss at about 10, 20 and 30 %, respectively. Data contaminations in C4 and C5 are done after each sample

with missing values is obtained, and the percent of contamination is proportional to the number of cases in each observed pattern. For each combination of population distribution, sample size and missing data proportion, $N_r = 1000$ replications are used.

Notice that, under the assumption of an elliptical population distribution and without missing value, an M- or S-estimator $\hat{\Sigma}$ is known to converge to a matrix that is proportional to Σ_0 (Lopuhaä 1989; Maronna 1976), and the proportional factor depends on the underlying population and the weights used in the estimation process. It is not proper to compare the bias in different estimates of the dispersion matrix. In our evaluation of different estimators, we put all the estimators on the same scale by obtaining the corresponding correlation matrix following each estimate of the dispersion matrix. Let $\hat{\boldsymbol{\mu}}$ be the vector of the estimates of 5 means and $\hat{\boldsymbol{\rho}}$ be the vector of estimates of the 10 correlations at each sample by one of the 7 estimation methods. With $\hat{\boldsymbol{\gamma}}_i = (\hat{\boldsymbol{\mu}}_i', \hat{\boldsymbol{\rho}}_i')'$ for the ith replication, the bias, variance and mean square error (MSE) for the jth element of $\hat{\boldsymbol{\gamma}}$ are calculated as

$$\begin{aligned} \text{Bias}_j &= \bar{\gamma}_j - \gamma_{j0}, \\ \text{Var}_j &= \frac{1}{N_r - 1} \sum_{i=1}^{N_r} (\hat{\gamma}_{ij} - \bar{\gamma}_j)^2, \end{aligned}$$

and

$$\text{MSE}_j = \frac{1}{N_r} \sum_{i=1}^{N_r} (\hat{\gamma}_{ij} - \gamma_{j0})^2,$$

respectively, where $\bar{\gamma}_j = \sum_{i=1}^{N_r} \hat{\gamma}_{ij}/N_r$. Notice that our study includes 3 missing-data conditions, 3 sample-size conditions, 5 distribution conditions, and 7 estimation methods. With a total of $3 \times 3 \times 5 \times 7 = 315$ conditions and 15 parameter estimates, many tables are needed if we report the biases, variances and MSEs of the estimates for individual parameters. To save space, we choose to report the average of absolute bias, variance and MSE across the 15 parameters according to

$$\text{Bias} = \frac{1}{15} \sum_{j=1}^{15} |\text{Bias}_j|, \quad \text{Var} = \frac{1}{15} \sum_{j=1}^{15} \text{Var}_j, \quad \text{MSE} = \frac{1}{15} \sum_{j=1}^{15} \text{MSE}_j.$$

These are contained in 5 tables corresponding to the 5 distribution conditions. Because most of the quantities are in the 3rd decimal place, they are multiplied by 10 in the tables for us to see more details and to save space.

Table 1 contains the results of bias, variance and MSE of $\hat{\boldsymbol{\gamma}}$ when $\mathbf{x}$ is normally distributed (C1). For easy comparison, the smallest entry among the 7 estimation methods is underlined and the largest entry is put in bold. For each method, the average across the 9 conditions (3-sample-size by 3-missing-proportion) is also included on the right of the table. It is clear that NMLEs (those following Nm) enjoy the smallest bias, variance and MSE across all the conditions. The estimates following B(0.2) have the largest bias; whereas those following $t(1)$ have the largest variance and also the largest MSE on average. But estimates following B(0.2) have the largest MSE at $n = 300$ and

Table 1 Averages of empirical absolute bias × 10, variance × 10 and MSE × 10 of $\hat{\boldsymbol{\mu}}$ and $\hat{\rho}_{ij}$ by seven methods, (C1) normally distributed population

	$n = 100$			$n = 300$			$n = 500$			
Mis prop	10 %	20 %	30 %	10 %	20 %	30 %	10 %	20 %	30 %	Ave
Bias										
Nm	0.054	0.062	0.080	0.019	0.025	0.036	0.011	0.015	0.026	0.036
t(3)	0.182	0.228	0.238	0.159	0.197	0.196	0.153	0.190	0.191	0.192
t(1)	0.287	0.378	0.367	0.266	0.353	0.329	0.260	0.348	0.328	0.324
H(0.1)	0.078	0.079	0.094	0.050	0.048	0.056	0.043	0.039	0.048	0.059
H(0.2)	0.124	0.115	0.124	0.098	0.082	0.085	0.092	0.074	0.078	0.097
B(0.1)	0.344	0.408	0.372	0.325	0.386	0.345	0.320	0.381	0.341	0.358
B(0.2)	**0.416**	**0.512**	**0.480**	**0.396**	**0.489**	**0.450**	**0.391**	**0.484**	**0.446**	**0.452**
Var										
Nm	0.085	0.100	0.128	0.028	0.033	0.041	0.017	0.019	0.024	0.053
t(3)	0.098	0.116	0.147	0.031	0.037	0.047	0.019	0.022	0.027	0.060
t(1)	**0.110**	**0.132**	**0.171**	**0.035**	**0.041**	**0.054**	**0.021**	**0.025**	**0.031**	**0.069**
H(0.1)	0.087	0.102	0.130	0.028	0.033	0.042	0.017	0.020	0.025	0.054
H(0.2)	0.089	0.104	0.133	0.029	0.034	0.043	0.017	0.020	0.025	0.055
B(0.1)	0.086	0.102	0.130	0.028	0.033	0.042	0.017	0.020	0.025	0.054
B(0.2)	0.087	0.104	0.133	0.028	0.034	0.043	0.017	0.020	0.025	0.055
MSE										
Nm	0.086	0.101	0.129	0.028	0.033	0.041	0.017	0.019	0.024	0.053
t(3)	0.101	0.121	0.153	0.034	0.041	0.051	0.021	0.026	0.031	0.064
t(1)	**0.118**	**0.146**	**0.185**	0.042	0.054	0.065	0.028	0.037	0.042	**.080**
H(0.1)	0.088	0.103	0.131	0.029	0.033	0.042	0.017	0.020	0.025	0.054
H(0.2)	0.091	0.106	0.134	0.030	0.034	0.043	0.018	0.021	0.026	0.056
B(0.1)	0.102	0.124	0.148	0.042	0.053	0.057	0.030	0.039	0.039	0.071
B(0.2)	0.108	0.135	0.159	**0.047**	**0.061**	**0.066**	**0.035**	**0.047**	**0.047**	0.078

Each underlined number is the smallest entry (bias, variance or MSE) among the 7 estimation methods, whereas the largest is in bold

500. Missing data proportion has little effect on the performance of the 7 estimation methods.

Table 2 contains the results of bias, variance and MSE of $\hat{\boldsymbol{\gamma}}$ when $\mathbf{x}$ is elliptically distributed (C2). While estimates following B(0.2) continue to have the largest bias in 8 out of the 9 conditions, the method yielding estimates with the smallest bias varies across conditions of sample size and missing data proportion. Estimates following *Nm* have the largest variance and MSE on average, whereas estimates with the smallest variance and MSE are given by ML based on $\mathbf{x} \sim Mt_5(\boldsymbol{\mu}, \Sigma, 1)$.

The results under a skew population distribution (C3) are in Table 3. Like in Table 2, estimates following *Nm* have the largest variance and MSE across the 9 conditions. But estimates following *Nm* have the smallest bias at $n = 100$; and B(0.2) enjoys the smallest bias at $n = 300$ and 500. Estimates following $t(1)$ have the smallest variances and also smallest MSE in 8 out of the 9 conditions. Missing data proportion

Table 2 Averages of empirical absolute bias × 10, variance ×10 and MSE ×10 of $\hat{\boldsymbol{\mu}}$ and $\hat{\rho}_{ij}$ by seven methods, (C2) elliptically distributed population

	$n = 100$			$n = 300$			$n = 500$			
Mis prop	10 %	20 %	30 %	10 %	20 %	30 %	10 %	20 %	30 %	Ave
Bias										
Nm	0.064	0.054	0.060	0.155	0.153	0.116	0.192	0.195	0.159	0.128
t(3)	0.021	0.108	0.170	0.047	0.153	0.217	0.053	0.161	**0.232**	0.129
t(1)	0.098	0.057	0.055	0.074	0.027	0.083	0.068	0.017	0.096	0.064
H(0.1)	0.025	0.093	0.121	0.059	0.160	0.192	0.067	0.169	0.206	0.121
H(0.2)	0.073	0.061	0.103	0.037	0.111	0.169	0.032	0.123	0.185	0.099
B(0.1)	0.374	0.357	0.280	0.308	0.271	0.175	0.297	0.257	0.155	0.275
B(0.2)	**0.402**	**0.413**	**0.308**	**0.353**	**0.352**	**0.234**	**0.350**	**0.347**	0.226	**0.331**
Var										
Nm	**0.292**	**0.346**	**0.398**	**0.121**	**0.143**	**0.165**	**0.078**	**0.094**	**0.109**	**0.194**
t(3)	0.074	0.086	0.098	0.023	0.026	0.030	0.014	0.016	0.018	0.043
t(1)	0.068	0.079	0.089	0.021	0.025	0.028	0.013	0.015	0.017	0.040
H(0.1)	0.094	0.110	0.131	0.030	0.034	0.039	0.018	0.021	0.024	0.056
H(0.2)	0.089	0.105	0.125	0.028	0.032	0.037	0.018	0.020	0.023	0.053
B(0.1)	0.129	0.158	0.192	0.041	0.049	0.058	0.024	0.029	0.033	0.079
B(0.2)	0.102	0.123	0.148	0.033	0.039	0.045	0.020	0.024	0.027	0.062
MSE										
Nm	**0.292**	**0.346**	**0.398**	**0.124**	**0.147**	**0.167**	**0.084**	**0.100**	**0.113**	**0.197**
t(3)	0.074	0.087	0.102	0.023	0.029	0.035	0.015	0.019	0.025	0.045
t(1)	0.070	0.080	0.090	0.022	0.025	0.029	0.014	0.015	0.018	0.040
H(0.1)	0.094	0.111	0.133	0.030	0.037	0.044	0.019	0.025	0.030	0.058
H(0.2)	0.090	0.106	0.126	0.028	0.034	0.041	0.018	0.022	0.028	0.055
B(0.1)	0.147	0.174	0.201	0.054	0.059	0.062	0.036	0.038	0.037	0.090
B(0.2)	0.123	0.145	0.159	0.049	0.055	0.052	0.036	0.039	0.033	0.077

Each underlined number is the smallest entry (bias, variance or MSE) among the 7 estimation methods, whereas the largest is in bold

has little effect on the performance of different estimation methods. Notice that the robust estimators may not be consistent when the population distribution is skewed. However, the average MSEs in Table 3 following all the robust methods are smaller than those following *Nm*. Biases following B(0.2) at $n = 300$ and 500 are also smaller than those following *Nm*.

Table 4 contains the results when 10 % of the sample from a normally distributed population is contaminated (C4). The estimates following *Nm* are most biased, least efficient and consequently have the largest MSEs. Least biased estimates are given by *t*(1) or *t*(3), whereas most efficient estimates are given by *t*(3) or H(0.2), depending on missing data proportion. Estimates with least MSEs are given by *t*(3) in 8 out of the 9 conditions, and *t*(1) enjoys the smallest MSE at $n = 500$ and 10 % of missing data.

Table 5 contains the results when 20 % of the sample from a normally distributed population is contaminated (C5). Again, the NMLEs (those following *Nm*) are most

Table 3 Averages of empirical absolute bias × 10, variance ×10 and MSE ×10 of $\hat{\boldsymbol{\mu}}$ and $\hat{\rho}_{ij}$ by seven methods, (C3) skew distributed population

	$n = 100$			$n = 300$			$n = 500$			
Mis prop	10 %	20 %	30 %	10 %	20 %	30 %	10 %	20 %	30 %	Ave
Bias										
Nm	0.063	0.050	0.045	0.135	0.125	0.085	0.168	0.160	0.121	0.106
t(3)	**0.282**	**0.383**	**0.429**	**0.299**	**0.400**	**0.461**	**0.309**	**0.414**	**0.474**	**0.383**
t(1)	0.192	0.254	0.339	0.204	0.264	0.363	0.213	0.277	0.375	0.276
H(0.1)	0.165	0.254	0.270	0.193	0.279	0.312	0.204	0.294	0.327	0.255
H(0.2)	0.156	0.282	0.310	0.181	0.308	0.355	0.191	0.324	0.372	0.275
B(0.1)	0.190	0.142	0.091	0.167	0.123	0.119	0.150	0.114	0.134	0.137
B(0.2)	0.144	0.121	0.079	0.135	0.116	0.084	0.125	0.109	0.095	0.112
Var										
Nm	**0.367**	**0.424**	**0.476**	**0.160**	**0.188**	**0.214**	**0.106**	**0.124**	**0.143**	**0.245**
t(3)	0.068	0.077	0.087	0.023	0.026	0.029	0.014	0.015	0.017	0.039
t(1)	0.063	0.072	0.081	0.022	0.024	0.027	0.012	0.014	0.016	0.037
H(0.1)	0.087	0.098	0.115	0.029	0.033	0.037	0.017	0.019	0.022	0.051
H(0.2)	0.083	0.093	0.110	0.028	0.032	0.036	0.017	0.018	0.021	0.049
B(0.1)	0.132	0.155	0.185	0.040	0.047	0.053	0.024	0.028	0.032	0.077
B(0.2)	0.098	0.114	0.135	0.033	0.039	0.043	0.019	0.022	0.025	0.059
MSE										
Nm	**0.367**	**0.424**	**0.476**	**0.163**	**0.190**	**0.215**	**0.110**	**0.128**	**0.145**	**0.246**
t(3)	0.076	0.092	0.106	0.033	0.042	0.051	0.023	0.033	0.040	0.055
t(1)	0.067	0.079	0.093	0.026	0.032	0.041	0.017	0.022	0.031	0.045
H(0.1)	0.091	0.106	0.124	0.034	0.043	0.050	0.023	0.030	0.035	0.060
H(0.2)	0.086	0.102	0.121	0.032	0.043	0.051	0.021	0.031	0.037	0.058
B(0.1)	0.137	0.158	0.186	0.044	0.049	0.055	0.027	0.031	0.035	0.080
B(0.2)	0.101	0.117	0.137	0.036	0.042	0.045	0.022	0.026	0.027	0.062

Each underlined number is the smallest entry (bias, variance or MSE) among the 7 estimation methods, whereas the largest is in bold

biased, least efficient and consequently have the largest MSEs. The estimates following *t*(1) are least biased and have the smallest MSEs, whereas estimates following *t*(3) have the least variances. Missing data proportion or sample size has little effect on the performance of the different methods.

Comparing the averaged numbers (the last column) in each table and across the 5 tables, we may notice that NMLEs differ from each of the robust estimates substantially in both bias and variance. The robust estimates differ more in bias than in variance. Except those following *t*(1), all the other estimates attain the largest averaged bias and MSE at C5, implying that the contaminated distribution is their least favorite among the 5 distribution conditions. Although the population distribution in C3 is skewed, the sizes of the average MSEs in Table 3 following all the robust methods are comparable to those in Tables 1 and 2. The average biases corresponding to B(0.1) and B(0.2) in Table 3 are even smaller than those in Tables 1 and 2.

Table 4 Averages of empirical absolute bias × 10, variance ×10 and MSE ×10 of $\hat{\boldsymbol{\mu}}$ and $\hat{\rho}_{ij}$ by seven methods, (C4) 10 % of normally distributed samples are contaminated

	$n = 100$			$n = 300$			$n = 500$			
Mis prop	10 %	20 %	30 %	10 %	20 %	30 %	10 %	20 %	30 %	Ave
Bias										
Nm	**1.252**	**1.281**	**1.245**	**1.272**	**1.303**	**1.312**	**1.273**	**1.308**	**1.319**	**1.285**
$t(3)$	0.134	0.103	0.104	0.156	0.126	0.134	0.154	0.123	0.130	0.129
$t(1)$	0.076	0.138	0.161	0.049	0.109	0.120	0.047	0.105	0.115	0.102
H(0.1)	0.501	0.521	0.538	0.520	0.540	0.569	0.513	0.533	0.563	0.533
H(0.2)	0.401	0.422	0.438	0.423	0.445	0.472	0.418	0.440	0.467	0.436
B(0.1)	0.497	0.442	0.485	0.506	0.414	0.442	0.496	0.411	0.436	.459
B(0.2)	0.226	0.303	0.340	0.197	0.265	0.284	0.192	0.263	0.273	0.260
Var										
Nm	**0.175**	**0.224**	**0.301**	**0.056**	**0.068**	**0.093**	**0.033**	**0.043**	**0.059**	**0.117**
$t(3)$	0.101	0.119	0.152	0.032	0.037	0.048	0.019	0.022	0.028	0.062
$t(1)$	0.112	0.132	0.169	0.035	0.041	0.053	0.021	0.025	0.032	0.069
H(0.1)	0.103	0.125	0.167	0.032	0.039	0.051	0.019	0.023	0.031	0.066
H(0.2)	0.100	0.120	0.157	0.031	0.038	0.049	0.019	0.023	0.030	0.063
B(0.1)	0.141	0.175	0.229	0.043	0.052	0.068	0.026	0.032	0.043	0.090
B(0.2)	0.117	0.141	0.183	0.036	0.043	0.055	0.021	0.026	0.033	0.073
MSE										
Nm	**0.371**	**0.435**	**0.514**	**0.253**	**0.278**	**0.312**	**0.230**	**0.254**	**0.279**	**0.325**
$t(3)$	0.104	0.121	0.154	0.034	0.039	0.050	0.022	0.024	0.031	0.064
$t(1)$	0.112	0.135	0.172	0.035	0.043	0.055	0.021	0.026	0.033	0.070
H(0.1)	0.140	0.166	0.213	0.069	0.079	0.096	0.055	0.063	0.075	0.106
H(0.2)	0.126	0.148	0.189	0.057	0.066	0.080	0.044	0.050	0.060	0.091
B(0.1)	0.188	0.218	0.277	0.087	0.091	0.109	0.068	0.070	0.083	0.132
B(0.2)	0.125	0.152	0.197	0.042	0.052	0.066	0.028	0.035	0.043	0.082

Each underlined number is the smallest entry (bias, variance or MSE) among the 7 estimation methods, whereas the largest is in bold

In summary, NML is most preferable when $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$, but it can perform badly when data are nonnormally distributed or contaminated. Each robust method also has its pros and cons, depending on the underlying population distribution of the sample. With 20 % of the sample from a normally distributed population being contaminated, we may expect B(0.2) to perform better than the results presented in Table 5. The under-expectation of B(0.2) might be understood from the definition of the breakdown point, which is the proportion of extreme observations an estimator can take before becoming arbitrarily large or small, not related to optimizing bias, variance or MSE. The reason for not observing the advantage of B(0.2) in Table 5 might not be because the contaminated observations are not extreme enough, since we also studied the condition of multiplying 20 % of normally distributed samples from C1 by 5 and found results similar to those in Table 5.

Table 5 Averages of empirical absolute bias × 10, variance ×10 and MSE ×10 of $\hat{\boldsymbol{\mu}}$ and $\hat{\rho}_{ij}$ by seven methods, (C5) 20 % of normally distributed samples are contaminated

	$n = 100$			$n = 300$			$n = 500$			
Mis prop	10 %	20 %	30 %	10 %	20 %	30 %	10 %	20 %	30 %	Ave
Bias										
Nm	**2.057**	**2.108**	**2.112**	**2.069**	**2.114**	**2.129**	**2.061**	**2.123**	**2.154**	**2.103**
$t(3)$	0.478	0.455	0.462	0.493	0.472	0.487	0.489	0.467	0.484	0.476
$t(1)$	0.230	0.148	0.149	0.251	0.170	0.175	0.252	0.169	0.171	0.191
H(0.1)	1.173	1.194	1.235	1.171	1.201	1.253	1.157	1.190	1.251	1.203
H(0.2)	1.001	1.015	1.046	1.005	1.026	1.065	.995	1.018	1.064	1.026
B(0.1)	1.262	1.114	1.059	1.275	1.121	1.071	1.251	1.108	1.068	1.148
B(0.2)	0.802	0.620	0.597	0.797	0.624	0.567	0.779	0.607	0.556	0.661
Var										
Nm	**0.186**	**0.231**	**0.316**	**0.059**	**0.072**	**0.098**	**0.036**	**0.044**	**0.060**	**0.122**
$t(3)$	0.103	0.122	0.155	0.033	0.038	0.049	0.020	0.023	0.029	0.064
$t(1)$	0.113	0.133	0.168	0.036	0.042	0.053	0.022	0.025	0.032	0.069
H(0.1)	0.120	0.150	0.200	0.038	0.046	0.060	0.023	0.028	0.037	0.078
H(0.2)	0.111	0.137	0.180	0.035	0.042	0.055	0.022	0.026	0.034	0.071
B(0.1)	0.160	0.196	0.255	0.050	0.061	0.079	0.031	0.037	0.049	0.102
B(0.2)	0.139	0.170	0.218	0.043	0.051	0.066	0.026	0.031	0.041	0.087
MSE										
Nm	**0.836**	**0.922**	**1.037**	**0.705**	**0.754**	**0.804**	**0.675**	**0.726**	**0.773**	**0.804**
$t(3)$	0.129	0.147	0.182	0.059	0.063	0.077	0.046	0.048	0.057	0.090
$t(1)$	0.120	0.137	0.173	0.043	0.046	0.059	0.029	0.029	0.037	0.075
H(0.1)	0.325	0.370	0.444	0.236	0.258	0.296	0.217	0.235	0.269	0.294
H(0.2)	0.260	0.295	0.354	0.181	0.196	0.225	0.165	0.177	0.201	0.228
B(0.1)	0.452	0.466	0.529	0.336	0.323	0.341	0.308	0.293	0.307	0.373
B(0.2)	0.247	0.254	0.297	0.144	0.129	0.139	0.124	0.105	0.111	.172

Each underlined number is the smallest entry (bias, variance or MSE) among the 7 estimation methods, whereas the largest is in bold

For contaminated data in C5, it is very likely that B(α) with a greater α than 0.2 may work better than B(0.2). Similarly, other degrees of freedom corresponding to the multivariate t-distribution and other tuning parameters in the Huber-type weights may yield more efficient and less biased estimates. Our results are consistent with what Richardson and Welsh (1995) have found in the context of mixed linear models, where no method performs the best across all the conditions.

4 Applications

As mentioned in the introduction, means and covariance/dispersion matrix are behind many commonly used statistical methods. In this section, we discuss applications of

robust means and dispersion matrix in linear regression and growth curve models, both are widely used in various disciplines.

4.1 Regression models

Consider the regression model

$$y_i = \alpha + \mathbf{u}_i'\boldsymbol{\beta} + e_i, \quad i = 1, 2, \ldots, n. \tag{23}$$

When all the n observations are completely observed, let s_{yy}, $\mathbf{s}_{uy}$, $\mathbf{S}_{uu}$ be the sample variance of y_i, vector of covariances of $\mathbf{u}_i$ with y_i, and covariance matrix of $\mathbf{u}_i$, respectively. Then

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{uu}^{-1}\mathbf{s}_{uy} \tag{24}$$

is the MLE of $\boldsymbol{\beta}$ under the assumption $e_i \sim N(0, \sigma^2)$ and that $\mathbf{u}_i$ and e_i are independent. Without missing values and when the $\mathbf{u}_i$s are not subject to data contamination, a robust estimate of $\boldsymbol{\beta}$ can be defined using estimating equations by assigning smaller weights to cases with larger residuals $e_i = y_i - \alpha - \mathbf{u}_i'\boldsymbol{\beta}$ (e.g., Hampel et al. 1986, pp. 311–312). However, with real data, both y_i and $\mathbf{u}_i$ may contain missing values. If y_i is missing in (23), then e_i is not available even when α and $\boldsymbol{\beta}$ are known. If certain elements of $\mathbf{u}_i$ are missing and y_i is observed, then the meaning of e_i based on observed $\mathbf{u}_i$ is different from that in (23). Thus, it is not clear how to generalize robust regression from complete data to missing data by downweighting large residuals.

Notice that a robust estimate of $\boldsymbol{\beta}$ parallel to (24) can still be obtained as long as robust estimates of $\Sigma_{uu} = \mathrm{Cov}(\mathbf{u}_i)$ and $\boldsymbol{\sigma}_{uy} = \mathrm{Cov}(\mathbf{u}_i, y_i)$ or the corresponding dispersion matrices are available. With missing values, Little (1988) parameterizes $\boldsymbol{\sigma}_{uy} = \Sigma_{uu}\boldsymbol{\beta}$ in formulating the EM algorithms for robust regression based on multivariate t-distributions. Such a parameterization is mathematically equivalent to letting $\boldsymbol{\sigma}_{uy}$ and Σ_{uu} be free parameters. Similarly, with $\mathbf{x}_i = (y_i, \mathbf{u}_i')'$, the $\boldsymbol{\mu}_i$ and Σ_i in (8) and (9) can also be reparameterized using α, $\boldsymbol{\mu}_u = E(\mathbf{u}_i)$, $\boldsymbol{\beta}$, Σ_{uu}, and $\sigma^2 = \mathrm{Var}(e_i)$. Since the ER algorithm corresponding to (8) and (9) is easier to program, there is no foreseeable advantage of using the regression parameterization. In particular, at the convergence of the ER algorithm, we obtain robust estimates of α and $\boldsymbol{\beta}$ as

$$\hat{\alpha} = \hat{\mu}_y - \hat{\boldsymbol{\mu}}_u'\hat{\boldsymbol{\beta}}, \quad \text{and} \quad \hat{\boldsymbol{\beta}} = \hat{\Sigma}_{uu}^{-1}\hat{\boldsymbol{\sigma}}_{uy}.$$

Since $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are functions of $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$, they will be consistent as long as $\hat{\boldsymbol{\mu}}$ is consistent and $\hat{\Sigma}$ converges to $\kappa\Sigma$ for certain $\kappa > 0$, as $n \to \infty$. With missing values that are MAR, although it is not clear under what conditions robust estimators are consistent beyond those noted in Sect. 2, the Monte Carlo results in Sect. 3 imply that robust estimators can perform a lot better than NMLEs for many conditions. With the $\hat{\Gamma}$ in (22), consistent SEs of $\hat{\boldsymbol{\beta}}$ can be obtained from the so-called delta-method. Notice that $\boldsymbol{\sigma} = \mathrm{vech}(\Sigma) = (\sigma_{yy}, \boldsymbol{\sigma}_{uy}', \mathrm{vech}'(\Sigma_{uu}))'$, there exist

$$\dot{\boldsymbol{\beta}}_1(\Sigma) = \partial\boldsymbol{\beta}(\Sigma)/\partial\boldsymbol{\mu}' = \mathbf{0}, \quad \dot{\boldsymbol{\beta}}_2(\Sigma) = \partial\boldsymbol{\beta}(\Sigma)/\partial\boldsymbol{\sigma}' \\ = (\mathbf{0}, \Sigma_{uu}^{-1}, -(\boldsymbol{\beta}' \otimes \Sigma_{uu}^{-1})\mathbf{D}_{p-1}), \tag{25}$$

where $\mathbf{D}_{p-1} = \partial\text{vec}(\Sigma_{uu})/\partial\text{vech}'(\Sigma_{uu})$ is the $(p-1)^2 \times [p(p-1)/2]$ duplication matrix (see e.g., Schott 2005, p. 313). It follows from the delta-method that the covariance matrix of $\hat{\boldsymbol{\beta}}$ is consistently estimated by

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \dot{\boldsymbol{\beta}}(\hat{\Sigma})\hat{\Gamma}\,\dot{\boldsymbol{\beta}}'(\hat{\Sigma}) \quad \text{or} \quad \text{Cov}(\hat{\boldsymbol{\beta}}) = \frac{n}{n-p}\dot{\boldsymbol{\beta}}(\hat{\Sigma})\hat{\Gamma}\dot{\boldsymbol{\beta}}'(\hat{\Sigma}), \tag{26}$$

where $\dot{\boldsymbol{\beta}}(\hat{\Sigma}) = (\dot{\boldsymbol{\beta}}_1(\hat{\Sigma}), \dot{\boldsymbol{\beta}}_2(\hat{\Sigma}))$.

When dummy coded categorical variables such as experimental condition, gender or race are present and are completely observed, robust means and dispersion matrix can be estimated for each group. After $\hat{\boldsymbol{\mu}}$s and $\hat{\Sigma}$s are obtained for all the groups, regression analysis can be done for each group separately when there is no constraint on parameters across the groups. With constraints on parameters across the groups, regression analysis can be done by fitting the regression models simultaneously to the $\hat{\boldsymbol{\mu}}$s and $\hat{\Sigma}$s under the constraints. When the $\mathbf{u}_i$s in (23) contain categorical variables that are missing, the robust methods described here may not be appropriate. Maximum likelihood estimates of the regression coefficients can be obtained if one can correctly specify the distribution of the categorical variables and the conditional distribution of the continuous variables given the categorical variables (see e.g., Little and Schluchter 1985). The resulting estimators will enjoy certain robust properties if the specified distribution accounts for heavy tails in the observed data. More studies in this direction are needed.

4.2 Growth curve models

Let y_{it} be the observed outcome of person i at time t, $t = 1, 2, \ldots, T$; $i = 1, 2, \ldots, n$; $\mathbf{u}_i$ be a vector that contains background variables (e.g., treatment conditions) for person i. For complete data, let the linear growth curve model be

$$y_{it} = \beta_{i0} + \beta_{i1}t + \varepsilon_{it}, \quad \beta_{i0} = \boldsymbol{\gamma}_0'\mathbf{u}_i + \delta_{i0}, \quad \beta_{i1} = \boldsymbol{\gamma}_1'\mathbf{u}_i + \delta_{i1}, \tag{27}$$

where $E(\varepsilon_{it}) = 0$, $\text{Var}(\varepsilon_{it}) = \psi_{tt}$, $\text{Cov}(\varepsilon_{is}, \varepsilon_{it}) = \psi_{st} = 0$ when $s \neq t$; $E(\delta_{i0}) = E(\delta_{i1}) = 0$, $\text{Var}(\delta_{i0}) = \phi_{00}$, $\text{Var}(\delta_{i1}) = \phi_{11}$, $\text{Cov}(\delta_{i0}, \delta_{i1}) = \phi_{01}$; and $\mathbf{u}_i$, ε_{it} and $(\delta_{i0}, \delta_{i1})$ are independent. Then the structured means and covariances of $\mathbf{x}_i = (\mathbf{y}_i', \mathbf{u}_i')'$, with $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, are

$$E(y_{it}) = \boldsymbol{\gamma}_0'\boldsymbol{\mu}_u + \boldsymbol{\gamma}_1'\boldsymbol{\mu}_u t, \quad E(\mathbf{u}_i) = \boldsymbol{\mu}_u, \\ \text{Cov}(y_{is}, y_{it}) = (\boldsymbol{\gamma}_0 + s\boldsymbol{\gamma}_1)'\Sigma_{uu}(\boldsymbol{\gamma}_0 + t\boldsymbol{\gamma}_1) + \phi_{00} + (s+t)\phi_{01} + st\phi_{11} + \psi_{st}, \\ \text{Cov}(\mathbf{u}_i) = \Sigma_{uu}, \quad \text{Cov}(y_{it}, \mathbf{u}_i) = (\boldsymbol{\gamma}_0 + t\boldsymbol{\gamma}_1)'\Sigma_{uu};$$

and the vector of model parameters is

$$\boldsymbol{\theta} = (\boldsymbol{\gamma}_0', \boldsymbol{\gamma}_1', \phi_{00}, \phi_{01}, \phi_{11}, \psi_{11}, \psi_{22}, \ldots, \psi_{TT}, \boldsymbol{\mu}_u', \text{vech}'(\Sigma_{uu}))'.$$

With missing values, let $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\Sigma(\boldsymbol{\theta})$ represent the mean and covariance structural models given above, and $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ be proper robust estimates that satisfy (8) and (9). Robust estimates of $\boldsymbol{\theta}$ can be obtained by minimizing a discrepancy function between $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ and $(\boldsymbol{\mu}(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$. A commonly used discrepancy function is derived from the likelihood ratio statistic of testing $(\boldsymbol{\mu}(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$ nested within $(\boldsymbol{\mu}, \Sigma)$ by assuming $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ as the sample means and covariance matrix based on $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$. Consistent SEs of the resulting $\hat{\boldsymbol{\theta}}$ can be obtained using a sandwich-type covariance matrix involving the $\hat{\Gamma}$ in (22). Details of fitting mean and covariance structural models in general are given in Yuan and Zhang (2012), which contains multiple test statistics for overall model evaluation.

With complete data, we can define a robust M-estimator of $\boldsymbol{\theta}$ for (27) by estimating equations in which cases with large ε_{it} and/or δ_{ij} are downweighted. Estimating equations for the structural model can also be formulated when y_{it}s are partially observed and the $\mathbf{u}_i$s contain no missing values. When the $\mathbf{u}_i$s contain missing values, however, it is not clear how to define estimating equations by downweighting cases with large ε_{it} or δ_{ij}. The two-stage approach by obtaining $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ first and then fitting $(\boldsymbol{\mu}(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$ to $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ provides a robust procedure for growth curve modeling with missing data. In mean and covariance structure analysis with missing values, it has been shown that a two-stage approach by estimating the saturated means and covariances using NML first and then fitting them by the structural models works better than direct NML (Savalei and Falk 2014). We expect that fitting $(\boldsymbol{\mu}(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$ to $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ will work equally well, if not better, than robust estimators of $\boldsymbol{\theta}$ for (27) directly defined through estimating equations.

Similar to regression, when dummy coded variables such as group membership or treatment conditions exist, means and dispersion matrix can be robustly estimated for each group. Growth curve modeling can be done separately for each group or simultaneously when across-group constraints exist. When the $\mathbf{u}_i$s contain categorical variables that are missing, method based on mixed distribution of continuous and categorical variables is needed to properly model the joint distribution of the observed $\mathbf{u}_i$ and $\mathbf{y}_i$. Robustness of the method depends on the extent the mixed distribution can account for heavy tails in the observed $\mathbf{x}_i = (\mathbf{y}_i', \mathbf{u}_i')'$. More development in this direction is needed.

5 Discussions

Most classical estimation methods generate consistent parameter estimates when data are complete. With missing data that are MAR, only MLEs are known to be consistent in general. However, with real data, it is hard to specify a correct likelihood function to generate true MLEs. When data have heavy tails, by adjusting the degrees of freedom (Liu 1997), a t-distribution might better describe the underlying population than the normal distribution. Estimating equations provide even more flexibility in modeling the distribution of the data, and they have become important tools in many areas when modeling practical data whose population distributions are unknown or cannot be described by a familiar parametric family (e.g., Godambe 1991; Liang and Zeger 1986; Prentice

and Zhao 1991). The flexibility with estimating Eqs. (8) and (9) lies in that w_{i1}, w_{i2}, and w_{i3} can have different forms. By properly choosing these weights, the estimating equations may closely approximate those yielded by setting the true but unknown score functions at zero. Then, the resulting estimates will be close to being consistent and asymptotically most efficient. Since it is unlikely to know the size of the bias in parameter estimates with real data, we may select the weights according to the size of the variances corresponding to a set of invariant model parameters (see Yuan et al. 2004), with the hope that the estimates also have minimal biases when their variances are close to smallest. The variances can be estimated using the asymptotic covariance matrix in (22) or the bootstrap (Efron and Tibshirani 1993). Of course, to identify nearly optimal weights for a given data set, one needs to include a variety of procedures. In particular, when a high percentage of data contamination is suspected, S-estimator or other high-breakdown-point estimators need to be included in the comparison. For most of the conditions in Tables 1, 2, 3, 4 and 5, the methods achieve the smallest variances also generate either the smallest biases or the smallest MSEs. For the few exceptions that the smallest variances and MSEs or biases do not go with the same method, the biases or MSEs corresponding to the smallest variances are close to being the smallest.

When data contamination or outliers are suspected, an alternative procedure is to use NML following outlier removal with influential analysis (Poon and Poon 2002). Such a procedure may generate more efficient estimates if the population is normally distributed without contamination. If the heavy tails in a sample are not just due to outliers, a robust method might perform better.

Statistical theory for robust estimation with complete data is primarily developed under the assumption of symmetric or elliptical distributions, mainly because the resulting parameter estimates are consistent. However, with missing data that are MAR, it is not clear whether the consistency property still holds when the population distribution is elliptical but the left sides of (8) and (9) are not the score functions derived from the elliptical-distribution-based log likelihood function. Even if the consistency property can be established within the class of elliptical distributions, it is not clear how to use such a result in practice. This is because, even when the population is elliptically distributed, the observed data can be skewed under MAR mechanism, and there does not exist an effective procedure to tell the difference between skewness caused by the MAR mechanism and that caused by a skew underlying population distribution. When it is not clear which method to choose for a given data set, many applied researchers just go with NML, and robust methods offer viable alternatives to NML. Monte Carlo results in Sect. 3 of this article showed that robust estimates are more accurate than the NMLEs when the population distribution is either of heavy tails or data are contaminated. References cited in this article and elsewhere have repeatedly shown the advantage of robust methods over NML with real data.

Appendix A

This appendix shows that the converged values of the ER algorithm in (2), (3), (12) and (13) satisfy (8) and (9). For simple notation we use $\boldsymbol{\mu}$ and Σ to denote the converged values and rewrite (12) and (13) as

$$\sum_{i=1}^{n} w_{i1}(\hat{\mathbf{x}}_{ic} - \boldsymbol{\mu}) = \mathbf{0} \tag{28}$$

and

$$\sum_{i=1}^{n} \left[w_{i2}(\hat{\mathbf{x}}_{ci} - \boldsymbol{\mu})(\hat{\mathbf{x}}_{ic} - \boldsymbol{\mu})' + w_{i3}(\mathbf{C}_i - \Sigma) \right] = \mathbf{0}, \tag{29}$$

where

$$\mathbf{C}_i = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{imm} \end{pmatrix}$$

with $\mathbf{C}_{imm} = \Sigma_{imm} - \Sigma_{imo}\Sigma_i^{-1}\Sigma_{iom}$ being the converged $\mathbf{C}_{imm}^{(j)}$ in (3). Notice that

$$\hat{\mathbf{x}}_{ic} - \boldsymbol{\mu} = \begin{pmatrix} \mathbf{x}_i - \boldsymbol{\mu}_i \\ \Sigma_{imo}\Sigma_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i) \end{pmatrix}$$

and

$$\Sigma^{-1} = \begin{pmatrix} \mathbf{I}_{p_i} & -\Sigma_i^{-1}\Sigma_{iom} \\ \mathbf{0} & \mathbf{I}_{q_i} \end{pmatrix} \begin{pmatrix} \Sigma_i^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{i(m|o)}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{p_i} & \mathbf{0} \\ -\Sigma_{imo}\Sigma_i^{-1} & \mathbf{I}_{q_i} \end{pmatrix},$$

where $q_i = p - p_i$ and $\Sigma_{i(m|o)} = \mathbf{C}_{imm}$. Direct matrix multiplication yields

$$\Sigma^{-1}(\hat{\mathbf{x}}_{ic} - \boldsymbol{\mu}) = \begin{pmatrix} \Sigma_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i) \\ \mathbf{0} \end{pmatrix}. \tag{30}$$

The equivalence of (8) and (28) follows from (30) and $\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\mu}' = (\mathbf{I}_{p_i}, \mathbf{0})$.

For showing equivalence of (9) and (29), let $\mathbf{H}_i = (\hat{\mathbf{x}}_{ic} - \boldsymbol{\mu})(\hat{\mathbf{x}}_{ic} - \boldsymbol{\mu})'$. When $p_i < p$,

$$\mathbf{H}_i = \begin{pmatrix} \mathbf{H}_{ioo} & \mathbf{H}_{ioo}\Sigma_i^{-1}\Sigma_{iom} \\ \Sigma_{imo}\Sigma_i^{-1}\mathbf{H}_{ioo} & \Sigma_{imo}\Sigma_i^{-1}\mathbf{H}_{ioo}\Sigma_i^{-1}\Sigma_{iom} \end{pmatrix},$$

where $\mathbf{H}_{ioo} = (\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_i - \boldsymbol{\mu}_i)'$. Matrix multiplications yield

$$\Sigma^{-1}\mathbf{H}_i\Sigma^{-1} = \begin{pmatrix} \Sigma_i^{-1}\mathbf{H}_{ioo}\Sigma_i^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{31}$$

and

$$\Sigma^{-1}\mathbf{C}_i\Sigma^{-1} = \begin{pmatrix} \Sigma_i^{-1}\Sigma_{iom}\Sigma_{i(m|o)}^{-1}\Sigma_{imo}\Sigma_i^{-1} & -\Sigma_i^{-1}\Sigma_{iom}\Sigma_{i(m|o)}^{-1} \\ -\Sigma_{i(m|o)}^{-1}\Sigma_{imo}\Sigma_i^{-1} & \Sigma_{i(m|o)}^{-1} \end{pmatrix}.$$

Notice that

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_i^{-1} + \Sigma_i^{-1}\Sigma_{iom}\Sigma_{i(m|o)}^{-1}\Sigma_{imo}\Sigma_i^{-1} & -\Sigma_i^{-1}\Sigma_{iom}\Sigma_{i(m|o)}^{-1} \\ -\Sigma_{i(m|o)}^{-1}\Sigma_{imo}\Sigma_i^{-1} & \Sigma_{i(m|o)}^{-1} \end{pmatrix}.$$

There exists

$$\Sigma^{-1} = \Sigma^{-1}\mathbf{C}_i\Sigma^{-1} + \begin{pmatrix} \Sigma_i^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{32}$$

The equivalence of (9) and (29) follows from (31), (32), and by noticing that (9) can be rewritten as

$$\sum_{i=1}^{n} \text{tr}\left\{\left[w_{i2}\Sigma_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_i - \boldsymbol{\mu}_i)'\Sigma_i^{-1} - w_{i3}\Sigma_i^{-1}\right](d\Sigma_i)\right\} = 0,$$

where $d\Sigma_i$ is the differential of Σ_i.

References

Cheng, T. C., Victoria-Feser, M. P. (2002). High-breakdown estimation of multivariate mean and covariance with missing observations. *British Journal of Mathematical and Statistical Psychology*, *55*, 317–335.

Devlin, S. J., Gnanadesikan, R., Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, *76*, 354–362.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the bootstrap*. New York: Chapman & Hall.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, *31*, 1208–1211.

Godambe, V. P. (Ed.). (1991). *Estimating functions*. New York: Oxford University Press.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistent alternatives (with discussion). *Journal of the Royal Statistical Society B*, *46*, 149–192.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

Heritier, S., Cantoni, E., Copt, S., Victoria-Feser, M. P. (2009). *Robust methods in biostatistics*. Southern Gate: Wiley.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. I, pp. 221–233). Oakland: University of California Press.

Huber, P. J. (1981). *Robust statistics*. New York: John Wiley.

Johnson, R. A., Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). New Jersey: Prentice-Hall.

Kano, Y., Berkane, M., Bentler, P. M. (1993). Statistical inference based on pseudo-maximum likelihood estimators in elliptical populations. *Journal of the American Statistical Association*, *88*, 135–143.

Kelley, C. T. (2003). *Solving nonlinear equations with Newton's method*. Philadelphia: SIAM.

Kent, J. T., Tyler, D. E., Vardi, Y. (1994). A curious likelihood identity for the multivariate t-distribution. *Communications in Statistics Simulation and Computation*, *23*, 441–453.

Liang, K. Y., Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics*, *37*, 23–38.

Little, R. J. A., Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, *72*, 497–512.

Little, R. J. A., Smith, P. J. (1987). Editing and imputing for quantitative survey data. *Journal of the American Statistical Association*, *82*, 58–68.
Liu, C. (1997). ML estimation of the multivariate *t* distribution and the EM algorithm. *Journal of Multivariate Analysis*, *63*, 296–312.
Lopuhaä, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariances. *Annals of Statistics*, *17*, 1662–1683.
Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Annals of Statistics*, *4*, 51–67.
Maronna, R. A., Martin, R. D., Yohai, V. J. (2006). *Robust statistics: theory and methods*. New York: Wiley.
Maronna, R., Zamar, R. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, *44*, 307–317.
Mehrotra, D. V. (1995). Robust elementwise estimation of a dispersion matrix. *Biometrics*, *51*, 1344–1351.
Meng, X. L., van Dyk, D. A. (1997). The EM algorithm: an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society B*, *59*, 511–567.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.
Poon, W. Y., Poon, Y. S. (2002). Influential observations in the estimation of mean vector and covariance matrix. *British Journal of Mathematical and Statistical Psychology*, *55*, 177–192.
Prentice, R. L., Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, *47*, 825–839.
Richardson, A. M., Welsh, A. H. (1995). Robust restricted maximum likelihood in mixed linear models. *Biometrics*, *51*, 1429–1439.
Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *Annals of Statistics*, *24*, 1327–1345.
Rubin, D. B. (1976). Inference and missing data (with discussions). *Biometrika*, *63*, 581–592.
Ruppert, D. (1992). Computing S estimators for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics*, *1*, 253–270.
Savalei, V., Falk, C. (2014). Robust two-stage approach outperforms robust FIML with incomplete nonnormal data. *Structural Equation Modeling*, *21*, 280–302.
Schott, J. (2005). *Matrix analysis for statistics* (2nd ed.). New York: Wiley.
Sen, P. K. (1968). Estimates of the regression coefficient based on Kendalls tau. *Journal of the American Statistical Association*, *63*, 1379–1389.
Theil, H. (1950). Rank invariant method for linear and polynomial regression analysis. *Indagationes Mathematicae*, *12*, 85–91.
Tyler, D. E. (1991). Some issues in the robust estimation of multivariate location and scatter. In W. Stahel S. Weisberg (Eds.), *Directions in robust statistics and diagnostics part II* (pp. 327–336). New York: Springer-Verlag.
Wilcox, R. R. (1998). A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. *Biometrical Journal*, *40*, 261–268.
Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Waltham: Academic Press.
Yuan, K.-H., Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, *65*, 245–260.
Yuan, K.-H., Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika*, *77*, 803–826.
Yuan, K.-H., Bentler, P. M., Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*, 421–436.
Yuan, K.-H., Wallentin, F., Bentler, P. M. (2012). ML versus MI for missing data with violation of distribution conditions. *Sociological Methods & Research*, *41*, 598–629.