Michael D. Regier* and Erica E. M. Moodie

The Orthogonally Partitioned EM Algorithm: Extending the EM Algorithm for Algorithmic Stability and Bias Correction Due to Imperfect Data

DOI 10.1515/ijb-2015-0016

Abstract: We propose an extension of the EM algorithm that exploits the common assumption of unique parameterization, corrects for biases due to missing data and measurement error, converges for the specified model when standard implementation of the EM algorithm has a low probability of convergence, and reduces a potentially complex algorithm into a sequence of smaller, simpler, self-contained EM algorithms. We use the theory surrounding the EM algorithm to derive the theoretical results of our proposal, showing that an optimal solution over the parameter space is obtained. A simulation study is used to explore the finite sample properties of the proposed extension when there is missing data and measurement error. We observe that partitioning the EM algorithm into simpler steps may provide better bias reduction in the estimation of model parameters. The ability to breakdown a complicated problem in to a series of simpler, more accessible problems will permit a broader implementation of the EM algorithm, permit the use of software packages that now implement and/or automate the EM algorithm, and make the EM algorithm more accessible to a wider and more general audience.

Keywords: EM algorithm, missing data, measurement error, imperfect data, bias correction

1 Introduction

The Expectation-Maximization (EM) algorithm, introduced by Dempster, Laird, and Rubin [1], is a well-known and flexible method for obtaining a maximum likelihood estimate (MLE) when data are incomplete. It is a two-step iterative algorithm. It estimates the expectation of the *complete data* log-likelihood given the observed data and the current parameter estimates in the first step. In the second step, it maximizes the log-likelihood from the expectation step giving new values of the parameter estimates. The algorithm iterates through the Expectation step (E-step) and the Maximization step (M-step) until the sequence of parameter estimates converges to a stationary point [1, 2].

For any situation in which maximum likelihood estimation is desirable and feasible for fully observed data, estimation in the analogous incomplete data situation can be accomplished with the EM algorithm. It is this ability to associate an incomplete data problem with a complete data problem that makes it an attractive approach for missing data analysis. Unfortunately the EM algorithm has fundamental drawbacks. For example, convergence may be slow and the covariance matrix is not an instant algorithmic by-product [2]. Fortunately, there are proposals for speeding up EM convergence, such as the Parameter-Expanded EM algorithm [3], and for obtaining the covariance matrix [4, 5].

The EM algorithm was originally formulated as an approach to obtain MLEs when there are missing data, but it has also been applied to analyses where there are mismeasured variables. Dawid and Skene [6] used the EM algorithm to correct misclassification when the true value was never observed, but multiple measures for each subject were obtained. In this situation, the multiple measurements permit the

*Corresponding author: Michael D. Regier, Department of Biostatistics, West Virginia University, Morgantown, WV, USA, E-mail: mregier@hsc.wvu.edu

Erica E. M. Moodie, Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montreal, QC, Canada, E-mail: erica.moodie@mcgill.ca

estimation of the misclassification rates allowing for the estimation of the underlying true population proportion. Drews et al. [7] consider the situation where the exposure is misclassified not by one classification scheme, but by two schemes (patient interviews and medical records). They use the EM algorithm to obtain corrected odds ratio estimates in this situation. Ganguli et al. [8] proposed a variation of the Monte Carlo EM (MCEM) algorithm specific to structural nonparametric regression problems with measurement error in covariates for additive models; it is referred to as the Ganguli, Staudenmayer, Wand EM.

Despite the transportation of the EM approach from missing data problems to the problem of measurement error, and the wealth of literature about the EM algorithm for missing data problems, there has been limited application of the EM algorithm to measurement error problems. Gustafson [9] comments that the lack of a closed form solution of the E-step for many problems, the requirement of an approximate solution, and the additional work required for covariance estimation has contributed to the scarceness of EM approaches as a proposed solution when mismeasured variables are present. This is echoed by Buonaccorsi [10] who comments on the unwieldy nature of the EM algorithm for the calibration adjustment of measurement error in the response of a linear model.

In many situations, there are alternate approaches that can be used. For missing data, weighting [11] or multiple imputation methods [12–14] often are viable options. For measurement error, alternative methods include simulation-extrapolation [15–17], instrumental variables [18], calibration [19], and imputation [20, 21]. Our interest in the likelihood approach is that it is a method of correction for *both* missing data and measurement error problems. Thus, there is considerable appeal to this approach, which can handle both aspects of “imperfect data” in the same model, within a unified methodological framework.

We consider the situation where the EM algorithm is a viable and potentially attractive approach for obtaining corrected parameter estimates for imperfect data. Of particular interest (Section 2.2) is the situation where there are missing responses, mismeasured continuous explanatory variables, no closed form solution for the E-step, and the algorithm has an unacceptably low probability of convergence. We propose a new variation of the EM algorithm, inspired by the expectation-conditional maximization (ECM) of Meng and Rubin [22] and the use of tangent spaces by Tsiatis [23], that maintains the ability to correct for both missing data and measurement error while overcoming non-convergence. This makes it a practical option for applied statistical research. With the emergence of software packages that now implement and/or automate the EM algorithm such as EMCluster [24], our proposal may permit the direct use of these new packages for increasingly complex problems as well as making the algorithm accessible to a wider audience.

In Section 2, we present the EM algorithm, motivating problem and imperfect data. In Section 3, we present the proposed variation of the EM algorithm and the general theory for maximization using a sequence of self-contained, smaller and simpler EM algorithms. We use a simulation study to consider the finite sample performance of the proposed variation of the EM algorithm in Section 4. Section 5 discusses possible future applications of the proposed method.

2 The EM algorithm, imperfect data and the motivating problem

2.1 The EM algorithm

When there are imperfect data (i.e. missing and mismeasured data), we can obtain MLEs by maximizing the log-likelihood of the observed portion of the data. The log-likelihood associated with the observed data, ℓ_o , is

$$\ell_o(\theta^Y | \mathbf{Y}_o = \mathbf{y}_o) = \int_{\mathbf{y}_m \in \mathcal{Y}_m} \log \left[\prod_{i=1}^n f(\mathbf{Y}_i | \theta^Y) \right] d\mathbf{y}_m.$$

where $\mathbf{Y}_i$ denotes the vector of variables intended for collection from the i^{th} observational unit, $f(\mathbf{Y}_i | \theta^Y)$ is the pdf of $\mathbf{Y}_i$ parameterized by θ^Y , and the variable $\mathbf{Y}_i = \{\mathbf{Y}_{i,o}, \mathbf{Y}_{i,m}\}$ such that $\mathbf{Y}_{i,o}$ refers to the elements of $\mathbf{Y}_i$ that are observable and $\mathbf{Y}_{i,m}$ refers to the elements of $\mathbf{Y}_i$ that are missing. We let $\mathbf{y}_i = \{\mathbf{y}_{i,o}, \mathbf{y}_{i,m}\}$ denote the

realization of $\mathbf{Y}_i = \{\mathbf{Y}_{i,o}, \mathbf{Y}_{i,m}\}$ where $\mathbf{y}_{i,o}$ denote the subset of observed variables and $\mathbf{y}_{i,m}$ denote the subset of variables for which realizations are not observable [25, 26]. Finally, $\mathscr{Y}_m$ is the space associated with $\mathbf{y}_m$. Without a closed form solution, the maximization of $\ell_o(\theta^Y | \mathbf{Y}_o = \mathbf{y}_o)$ can be cumbersome.

The EM algorithm provides an indirect approach for obtaining the MLE for θ^Y . The process begins by obtaining an estimate of θ^Y using only the completely observed portion of the data, denoted $\theta^{Y(0)}$. Given this initial estimate of the parameters, the E-step finds the expectation of $\ell(\theta^Y | \mathbf{y})$ given the observed data, $\mathbf{y}_o$, and the current estimate of θ^Y , $\theta^{Y(0)}$. The M-step maximizes $E\{\ell(\theta^Y | \mathbf{Y}) | \mathbf{y}_o; \theta^{Y(0)}\}$ from the E-step, yielding an updated estimate of the parameter, $\theta^{Y(1)}$. The E- and M-steps are repeated with $\theta^{Y(1)}$ used in place of $\theta^{Y(0)}$.

The algorithm iterates through E- and M-steps; the E-step for the $(t+1)$ th iteration is defined as

$$Q(\theta^Y | \theta^{Y(t)}) = E\{\ell(\theta^Y | \mathbf{Y}) | \mathbf{y}_o; \theta^{Y(t)}\}, \quad (1)$$

where $\theta^{Y(t)}$ is the current parameter estimate obtained from the t th iteration. The M-step, requires that we select the updated parameter, $\theta^{Y(t+1)}$ that maximizes the Q-function (Equation (1)) over the parameter space, Θ^Y , such that

$$Q(\theta^{Y(t+1)} | \theta^{Y(t)}) \geq Q(\theta^Y | \theta^{Y(t)}). \quad (2)$$

Under the regularity conditions of Dempster et al. [1] and Wu [27], the likelihood, $L(\cdot)$, increases with each iteration

$$L(\theta^{Y(t+1)}) > L(\theta^{Y(t)}),$$

requiring only that Equation (1) be continuous in θ^Y . This is a weak condition that is held in many practical situations, and it holds for the exponential family [2].

The E- and M-steps are repeated until the difference between sequential steps in the likelihood changes by an arbitrarily small amount, δ ,

$$\|L(\theta^{Y(t+1)}) - L(\theta^{Y(t)})\| < \delta.$$

Pragmatically, the stopping criterion

$$\|\theta^{Y(t+1)} - \theta^{Y(t)}\| < \delta. \quad (3)$$

is used, probably motivated by Wu [27] who showed

$$\|\theta^{Y(t+1)} - \theta^{Y(t)}\| \rightarrow 0$$

as $t \rightarrow \infty$, if the limit points are in a compact, connected subset of all stationary points.

The generalized EM algorithm (GEM) relaxes the M-step condition. It requires that Equation (2) holds, permitting the selection of $\theta^{Y(t+1)}$ such that the Q-function increases. This avoids maximizing over all $\theta^Y \in \Theta^Y$, while ensuring that the likelihood does not decrease after each iteration [1]. The limit point of the sequence of estimated parameters is a stationary point of the likelihood, thus providing a local maximum for θ^Y .

Three relevant variations of the EM algorithm are the Monte Carlo EM (MCEM) [28], expectation-conditional maximization (ECM) of Meng and Rubin [22], and the multi-cycle ECM algorithms [22]. The MCEM algorithm replaces a complicated and analytically intractable expectation in the E-step by approximating the integral with a finite sum,

$$Q_m(\theta^Y | \theta^{Y(t)}) = m^{-1} \sum_{r=1}^m \ell(\theta^Y | y_o, y_{m,r}). \quad (4)$$

where m denotes the Monte Carlo sample size. In the $(t+1)$ th E-step, the missing data, Y_m , is simulated from the current conditional distribution $f(Y_m | y_o; \theta^{Y(t)})$ [28]. For each iteration, the ECM has a single E-step followed by a sequence of conditional maximization steps, CM-steps. The constraint function for the CM-

steps is defined such that maximization occurs over the entire parameter space [22]. The multi-cycle ECM extends the ECM algorithm. For each iteration, the multi-cycle ECM permits an E-step before a few select CM-steps. The extra computational effort within each iteration results in larger increases in the likelihood, but does introduce the potential for the algorithm to not be a GEM [22].

2.2 Imperfect data and the motivating problem

For subjects $i = 1, \dots, n$, let X_i denote imperfect data variables as (X_i^*, X_i) . If there is only missing data then $(X_i^*, X_i) = (X_{i,o}, X_{i,m})$ so that $X_i^* = \emptyset$. On the other hand, if there is only measurement error then $(X_i^*, X_i) = (X_{i,o}^*, X_{i,o})$ so that $X_{i,m}^* = X_{i,m} = \emptyset$. Finally, if there is both missing data and measurement error, then $(X_i^*, X_i) = (X_{i,o}^*, X_{i,m}^*, X_{i,o}, X_{i,m})$. Realizations of measurement error and missing data have similarities; for example, both have at least one variable for which direct observation may not be possible. For measurement error, this is the true value, x_i , and for missing data, this is $x_{i,m}$.

In what follows, we assume that X_i is only measured with error, $(X_i^*, X_i) = (X_{i,o}^*, X_{i,o})$, and that all the measurement error of all subjects' follows the same distribution and a classical, unbiased, additive, nondifferential measurement error model. The classical, unbiased, additive, nondifferential measurement error model for the i th subject and a random variable, X , is defined as $X_{i,o}^* = X_{i,o} + \varepsilon_i$ where $\varepsilon_i \sim N(0, \theta^M)$ for $i = 1, \dots, n$ and where θ^M parameterizes the distribution of ε_i [29]. We observe $x_{i,o}^*$ and not $x_{i,o}$ for each of the i units of interest.

The measurement error is unbiased, $E(X^*|x) = x$, and non-differential (surrogate) defined by the assumption that $X^*|x, y = X^*|x$ or equivalently X^* and Y are independent given x [10, 29]. This suggests that a surrogate is the measured substitute for the desired variable, but it gives no information beyond that which would have been given by the true variable. When there is more than one mismeasured variable in the model, then we add the further assumption that the associated errors are independent, that is for mismeasured variables $X_{j,o}^*$ and $X_{k,o}^*$ we assume that $\varepsilon_{ij} \perp \varepsilon_{ik}$. A measurement error indicator, $C_i^M \in \{0, 1\}$ with $C_i^M = 0$ defining the observation of the true value, may be defined as the probability of measurement error in a manner analogous to the following definition for the probability of missingness.

We let Y_i denote the imperfect response and assume that the response is subject to missing data, $Y_i = (Y_{i,o}, Y_{i,m})$. Rubin [30] identified three basic types of missing data mechanisms: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). For the i th subject, define the missing data indicator as $C_i \in \{0, 1\}$ with $C_i = 0$ defining observation. MCAR occurs when the probability of not observing a response does not depend either the observed or unobserved data, $f(C_i|Y_i = y_i, \theta^C) = f(C_i|\theta^C)$, where θ^C parameterizes the missing data mechanism. MAR occurs when the probability of an observation being missing is independent of the unobserved values given the observed values, $f(C_i|Y_i = y_i, \theta^C) = f(C_i|Y_{i,o} = y_{i,o}, \theta^C)$, where $y_{i,o}$ is the observed value of y_i . An MNAR missing data mechanism occurs when the probability of an observation being missing depends on both the observed and unobserved values, $f(C_i|Y_i = y_i, \theta^C)$.

Assuming unique parameterization of the conditional distributions [9, 26] while restricting our exposition to classical unbiased nondifferential measurement error, MCAR and MAR missing data mechanisms, the joint distribution may be expressed as

$$\begin{aligned} f(Y_i, C_i, X_i, X_i^*, Z_i|\theta) &= f(Y_i|x_i, \mathbf{z}_i; \theta^Y) \\ &\quad \times f(C_i|x_i, \mathbf{z}_i; \theta^C) \\ &\quad \times f(X_i^*|x_i; \theta^M) \\ &\quad \times f(X_i|\mathbf{z}_i; \theta^X) \\ &\quad \times f(Z_i|\theta^Z) \end{aligned} \quad (5)$$

where $\mathbf{Z}_i$ denote complete variables (i.e. no missing data and no measurement error), and $\theta = \{\theta^Y, \theta^C, \theta^M, \theta^X, \theta^Z\}$.

The joint distribution (Equation 5) has been decomposed into the product of four conditional distributions and the marginal distribution of the perfect covariates $\mathbf{Z}$. The *outcome model*, $f(Y_i|x_i, \mathbf{z}_i; \theta^Y)$, describes the relationship between the variables of interest, $(X, \mathbf{Z})$ and the outcome. The *missing data mechanism*, $f(C_i|x_i, \mathbf{z}_i; \theta^C)$ is parameterized by θ^C . The *measurement model*, $f(X_i^*|x_i; \theta^M)$ is the conditional density of X^* given the unobserved, true value, $X = x$. The final conditional and marginal distributions comprise the joint distribution of the covariates, sometimes referred to as the *exposure model* in epidemiological applications.

The outcome model $f(Y_i|x_i, \mathbf{z}_i; \theta^Y)$ is of primary interest as is an estimate of θ^Y . Typically, the remaining parameters, $\{\theta^C, \theta^M, \theta^X, \theta^Z\}$, are treated as nuisance parameters. If the nuisance parameter space is orthogonal to the parameter space of interest, then an unbiased estimator for θ^Y can be obtained provided the nuisance parameter estimator is root- n consistent [23].

We have a primary interest in the outcome model, $f(Y_i|x_i, \mathbf{z}_i; \theta^Y)$, and a secondary interest in the missing data mechanism, $f(C_i|x_i, \mathbf{z}_i; \theta^C)$. As such, we are interested in the unbiased estimation of $\{\theta^Y, \theta^C\}$. Under the specification of unique parameterization, the assumption is that $\{\theta^Y, \theta^C\}$ is orthogonal to $\{\theta^M, \theta^X, \theta^Z\}$. The challenge is that we are interested in a parameter that is commonly relegated to the nuisance space. An example of possible application (Section 5) is the use of inverse probability weights (IPW) when there is mismeasurement in at least one variable used to model the weights and the response, and the response may be missing [31]. In this scenario, we wish to obtain “good” estimates of the parameters in the weighting models to ensure that the inverse probability weighted estimators exhibit reduced or no bias.

Motivated by this example, we used a simulation study to examine the finite-sample bias reduction for both $\{\theta^Y, \theta^C\}$, using known implementations of the EM algorithm. Due to lack of convergence of the algorithm for a standard implementation, we implemented the ECM and the multi-cycle ECM algorithm with Monte Carlo integration. By using a single EM algorithm to obtain estimates, we observed that θ^Y and θ^C were have problems converging. Having a unified approach resulted in a lack of convergence for the single EM algorithm. Convergence rates, across the three implementations of the EM algorithm, ranged from 0% to 60%, suggesting unstable algorithms. To overcome this, we sought to partition the parameter space orthogonally and implement separate EM algorithms for each partition. By having root- n consistent estimators in an particular subspace, we would have unbiased estimators that converge to a stationary point in the complementary subspace.

3 The orthogonally partitioned EM algorithm

Let $\mathbf{X}$ be a vector-valued random variable with a joint distribution $f(\mathbf{X}|\theta)$, where $\theta \in \Theta$ and Θ denotes the associated parameter space. The likelihood, $L(\theta|\mathbf{x})$, is said to be uniquely parameterized if it can be written as a product of conditional distributions,

$$L(\theta|\mathbf{x}) = \prod_{j=1}^p f(X_{p-j+1}|x_1, \dots, x_{p-j}; \theta^{p-j+1}), \quad (6)$$

such that $\theta^i \in \Theta^i$, $\theta^j \in \Theta^j$, $\Theta^i \subset \Theta$, $\Theta^j \subset \Theta$, $\Theta^i \perp \Theta^j$ for all $i \neq j$, $i = 1, \dots, p$ and $j = 1, \dots, p$, and $x_0 \equiv \emptyset$.

Unique parameterization has been used, without a formal definition as a means to specify complex joint distributions through a product of conditional ones [32–36]. Proposition 1 formalizes unique parameterization, ensuring that the parameter space is comprised of orthogonal subspaces.

Proposition 1 *A likelihood may be uniquely parameterized if and only if*

$$\Theta = \bigoplus_{j \in \{1, \dots, p\}} \Theta^j,$$

for $p \geq 1$, where $\bigoplus$ denotes the direct sum of $\Theta^j \subseteq \Theta$.

Proof: See Appendix.

The proposed extension of the EM algorithm begins with an initial estimation of θ denoted $\theta^{(0)} = \{\theta^{Y(0)}, \theta^{C(0)}, \theta^{M(0)}, \theta^{X(0)}, \theta^{Z(0)}\}$. Proposition 1 permits the construction of an orthogonal partitioning set of subspaces using direct-sum decomposition [49], $\{\Theta^k: k=1, \dots, K\}$. The ability to partition our parameter space for unique parameterization allows a general formulation for the E-step of k^{th} stage in the orthogonally partitioned EM (OPEM) algorithm,

$$\tilde{Q}(\theta^{(\tau+1)}|\theta^{(\tau)}) = \sum_{k=1}^K Q(\theta^{k(t_k+1)}|\theta^{k(t_k)}, \hat{\theta}^{1(t_1')}, \dots, \hat{\theta}^{k-1(t_{k-1}')} , \theta^{k+1(0)}, \dots, \theta^{K(0)}) \quad (7)$$

where $\tau+1$ is the $(t_k+1)^{\text{th}}$ step for the k^{th} partition given the t^{th} step of the of the k^{th} partition and the maximized values $\hat{\theta}^{1(t_1')}, \dots, \hat{\theta}^{k-1(t_{k-1}')}$ obtained on steps $t_1', \dots, t_{k-1}'$ of the respective $k-1$ EM algorithms, where $\tau = \{t_k, t_1', \dots, t_{k-1}'\}$ of the k^{th} stage. For the k^{th} stage, partitions from $k+1$ to K have their parameter estimates fixed at the initialized values $\{\theta^{k+1(0)}, \dots, \theta^{K(0)}\}$ and parameter estimates from stages 1 to $(k-1)$ fixed at their stage estimate values $\{\theta^{1(t_1')}, \dots, \theta^{k-1(t_{k-1}')} \}$

Equation (5) is used to illustrate Equation (7) and motivate the simulation study in Section 4. For the i^{th} subject, the $(t+1)^{\text{th}}$ E-step associated with Equation (5) is

$$\begin{aligned} & E \left[\ell(\theta|y_i, c_i, x_i, x_i^*, z_i) | y_{i,o}, c_i, x_i^*, z_i; \theta^{(t)} \right] \\ &= \int_{\mathcal{Y}_m} \int_{\mathcal{X}} \log f(Y_i | x_i, z_i; \theta^Y) f(Y_{i,m}, X_i | y_{i,o}, c_i, x_i^*, z_i; \theta^{(t)}) d\mathbf{y}_{i,m} d\mathbf{x}_i \\ &+ \int_{\mathcal{Y}_m} \int_{\mathcal{X}} \log f(C_i | x_i, z_i; \theta^C) f(Y_{i,m}, X_i | y_{i,o}, c_i, x_i^*, z_i; \theta^{(t)}) d\mathbf{y}_{i,m} d\mathbf{x}_i \\ &+ \int_{\mathcal{Y}_m} \int_{\mathcal{X}} \log f(X_i^* | x_i; \theta^M) f(Y_{i,m}, X_i | y_{i,o}, c_i, x_i^*, z_i; \theta^{(t)}) d\mathbf{y}_{i,m} d\mathbf{x}_i \\ &+ \int_{\mathcal{Y}_m} \int_{\mathcal{X}} \log f(X_i | z_i; \theta^X) f(Y_{i,m}, X_i | y_{i,o}, c_i, x_i^*, z_i; \theta^{(t)}) d\mathbf{y}_{i,m} d\mathbf{x}_i \\ &+ \int_{\mathcal{Y}_m} \int_{\mathcal{X}} \log f(Z_i | \theta^Z) f(Y_{i,m}, X_i | y_{i,o}, c_i, x_i^*, z_i; \theta^{(t)}) d\mathbf{y}_{i,m} d\mathbf{x}_i \end{aligned} \quad (8)$$

where $\mathcal{X}$ is the support of X . For simplicity, let $K=2$ and partition Θ such that $\Theta = \Theta^1 \oplus \Theta^2$ where $\Theta^1 = \bigoplus_{j \in \{C, M, X, Z\}} \Theta^j$ and $\Theta^2 = \Theta^Y$. The subspaces for each of the k -stages may be determined by the context of the research or to provide algorithmic stability.

For the first stage, $\theta^1 \in \Theta^1$, we have

$$f(Y_{i,m}, X_i | y_{i,o}, c_i, x_i^*, z_i; \theta^{(t)}) = \frac{f(C_i | x_i, z_i; \theta^C) f(X_i^* | x_i; \theta^M) f(X_i | z_i; \theta^X)}{\int_{\mathcal{X}} f(C_i | x_i, z_i; \theta^C) f(X_i^* | x_i; \theta^M) f(X_i | z_i; \theta^X) d\mathbf{x}_i} w_i^1$$

where

$$w_i^1 = \frac{\int_{\mathcal{X}} f(C_i | x_i, z_i; \theta^C) f(X_i^* | x_i; \theta^M) f(X_i | z_i; \theta^X) d\mathbf{x}_i}{\int_{\mathcal{X}} \int_{\mathcal{Y}_m} f(Y_i | x_i, z_i; \theta^Y) d\mathbf{y}_{i,m} f(C_i | x_i, z_i; \theta^C) f(X_i^* | x_i; \theta^M) f(X_i | z_i; \theta^X) d\mathbf{x}_i}.$$

The first stage Q-function at the $(t_1 + 1)^{\text{th}}$ step, $Q(\theta^{1(t_1+1)} | \theta^{1(t_1)})$ does not have a closed form; it is evaluated using Monte Carlo integration with a Gibbs sampler [33, 28]. Maximum likelihood estimates for the parameters are obtained using a Gibbs sampler and the adaptive rejection algorithm [37]. Sampling is from $f(Y_{i,m}, X_i | y_{i,o}, c_i, x_i^*, z_i; \theta^{(t)})$, where

$$f(Y_{i,m}, X_i | y_{i,o}, c_i, x_i^*, z_i; \theta^{(t)}) \propto f(C_i | x_i, z_i; \theta^C) f(X_i^* | x_i; \theta^M) f(X_i | z_i; \theta^X)$$

for θ^1 . The first stage Q-function for the i^{th} subject at the $(t_k + 1)^{\text{th}}$ iteration is written as,

$$Q(\theta^{1(t_k+1)} | \theta^{1(t_k)}) \approx \frac{1}{m_1} \sum_{r=1}^{m_1} [\log f(C_i | x_{ir}, z_i; \theta^C) + \log f(X_i^* | x_{ir}; \theta^M) + \log f(X_i | z_i; \theta^X)]$$

The argument for the second stage Q-function at the $(t_2 + 1)^{\text{th}}$ step,

$$Q(\theta^{2(t_2+1)} | \hat{\theta}^{1(t_1')}, \theta^{2(t_2)})$$

parallels that of the first stage.

Each stage can be thought of as a constrained maximization, where the constraint is to an orthogonal subspace of θ . An EM algorithm or a generalized EM algorithm requires the optimizing constraint to be space-filling [22] and to have the appropriate convergence properties. If the space-filling criterion holds, then we obtain an unconstrained maximum. If the algorithm is a generalized EM algorithm, then convergence to a stationary point is quickly obtained.

Proposition 2 *If the E-step can be written in the form of Equation (7) and $\tilde{Q}(\theta^{(\tau+1)} | \theta^{(\tau)})$ is maximized subject to the constraint function $G = \{g_k(\theta) = \theta/\theta^k | k = 1, \dots, K\}$, where $g_k(\theta)$ is the vector of all parameter subvectors in θ except θ^k , using K Monte Carlo EM stages, then we call the algorithm an orthogonally partitioned EM (OPEM) algorithm.*

Beginning with the space-filling property, the gradient of G , $\nabla g_k(\theta)$, is full rank at $\theta^{k(t)} \in \Theta$ for all k and t . It is a set of d_k elementary column vectors and a set of $d_\Theta - d_k$ zero column vectors, where $d_\Theta = \dim(\Theta)$ and $d_k = \dim(\theta^k)$. Taking $\eta \in \mathbb{R}^{d_k}$, the column space of the gradient is $\mathcal{C}_k(\theta) = \{\eta \nabla g_k(\theta) | \eta \in \mathbb{R}^{d_k}\}$. The intersection of the column spaces, using Proposition 1, is

$$\mathcal{C}(\theta) = \bigcap_k \mathcal{C}_k(\theta) = \emptyset,$$

with the complement

$$\begin{aligned} \mathcal{C}(\theta)^c &= \bigcup_k \{\eta \nabla g_k(\theta) | \eta \in \mathbb{R}^{d_k}\}^\perp \\ &= \mathbb{R}^{d_\Theta}. \end{aligned}$$

Meng and Rubin [22] showed that having the convex hull of all feasible directions determined by the constraint spaces, $\{\theta \in \Theta | g_k(\theta)\}$, $k = 1, \dots, K$, is equivalent to $\mathbb{R}^{d_\Theta}$, the Euclidean space of dimension d_Θ . This is equivalent to the space-filling criterion as defined using tangent cones.

If each of the k , $k = 1, \dots, K$, Q-functions is a generalized EM algorithm, then Proposition 3 states that the entire algorithm is a generalized EM algorithm. To address the problem of non-monotonicity of the MCEM algorithm [38], we introduce the requirement that there exists an iteration, $t_k \geq 1$, for which the k^{th} Q-function increases with each iteration.

Proposition 3 *If we have an OPEM (Proposition 2) and there exists a $t_k \geq 1$ for which $Q(\theta^{k(t_k+1)} | \theta^{k(t_k)}) \geq Q(\theta^{k(t_k)} | \theta^{k(t_k)})$ for all subsequent steps then the algorithm is a generalized EM algorithm.*

Proof: See Appendix.

Dempster, Laird, and Rubin [1] and Wu [27] provide two methods to show convergence to stationary points. The simplest is to view the algorithm as a single process. By the Inverse Mapping Theorem and by Meng and Rubin [22], Theorem 2 proof part (i), it is a closed mapping. This directly implies that the limit points are stationary points (Wu [27], Theorem 1).

An alternate approach views the algorithm as the sum of K generalized EM algorithms. Each converges to a limiting point which is a stationary point [27]. In each stage, Monte Carlo integration is used with the relevant EM variant to obtain the maximizer of $Q(\theta^{k(t_k+1)}|\theta^{k(t_k)})$. With space-filling G , we have the following variation of Theorem 3 in Meng and Rubin [22].

Proposition 4 *If all K maximizations of Equation (7) are unique, then the limit points of the K OPEM stages $\{\theta^{k(t_k')}; t_k' \geq 0, k = 1, \dots, K\}$ are stationary points of $L_o(\theta|x_o)$ if G is space filling.*

By extending Wu's Theorem 6 [27] and replacing the uniqueness condition with the continuity of $[\partial Q(\theta^k|\theta^{k'})/\partial \theta^k]_{\theta^k=\theta^{k'}}$ in both θ^k and $\theta^{k'}$, and $\nabla g_k(\theta)$ for all $k, k = 1, \dots, K$, we can relax the assumption that all conditional maximizations are unique.

4 Numerical example

4.1 Simulation specification

We consider a finite-sample simulation study using the general model given in Equation (5). As noted in Section 2.2, we are interested in the unbiased estimation of $\{\theta^Y, \theta^C\}$. For the i th subject, we specify the exposure model (X_i, Z_i) , $i = 1, \dots, n$, to be independently and normally distributed with mean zero, $\sigma_{X,X} = \sigma_{Z,Z} = 1$, and $\sigma_{X,Z} = 0.2$, where $\{\theta^X, \theta^Z\} = \{\mu_X, \mu_Z, \sigma_{X,X}, \sigma_{Z,Z}, \sigma_{X,Z}\}$. The response Y_i is binary with $Y_i = 1$ denoting the occurrence of the event of interest. The missing data indicator for the response, C_i , denotes missing responses as $C_i = 1$.

The response is modeled as a logistic regression,

$$\text{logit}\{P(Y_i = 1|x_i, z_i; \theta^Y)\} = \theta_0^Y + \theta_1^Y x_i + \theta_2^Y z_i,$$

where $\theta^Y = (\theta_0^Y, \theta_1^Y, \theta_2^Y)$. The missing data mechanism is modeled as

$$\text{logit}\{P(C_i = 1|x_i, z_i; \theta^C)\} = \theta_0^C + \theta_1^C x_i + \theta_2^C z_i,$$

where $\theta^C = (\theta_0^C, \theta_1^C, \theta_2^C)$. We set $\theta_0^Y = 0.40$ and $\theta_0^C = 0.03$ and the remaining parameters were randomly generated and restricted to lie in the range $(-3, 3)$ for each dimension (Table 1).

Table 1: Simulation parameterization for the response and missing data mechanism models.

Model	θ_0^C	θ_1^C	θ_2^C	θ_0^Y	θ_1^Y	θ_2^Y
1	0.03	0.38	-1.88	0.40	2.63	2.31
2	0.03	-0.18	-2.82	0.40	2.45	0.86
3	0.03	1.04	0.89	0.40	1.86	-0.68

Imperfect data are introduced in the following manner. Let X be measured with error using a classical unbiased nondifferential measurement error model, $X^* = X + \varepsilon$. Here, $\varepsilon \sim N(0, \theta^M)$. Bias reduction as a function of θ^M is explored to consider the effect of imprecision. We let θ^M take values $\{0.1, 0.3, 0.5, 0.7\}$. With $\sigma_{X,X} = 1$, θ^M is directly interpreted as the proportion of imprecision in the measurement of X [9].

In applied settings, θ^M may be estimated through the collection of auxiliary information or by inferring a measurement error model from prior studies. This approach prevents the problem of having a nonidentifiable model ([9], section 4.1). Given that we are not generating auxiliary information from which estimates of θ^M may be derived, we specify θ^M for all simulations.

We let $K=2$ for our simulation study for a two-stage OPEM. The first stage targets θ^C with $\{\theta^M, \theta^X, \theta^Z\}$ as nuisance parameters and θ^Y fixed at its initialization value, θ_0^Y . This stage is an implementation of the MCEM algorithm, given θ_0^Y . The second stage estimates θ^Y given the stage one estimates.

For each parameterization, we implemented $S=100$ simulations and each simulation had $n=500$ observations. We replicated this at each of the four specifications of measurement error noise, θ^M . For the Monte Carlo integration, we used a burn-in of 1,000 and a Monte Carlo sample of $m=2500$ as determined by a small Gibbs sampler convergence study and relevant recommendations [33, 39]. We used an MCEM algorithm for each stage of the 2-stage OPEM, and a dissimilarity metric based on a one step lag (e.g. Equation (3)) with measure less than 0.0025 to terminate the algorithm.

Due to the low probability of convergence using the MCEM, ECM and multi-cycle ECM algorithm with Monte Carlo integration (Section 2.2), we compare the OPEM with the naive analysis that assumes complete data. This is a reasonable contrast since the OPEM is not a direct extension of the ECM or multi-cycle ECM, but a new option for implementing an EM algorithm when convergence is problematic. The proposed OPEM algorithm has the advantage of convergence when alternate implementations of the EM algorithm have a low probability of convergence. All simulations were implemented in R [50].

4.2 Simulation results

For the estimated coefficients $\hat{\theta}_{1,\text{naive}}^C$ and $\hat{\theta}_{1,\text{naive}}^Y$ of the naive analysis, we expect either attenuation, estimates that are smaller in magnitude than the true value, or augmentation, estimates that are larger in magnitude than the true value ([9], section 2.6). Furthermore, we expect an enhanced effect due to the modest correlation between X and Z ([9], Section 2.2) and induced bias for $\hat{\theta}_{2,\text{naive}}^C$ and $\hat{\theta}_{2,\text{naive}}^Y$ ([10], section 5.3). If the OPEM algorithm performs as theory suggests in Section 3, we should observe bias reduction for the affected estimators without inducing bias in those that were largely unaffected by measurement error. Finally, we note that *all of the simulations converged* which is a substantial improvement over the convergence rates for standard implementations of the EM algorithm to this problem (Section 2.2).

Although our two-stage OPEM algorithm first estimated $\{\hat{\theta}^C, \hat{\theta}^M, \hat{\theta}^X\}$ and then estimated $\hat{\theta}^Y$ given $\{\hat{\theta}^C, \hat{\theta}^M, \hat{\theta}^X\}$, we first present the results for response model since it is of primary interest. We follow the response model results with those of secondary interest, the parameter estimation results for the missing data mechanism.

4.2.1 Response mechanism model results, θ^Y

The estimates of the outcome model were obtained using an MCEM algorithm, conditional on the parameter estimates from the first stage, $\{\hat{\theta}^C, \hat{\theta}^M, \hat{\theta}^X\}$ (Section 4.2.2). Table 2 shows that the OPEM provides bias reduction for θ_1^Y , across the levels of θ^M , when compared to the bias of the naive estimator. This ranges from a 33% reduction to a 98% reduction in the bias. There is one situation where a slight inflation of the OPEM bias occurs, model 3 when $\theta^M=0.1$. The proportion of imprecision in x^* is small and the associated confidence intervals suggest that naive and OPEM biases are similar. The same conclusion can be drawn for model 2 when $\theta^M=0.1$, even though the OPEM does reduce the bias in this situation.

Bias reduction for θ_0^Y and θ_2^Y is observed. In particular, the reduction is more notable than those observed for θ_0^C and θ_2^C (Section 4.2.2). Part of this is related to the magnitude of the naive estimator's bias (e.g. model 3, θ_2^Y when $\theta^M=0.7$), but it is also noticeable where the naive estimator's bias was modest (e.g. model 1, θ_0^Y when $\theta^M=0.7$). As before, there is a slight increase in the standard errors of the OPEM

Table 2: Bias assessment for simulated example: θ_{naive}^Y denotes the use of complete data and θ_{OPEM}^Y denotes the application of the OPEM to adjust for incomplete information.

Model	θ^M	θ_{naive}^Y			θ_{OPEM}^Y		
		θ_{0^Y} (SE)	θ_1^Y (SE)	θ_2^Y (SE)	θ_0^Y (SE)	θ_1^Y (SE)	θ_2^Y (SE)
1	0.1	0.01 (0.01)	-0.06 (0.02)	-0.01 (0.02)	-0.01 (0.01)	0.01 (0.2)	0.02 (0.02)
	0.3	0.02 (0.01)	-0.39 (0.02)	-0.09 (0.02)	-0.01 (0.02)	0.06 (0.03)	0.08 (0.03)
	0.5	0.06 (0.02)	-0.90 (0.02)	-0.28 (0.02)	0.01 (0.02)	0.07 (0.04)	0.05 (0.03)
	0.7	0.10 (0.01)	-1.30 (0.01)	-0.41 (0.02)	-0.01 (0.02)	0.20 (0.06)	0.12 (0.05)
2	0.1	0.02 (0.01)	-0.03 (0.02)	0.01 (0.02)	0.02 (0.01)	0.02 (0.02)	0.01 (0.02)
	0.3	0.04 (0.01)	-0.36 (0.02)	-0.03 (0.01)	0.02 (0.01)	0.02 (0.03)	-0.02 (0.02)
	0.5	0.07 (0.01)	-0.80 (0.02)	-0.07 (0.01)	-0.01 (0.02)	0.07 (0.04)	0.01 (0.02)
	0.7	0.09 (0.01)	-1.19 (0.01)	-0.07 (0.01)	0.01 (0.02)	0.10 (0.05)	0.02 (0.03)
3	0.1	-0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	-0.01 (0.01)	0.03 (0.02)	0.01 (0.01)
	0.3	0.02 (0.01)	-0.20 (0.01)	0.04 (0.01)	0.01 (0.01)	0.07 (0.02)	-0.03 (0.01)
	0.5	0.05 (0.01)	-0.55 (0.01)	0.13 (0.01)	0.01 (0.01)	0.01 (0.02)	-0.02 (0.01)
	0.7	0.06 (0.01)	-0.82 (0.01)	0.21 (0.01)	-0.01 (0.01)	0.06 (0.03)	-0.02 (0.02)

estimates. When compared against the bias reduction, we have overall gains in terms of the MSE and bias-variance trade-off.

4.2.2 Missing data mechanism results, θ^C

We observe the expected attenuation and augmentation in θ_1^C for the naive estimate. As expected, we observe bias reduction for the OPEM implementation since the first stage in any K -stage OPEM is a simple EM algorithm.

We observe modest bias reduction for model 1 across all levels of θ^M (Table 3). For models 2 and 3, we observe much larger bias reductions across the four levels of θ^M , ranging from a 33% reduction in the bias of θ_1^C , to a 95% reduction. The reduction is greatest when proportion of imprecision is large, $\theta^M = 0.5, 0.7$.

We observe a nominal inflation of the associated standard errors with the OPEM. In terms of mean squared error and bias-variance trade-off, the OPEM performs well when the proportion of imprecision in

Table 3: Bias assessment for simulated example: θ_{naive}^C denotes the use of complete data and θ_{OPEM}^C denotes the application of the OPEM to adjust for incomplete information.

Model	θ^M	θ_{naive}^C			θ_{OPEM}^C		
		θ_0^C (SE)	θ_1^C (SE)	θ_2^C (SE)	θ_0^C (SE)	θ_1^C (SE)	θ_2^C (SE)
1	0.1	-0.11 (0.04)	0.05 (0.02)	-0.08 (0.03)	-0.11 (0.04)	0.05 (0.02)	-0.08 (0.03)
	0.3	-0.10 (0.04)	-0.01 (0.02)	-0.06 (0.03)	-0.11 (0.04)	0.03 (0.02)	-0.07 (0.03)
	0.5	-0.11 (0.04)	0.05 (0.02)	-0.08 (0.03)	-0.11 (0.04)	0.05 (0.02)	-0.08 (0.03)
	0.7	-0.06 (0.04)	-0.10 (0.02)	0.01 (0.03)	-0.10 (0.04)	0.05 (0.03)	-0.05 (0.03)
2	0.1	-0.03 (0.03)	-0.02 (0.02)	-0.05 (0.03)	-0.03 (0.03)	-0.02 (0.02)	-0.05 (0.03)
	0.3	-0.08 (0.04)	0.01 (0.02)	-0.06 (0.04)	-0.08 (0.04)	-0.02 (0.02)	-0.06 (0.04)
	0.5	-0.13 (0.04)	0.04 (0.02)	-0.13 (0.04)	-0.13 (0.04)	0.01 (0.02)	-0.13 (0.04)
	0.7	-0.02 (0.04)	0.06 (0.01)	-0.04 (0.03)	-0.04 (0.04)	-0.01 (0.02)	-0.04 (0.04)
3	0.1	-0.05 (0.04)	0.01 (0.03)	0.01 (0.02)	-0.05 (0.04)	0.02 (0.03)	0.01 (0.02)
	0.3	-0.01 (0.04)	-0.06 (0.02)	0.02 (0.03)	-0.06 (0.04)	0.04 (0.03)	0.01 (0.03)
	0.5	0.03 (0.03)	-0.23 (0.03)	0.06 (0.02)	-0.09 (0.04)	0.01 (0.04)	0.04 (0.02)
	0.7	0.08 (0.04)	-0.37 (0.02)	0.06 (0.03)	-0.11 (0.04)	0.03 (0.03)	0.03 (0.03)

the measurement of X is large, $\theta^M = 0.5, 0.7$. The OPEM provides nominal gains when the proportion of imprecision in the measurement of X is small, $\theta^M = 0.1, 0.3$. This is not unreasonable as the observed values, x^* are relatively similar to the unobserved values x when the imprecision is slight.

The first stage of the OPEM permits us to compare the effect of estimation using a two-stage OPEM for bias reduction versus a one-stage approach (standard implementation of the EM algorithm), providing that the one stage EM algorithm converges. We observe minimal reduction of bias in θ_0^C and θ_2^C , but noticeable bias reduction in θ_0^Y and θ_2^Y which are the stage two analogues of θ_0^C and θ_2^C . This finite sample simulation result, suggests that partitioning the EM algorithm into simpler steps may provide better bias reduction in the estimation of parameters of the outcome model.

5 Discussion

The K -stage orthogonally partitioned EM (OPEM) algorithm provides an implementation that exploits the common assumption of unique parameterization, corrects for biases due to imperfect data, converges for the specified model when standard implementation of the EM algorithm has a low probability of convergence, and reduces a potentially complex algorithm into a sequence of smaller, simpler, self-contained EM algorithms. We showed, using the theoretical framework of the EM algorithm, that the OPEM can be applied to any situation where the joint probability distribution can be written as a product of conditional probability distributions. As well, it yields an optimal solution over the parameter space, and known extensions of the EM algorithm for accelerating convergence may be used for each stage of the OPEM [40]. The computation of standard errors may be performed stage-wise with an appropriate adaptation of Louis' method for MCEM approaches [4, 38, 28] or through the adaptation of Oakes' formula for standard errors from an MCEM by Casella [41, Appendix A.3]. The finite sample simulation study revealed that partitioning the EM algorithm into simpler steps may provide better bias reduction in the estimation of parameters of the outcome model. This suggests that positioning the primary model at or near the end of the OPEM sequence is a strategy for better bias reduction in the estimators of the parameters of interest.

A potential application of this method is for the correction of bias due to mismeasured variables that are used to construct a Horvitz-Thompson [42] type weighting mechanism or propensity score for model correction [43, 44]. Additionally, this methodology could be adapted for related approaches, such as the inverse probability of treatment weighted marginal structural model estimator [45, 46, 47, 48]. Given that a likelihood approach could provide a unified method to correct for imperfect data, the EM algorithm is a natural tool to explore.

Beyond the potential application of the K -stage OPEM to a wide variety of Horvitz-Thompson type estimators, Section 3 strongly suggests broad applicability to complex, multi-model settings. The OPEM can be directly used for any situation with missing or mismeasured data where the joint distribution is expressed as a product of conditional, uniquely parameterized distributions. The ability to breakdown a complicated problem in to a series of simpler, more accessible problems will permit a broader implementation of the EM algorithm, permit the use of software packages that now implement and/or automate the EM algorithm for complex situations of imperfect data, and make the EM algorithm more accessible to a wider and more general audience.

A Appendix

A.1 Proof of Proposition 1

$\Rightarrow$ By the definition of unique parameterization, the direct sum $\theta^1 + \dots + \theta^p = 0$, when $\theta^k = 0$ for all k , $\theta^k \in \Theta^k$, $k=1, \dots, p$. Thus Θ can be written as a direct sum of the orthogonal subspaces.

$\Leftarrow$ Now, if

$$\Theta = \bigoplus_{j \in \{1, \dots, p\}} \Theta^j$$

then there exist p linear operators $H_1, \dots, H_p$ on Θ such that each H_k , $k = 1, \dots, p$, is a projection of Θ . Since each H_k is a projection of Θ , a direct-sum decomposition yields $H_k H_l = 0$ if $k \neq l$, where $k, l \in \{1, \dots, p\}$ ([49], section 6.6). Any vector $\theta^l \in R^{H_l}$, the range of H_l , is in the orthogonal complement of R^{H_k} . $\square$

A.2 Proof of Proposition 3

The algorithm begins with an initial value $\theta^{(0)} = \{\theta^{1(0)}, \dots, \theta^{K(0)}\}$. Applying the MCEM algorithm on the k th Q -function under the assumption that there exists a $t_k \geq 1$ such that $Q(\theta^{k(t_k+1)} | \theta^{k(t_k)}) \geq Q(\theta^{k(t_k)} | \theta^{k(t_k)})$, for all $\theta \in \Theta^k \subseteq \Theta$ where

$$\Theta^k \cap \bigoplus_{j \in \{1, \dots, K\}, j \neq k} \Theta^j = \emptyset$$

results in the k th Q -function being a Generalized EM algorithm. There are K MCEM algorithms for which maximizers are obtained. By induction, the entire optimization procedure is itself a Generalized EM algorithm. $\square$

References

1. Dempster A, Laird N, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 1977;39:1–38.
2. McLachlan GJ, Krishnan T. The EM algorithm and extensions. Hoboken, NJ: Wiley, 2008.
3. Liu C, Rubin DB, Yu YN. Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* 1998;85:755–70.
4. Louis TA. Finding the observed information matrix when using the EM algorithm. *J R Stat Soc Ser B* 1982;44:226–33.
5. Oakes D. Direct calculation of the information matrix via the EM algorithm. *J R Stat Soc Ser B* 1999;61:479–82.
6. Dawid AP, Skeme AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Appl Stat* 1979;28:20–8.
7. Drews CD, Flanders WD, Kosinski AS. Use of two data sources to estimate odds-ratios in case-control studies. *Epidemiology* 1993;4:327–35.
8. Ganguli B, Staudenmayer J, Wand MP. Additive models with predictors subject to measurement error. *Aust N Z J Stat* 2005;47:193–202.
9. Gustafson P. Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. New York: Chapman & Hall/CRC, 2004.
10. Buonaccorsi JP. Measurement error: models, methods, and applications. Boca Raton, FL: Chapman & Hall, 2010.
11. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* 2013;22:278–95.
12. Carpenter JR, Kenward MG, Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J R Stat Soc Ser A* 2006;169:571–84.
13. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Stat Methods Med Res* 2007;16:199–218.
14. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010;29:2920–31.
15. Guolo A. The SIMEX approach to measurement error correction in meta-analysis with baseline risk as covariate. *Stat Med* 2014;33:2062–76.
16. He W, Xiong J, Yi GY. SIMEX R Package for accelerated failure time. *J Stat Software* 2012.
17. Shang Y. Measurement error adjustment using the SIMEX method: an application to student growth percentiles. *J Educ Meas* 2012;49:446–65.
18. Buzas JS, Stefanski LA. Instrumental variable estimation in generalized linear measurement error models. *J Am Stat Assoc* 1996;91:999–1006.
19. Strand M, Sillau S, Grunwald GK, Rabinovitch N. Regression calibration for models with two predictor variables measured with error and their interaction, using instrumental variables and longitudinal data. *Stat Med* 2014;33:470–87.
20. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *Int J Epidemiol* 2006;35:1074–81.

21. Edwards JK, Cole SR, Troester MA, Richardson DB. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *Am J Epidemiol* 2013;177:904–12.
22. Meng X-L, Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 1993;80:267–78.
23. Tsiatis AA. Semiparametric theory and missing data. New York: Springer, 2006.
24. Chen W, Maitra R, Melnykov V (2013): EM algorithm for model-based clustering of finite mixture gaussian distribution, Version 0.2-4.
25. Carpenter J, Kenward M. Multiple imputation and its application. Chichester, West Sussex: John Wiley & Sons, 2013.
26. Little RJA, Rubin DB. Statistical analysis with missing data, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2002.
27. Wu CFJ. On the convergence properties of the EM algorithm. *Ann Stat* 1983;11:95–103.
28. Wei GCG, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J Am Stat Assoc* 1990;85:699–704.
29. Carroll R, Ruppert D, Stefanski L, Crainiceanu C. Measurement error in non-linear models: a modern perspective, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2006.
30. Rubin D. Inference and missing data. *Biometrika* 1976;63:581–92.
31. Regier MD, Moodie EEM, Platt RW. The effect of error-in-confounders on the estimation of the causal parameter when using marginal structural models and inverse probability-of-treatment weights: a simulation study. *Int J Biostat* 2014;10:1–15.
32. Chen J, Zhg D, Davidian M. A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics* 2002;3:347–60.
33. Ibrahim JG, Chen MH, Lipsitz SR. Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* 1999;55:591–6.
34. Lipsitz SR, Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. *Biometrika* 1996;83:916–22.
35. Liu W, Wu L. Simultaneous inference for semiparametric nonlinear mixed-effects models with covariate measurement errors and missing responses. *Biometrics* 2007;63:342–50.
36. Stubbendick A, Ibrahim JG. Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics* 2003;59:1140–50.
37. Gilks W, Wild P. Adaptive rejection sampling for Gibbs sampling. *Appl Stat* 1992;41:337–48.
38. Robert CP, Casella G. Monte Carlo statistical methods, 2nd edition. New York: Springer, 2004.
39. Ibrahim JG, Chen MH, Lipsitz SR. Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika* 2001;88:551–64.
40. Daigle B, Roh M, Petzold L, Niemi J. Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. *BMC Bioinformatics* 2012;13:1–18.
41. Casella G. Empirical Bayes Gibbs sampling. *Biostatistics* 2001;2:485–500.
42. Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 1952;47:663–85.
43. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004;23:2937–60.
44. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. *J Econometrics* 2007;141:1281–301.
45. Cole SR, Hudgens MG, Tien PC, Anastos K, Kingsley L, Chmiel JS, et al.. Marginal structural models for case-cohort study designs to estimate the association of antiretroviral therapy initiation with incident AIDS or death. *Am J Epidemiol* 2012;175:381–90.
46. Hernán MA, Hernández-Daz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15:615–25.
47. Moodie EEM, Stephens D. Marginal structural models: unbiased estimation for longitudinal studies. *Int J Public Health* 2011;56:117–19.
48. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–60.
49. Hoffman K, Kunze R. Linear algebra, 2nd ed. Upple Sadle River, NJ: Prentice Hall, 1971.
50. R Development Core Team (2014): R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>, ISBN 3-900051-07-0.