The Annals of Applied Statistics
2014, Vol. 8, No. 3, 1713–1727
DOI: 10.1214/14-AOAS752

ON ANALYSIS OF INCOMPLETE FIELD FAILURE DATA

BY ZHISHENG YE[1] AND HON KEUNG TONY NG[2]

National University of Singapore and Southern Methodist University

Many commercial products are sold with warranties and indirectly through dealers. The manufacturer-retailer distribution mechanism results in serious missing data problems in field return data, as the sales date for an unreturned unit is generally unknown to the manufacturer. This study considers a general setting for field failure data with unknown sales dates and a warranty limit. A stochastic expectation–maximization (SEM) algorithm is developed to estimate the distributions of the sales lag (time between shipment to a retailer and sale to a customer) and the lifetime of the product under study. Extensive simulations are used to evaluate the performance of the SEM algorithm and to compare with the imputation method proposed by Ghosh [*Ann. Appl. Stat.* **4** (2010) 1976–1999]. Three real examples illustrate the methodology proposed in this paper.

1. Introduction. Field failure data contain rich information about product reliability and the operating conditions in actual use. The information is important for risk assessment of field failures, early detection of unanticipated reliability problems [Wu and Meeker (2002)], and prediction of operation costs. Since many commercial products are sold with warranties, field failure data usually come from warranty claims. Alternatively, for noncommercial products such as military products, field data may be extracted from maintenance reports [Coit and Jin (2000)], and this type of data is called field maintenance data.

The rich information contained in field failure data can be extracted by careful data analysis. However, the analysis is difficult because field data are generally coarse and of poor quality. Compared with lab data that are collected under well-controlled testing conditions, field data are collected from customers and are contaminated by customer behaviors. For instance, the data are often contaminated with heterogeneous use conditions [Ye, Hong and Xie (2013)], dormant period after purchase [Wu (2012)], delayed report after failure [Kalbfleisch, Lawless and Robinson (1991)], customer rush near warranty expiration [Rai and Singh (2006)], the failed-but-not-reported problem [Xie and Liao (2013)], and systematic error on the cause and time of failures due to report error. To address these issues, a number

Received November 2013; revised May 2014.

[1] Supported by the National Research Foundation Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE).

[2] Supported by a Grant from the Simons Foundation (# 280601 to Tony Ng).

Key words and phrases. EM algorithm, maximum likelihood estimator, lifetime data, missing information.

of statistical models have been developed for warranty data analysis. See Blischke, Karim and Murthy (2011) and Wu (2012) for a comprehensive overview.

Another important cause of the coarse data is the missing sales date of unreturned units. Nowadays, many products are sold to customers through multiple channels of distribution instead of direct sale from the manufacturers. Under the manufacturer-retailer distribution mechanism, if a product fails within warranty, it will be returned to the manufacturer as a warranty claim. Then, the lifetime and the sales lag, which is the time between shipment to a retailer and sale to a customer, can be easily obtained from the warranty card. For an unreturned unit, however, the sales date is generally unknown unless the product is expensive (e.g., cars). The unit might still be in a retailer's warehouse or it might have been sold to a customer at some date unknown to the manufacturer. Ghosh (2010) presented such an example, where residential furnace components were shipped to retailers in batches and then sold to customers through retailers. Because of the retailers, the exact release time of a furnace to a customer was generally masked unless a furnace was sold and failed before a fixed end-of-study date.

A common approach to the unknown sales lag problem is to carry out a sensitivity analysis by assuming that the sales lag is fixed [Lawless (1998)]. Another method is to obtain the sales-lag distribution using survey or past experience, and then this distribution is incorporated into the data analysis to improve estimation accuracy [Hu and Lawless (1996), Wilson, Joyce and Lisay (2009)]. This method does not make full use of the database, as the sales date for returned units can be read from the warranty card and the sales-lag information is available from these returned units. Some studies treat both the observed sales-lag data and the observed lifetime as right censored so that the two types of data can be analyzed separately [Akbarov and Wu (2013), Ion et al. (2007), Karim (2008)]. Given that the sales date of an unreturned unit is unknown, however, the sales-lag data are not right censored, and the lifetime data are neither left truncated nor right censored. To get an accurate estimate, the sales-lag data and the lifetime data have to be jointly analyzed. In an interesting study, Ghosh (2010) analyzed field failure data with unknown sales lags. However, the inference procedure in that work is not efficient. In addition, it does not allow for a warranty limit and, thus, it is not applicable to warranty data. In addition, previous research assumes independent sales lag and lifetime. This assumption is true for some products, for example, light bulbs, televisions, computers, etc. For seasonal products such as heaters, fans, and air purifiers, the sales lag and the lifetime are correlated due to the usage pattern. For instance, a heater sold in summer will last longer than one sold in winter due to the uneven usage. It is also possible that a longer sales lag introduces more damage to the product [Akbarov and Wu (2013)]. Moreover, most research on field data analysis emphasizes the field failure time distribution only. The sales-lag information reflects customer demand rate and is important in manufacturing and inventory decisions.

In this paper, we consider joint parametric inference of sales-lag and lifetime in the presence of unknown sales dates. In contrast to the work by Ghosh (2010), we allow for a warranty limit as well as dependency between sales lag and lifetime. In addition, we propose a more efficient algorithm for statistical inference. Section 2 presents a simplified problem setting for field data with a warranty limit and an end-of-study date. Section 3 proposes an inference framework based on the stochastic expectation–maximization (SEM) algorithm. Section 4 discusses how the SEM algorithm can be modified to handle more general situations. In Section 5 a simulation study examines the performance of the proposed algorithm, and we compare it with the imputation method proposed by Ghosh (2010). The proposed algorithm is demonstrated using three examples with different missing data patterns in Section 6. A concise conclusion is provided in Section 7.

2. Problem statement. Suppose that N identical units are produced in a batch and delivered to several retailers at the same time. The delivery time is set as the time origin in the analysis. These units are then sold to customers with a warranty of length τ, starting from the date of purchase. Let X be the sales lag (same as the sales date in this setting) and T the lifetime of the product from the date of sale, where both X and T are random. Let $\mathcal{T}_0$ be a fixed end-of-study date, which can be viewed as the date the analysis is performed. If a unit fails before $\mathcal{T}_0$ and is within warranty, we assume that a warranty claim is made to the manufacturer without delay. Then both the sales date X and the lifetime T are known to us. Otherwise, the sales date and the product lifetime are unavailable. Suppose that before $\mathcal{T}_0$, we observe C claims, and so we have C realizations of (X, T), denoted as (x_i, t_i), $i = 1, 2, \dots, C$. For the remaining $N - C$ units, the values of (X, T) are missing.

This study focuses on parametric inference. Denote the joint probability density function (PDF) of (X, T) as $f_{X,T}(x, t)$ and the joint cumulative distribution function (CDF) as $F_{X,T}(x, t)$, where $x, t > 0$. Let Θ be the parameter vector. Given the observed data (x_i, t_i), $i = 1, 2, \dots, C$, the likelihood function of Θ is given by

$$\mathcal{L}(\Theta) = [1 - \Pr(X + T < \mathcal{T}_0, T < \tau)]^{N-C} \prod_{i=1}^{C} f_{X,T}(x_i, t_i), \tag{2.1}$$

where $\Pr(X + T < \mathcal{T}_0, T < \tau)$ is the probability that a unit fails within warranty and is observed within $\mathcal{T}_0$. This probability can be written as

$$\Pr(X + T < \mathcal{T}_0, T < \tau) = \int_0^{\tau} \int_0^{\mathcal{T}_0 - t} f_{X,T}(x, t)\, dx\, dt.$$

If X and T are independent, this probability simplifies to

$$\Pr(X + T < \mathcal{T}_0, T < \tau) = \int_0^{\tau} F_X(\mathcal{T}_0 - t)\, dF_T(t).$$

In principle, the maximum likelihood estimator (MLE) of Θ can be obtained from direct maximization of the likelihood (2.1). Nevertheless, numerical evaluation of the integral would introduce computation error, which is magnified by the factor $N - C$ in (2.1). Due to the high missing data rate in our problem (i.e., large N and small C), the total computation error is significant, and the likelihood is flat near the maximum. These two factors lead to unstable estimates if direct maximization is used (i.e., convergence to values far from the optimal or failure to converge). The instability is observed in our simulation study (see Section 5) and Ghosh (2010). Therefore, alternative techniques are needed. In the next section we propose an efficient and easy-to-implement procedure based on the SEM algorithm.

3. The stochastic expectation–maximization framework.

3.1. *The SEM algorithm.* The EM algorithm is an iterative procedure that repeatedly fills the missing data in the complete-data log-likelihood with their conditional expected values (E-step) and maximizes the complete data log-likelihood to update the parameter estimates (M-step). The EM algorithm is efficient in finding the MLEs when computation of the expectation and the maximization are easy to perform. See McLachlan and Krishnan (2008) for a book-length account. Unfortunately, the E-step is intractable when the EM algorithm is applied to the problem in Section 2. Alternatively, the expectation can be approximated through Monte Carlo simulation, leading to the Monte Carlo EM (MCEM) algorithm. In our problem, the approximation error of the expectation leads to a breakdown of the MCEM algorithm because the likelihood is flat near the maximum.

The difficulty in executing the E-step can be efficiently addressed by the SEM algorithm proposed by Celeux and Diebolt (1985). The SEM algorithm replaces the E-step with a stochastic step (S-step), which is easy to implement as long as the missing data are easy to impute. Compared with the MCEM algorithm, the SEM algorithm completes the observed sample by replacing each missing datum with a value randomly drawn from the distribution conditional on results from the previous step. The SEM algorithm has been shown to be computationally less burdensome than the MCEM algorithm. Because of the stochastic nature, it is free of the saddle point problem, a serious problem for the EM algorithm [Bordes, Chauveau and Vandekerkhove (2007), Cariou and Chehdi (2008)]. It was shown by Diebolt and Celeux (1993), Chauveau (1995) and Nielsen (2000) that under suitable regularity conditions the SEM estimators are efficient in the sense that the variance approaches the Cramér–Rao lower bound. Some applications of the algorithm suggest that it is insensitive to starting values and performs well for small or moderate sample sizes. See, for example, Chauveau (1995), Cariou and Chehdi (2008), and Svensson and Sjöstedt-de Luna (2010).

3.2. *Implementation.* Let $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ be the sets of observed and missing data, respectively. Here, Ω includes the C observed values of (x_i, t_i) and the information that $N - C$ observations are missing. Given the parameter values $\Theta^{(k)}$ of Θ from the kth SEM cycle, the $(k + 1)$st cycle for the problem described in Section 2 evolves as follows:

S-step. Draw a random sample $\mathbf{\Gamma}^{(k)} = \{(x_j^{(k)}, y_j^{(k)}); j = 1, 2, \ldots, N - C\}$ from the conditional distribution of $\{\mathbf{\Gamma} | \mathbf{\Omega}, \Theta^{(k)}\}$ to update the pseudo Q-function

$$Q(\Theta; \mathbf{\Omega}, \mathbf{\Gamma}) = \sum_{i=1}^{C} \ln f_{X,T}(x_i, t_i) + \sum_{j=1}^{N-C} \ln f_{X,T}\bigl(x_j^{(k)}, t_j^{(k)}\bigr). \tag{3.1}$$

M-step. Maximize the pseudo Q-function (3.1), which is a complete data log-likelihood, to obtain $\Theta^{(k+1)}$ for the next cycle.

The M-step deals with a complete-data log-likelihood. It is easy to implement through direct optimization or with the help of statistical software if some common distributions are used for X and T, for example, independent exponential, Weibull, or bivariate lognormal. Under suitable regularity conditions, the sequence $\Theta^{(k)}$ converges to a random variable whose mean is an asymptotically efficient estimator of Θ. These conditions typically are satisfied if the complete data model and the missing data model are sufficiently smooth [Nielsen (2000), Section 2.3]. The simulation results in Section 5 support this argument for commonly used lifetime distributions. To obtain an estimate of Θ, we run the SEM algorithm to obtain $\Theta^{(k)}, k = 1, 2, \ldots, K$, discard the first few iterations for burn-in, and average over the estimates from the remaining iterations to get $\hat{\Theta}$. According to some reports [e.g., Marschner (2001)] as well as our experience, a burn-in period of 100 cycles is long enough under moderate missing data rates, while an additional 1000 iterations are sufficient to estimate Θ. Nevertheless, we suggest a trace plot of the $\{\Theta^{(k)}\}$ sequence versus the iterations for checking the sufficiency of the burn-in, and determining a more appropriate burn-in duration, if necessary.

There are several ways to impute the missing data in the S-step. The standard method is based on the conditional distribution of the unobserved (X, T), which is

$$g_{X,T}(x, t) = \frac{f_{X,T}(x, t)}{1 - \Pr(X + T < \mathcal{T}_0, T < \tau)}\bigl(1 - I\{x + t < \mathcal{T}_0, t < \tau\}\bigr), \tag{3.2}$$

where $I\{\cdot\}$ is the indicator function. Direct sampling from this conditional PDF is difficult. We might resort to the Markov chain Monte Carlo (MCMC) method. However, it is inefficient to imbed an iterative algorithm (MCMC) into another one (SEM). Due to the extremely high missing data rate in our problem, we impute missing data in a natural way, which is somewhat brute force, yet very straightforward, easy to implement, and efficient.

Recall that a unit is observed only when $X + T < \mathcal{T}_0$ and $T < \tau$, while the probability of being observed is typically low. This motivates us to impute the missing data $\mathbf{\Gamma}^{(k)}$ by using a simple acceptance-rejection method: an imputation (x, t) from $f_{X,T}(x, t|\Theta^{(k)})$ is rejected only when $x + t < \mathcal{T}_0$ and $t < \tau$. It can be easily shown that (X, T) imputed from this sampling scheme follows the distribution given in (3.2). To use this imputation scheme, a starting point $\Theta^{(0)}$ that leads to a large mean value of X or T is strongly recommended in order to avoid a high rejection rate at the outset of the SEM algorithm. According to our comprehensive simulation trials, this scheme is very efficient because the missing data rate, which approximately equals 1 minus the rejection rate, is high in our setting. The rejection rate should be low as long as $\Theta^{(k)}$ is not too far away from the true value. Therefore, the brute-force imputation is expected to be effective in the sense that the computational time for each SEM iteration is relatively small.

3.3. *Confidence intervals.* The log-likelihood based on full data $\mathbf{D} = \mathbf{\Omega} \cup \mathbf{\Gamma}$ is the same as (3.1). Because of the simple structure of the full data likelihood, the score function and the observed information matrix based on full data can be easily obtained by taking the first and second derivatives of (3.1) with respect to the parameters Θ. Denote the first and the negative of the second derivatives as $S(\Theta, \mathbf{D})$ and $B(\Theta, \mathbf{D})$, respectively. The observed information matrix based on incomplete data can be computed based on the missing information principle [Louis (1982)] as

$$\mathcal{I}(\Theta) = E[B(\Theta, \mathbf{D})|\mathbf{\Omega}] - E[S^2(\Theta, \mathbf{D})|\mathbf{\Omega}] + \{E[S(\Theta, \mathbf{D})|\mathbf{\Omega}]\}^2, \tag{3.3}$$

where $v^2 = v \cdot v'$ when v is an $m \times 1$ vector. To evaluate (3.3), we first impute M samples $\mathbf{\Gamma}^{(i)}, i = 1, 2, \dots, M$, for the missing data $\mathbf{\Gamma}$ conditional on the observed data and Θ. Let $\mathbf{D}^{(i)} = \mathbf{\Omega} \cup \mathbf{\Gamma}^{(i)}$. Then, the incomplete data information matrix can be approximated by [Wei and Tanner (1990)]

$$\begin{aligned}\hat{\mathcal{I}}(\Theta) \doteq {} & \frac{1}{M}\sum_{i=1}^{M} B(\Theta, \mathbf{D}^{(i)}) - \frac{1}{M}\sum_{i=1}^{M}[S(\Theta, \mathbf{D}^{(i)})]^2 \\ & + \left[\frac{1}{M}\sum_{i=1}^{M} S(\Theta, \mathbf{D}^{(i)})\right]^2.\end{aligned} \tag{3.4}$$

The SEM estimate, $\hat{\Theta}$, is plugged into (3.4) to obtain $\hat{\mathcal{I}}(\hat{\Theta})$, which is then used to obtain the asymptotic variances of $\hat{\Theta}$ as well as the confidence intervals. To ensure the accuracy of the simulation approximation, the number of samples M should be carefully chosen. The magnitude depends on the missing data rate.

4. Some further considerations. Usually, products are manufactured and shipped to retailers intermittently, meaning that the shipment dates for distinct units may differ. Under this circumstance, we can still observe (X, T) for a returned unit. For an unreturned unit, we can subtract the date of shipment from the end-of-study date to obtain the censored time for $X + T$. Then, the framework discussed in Section 3 applies.

In some situations, direct sale from the manufacturer is possible. The sales dates for units sold directly to customers are available in the database. The data do not have sales lag and the lifetimes are simply right censored. The contribution of an observed unit to the likelihood is exactly the PDF of T, while if a unit is censored, say, at time $\mathcal{T}_c$, the missing value can be easily imputed in the S-step as $t = F_T^{-1}(u + (1-u)F_T(\mathcal{T}_c|\Theta^{(k)}))$, where $F_T^{-1}(\cdot)$ is the quantile function of $F_T(\cdot)$, while u is a random draw from the uniform distribution on (0, 1).

Wilson, Joyce and Lisay (2009) considered a nonnegligible report delay after failure (denoted as Y) in addition to the sales lag X. When information about Y for a returned unit is available, we can work on the random vector (X, Y, T). In the S-step, the missing (X, Y, T) can be imputed similar to the acceptance-rejection method discussed in Section 3.2, after which the pseudo Q-function can be easily specified. The M-step can be implemented based on standard estimation procedures established for complete multivariate data. Analysis of such data will be demonstrated in Section 6.3.

5. Simulation study. In the simulation the number of units in a batch is assumed to be $N = 200$. Both dependent and independent (X, T) are examined. We first assume a bivariate lognormal distribution for (X, T):

$$(\ln X, \ln T) \sim \mathcal{N}\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right).$$

The biases and root mean square errors (RMSEs) of the SEM estimators under different parameter values and different combinations of $(\tau, \mathcal{T}_0)$ are estimated using 5000 MC replications, as shown in Table 1. We then consider independent T and X, each conforming to either an exponential distribution or a Weibull distribution. Different settings have been examined. The estimated biases and RMSEs are presented in Table 2. Code in Matlab® is presented in the supplementary materials [Ye and Ng (2014)]. From Tables 1 and 2, we can see that the SEM algorithm effectively estimates the model parameters in both dependent and independent cases. We can also observe that, on average, a longer warranty period leads to higher accuracy of the estimator. This observation agrees with our intuition as the missing data rate decreases with τ.

TABLE 1

Estimated biases and RMSEs of the SEM estimator when $(\ln X, \ln T) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with the consideration of a warranty period* (τ) *and a batch size* $N = 200$

			True values							
Scenario	τ	$\mathcal{T}_0$	μ	Σ		μ_1	μ_2	σ_{11}	σ_{22}	σ_{12}
S1	4	6	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$	Bias ($\times 10^2$)	−0.22	−0.21	1.51	1.19	−2.57
					RMSE ($\times 10$)	1.68	1.61	2.27	2.19	1.47
S2	3	4	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$	Bias ($\times 10^2$)	−2.97	−4.40	15.48	10.33	−17.94
					RMSE ($\times 10$)	3.16	2.96	3.82	3.55	3.11
S3	3	4	$\begin{pmatrix} 1 \\ 1.3 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$	Bias ($\times 10^2$)	−0.50	−3.03	6.13	1.36	−6.24
					RMSE ($\times 10$)	3.98	3.34	4.15	4.10	3.15

TABLE 2

Estimated biases and RMSEs of the parameter estimates obtained from the SEM algorithm when (X, T) are independent with the consideration of a warranty period (τ)

Setting			$\boldsymbol{\lambda}$		$\boldsymbol{\theta}$		$\boldsymbol{\beta}$	
τ	$\mathcal{T}_0$	(λ, θ, β)	**Bias (×10^2)**	**RMSE (×10)**	**Bias (×10^2)**	**RMSE (×10)**	**Bias (×10^2)**	**RMSE (×10)**
				$X \sim \text{Exp}(\lambda)$, $T \sim \text{Weibull}(\theta, \beta)$				
5	6	(0.7, 5, 2)	−0.02	1.17	−4.90	3.52	5.82	2.03
3	4	(0.7, 5, 2)	−5.79	2.71	−42.62	10.18	20.45	4.45
4	6	(0.7, 5, 2)	−0.06	1.25	−5.74	3.81	6.60	2.29
				$X \sim \text{Weibull}(\theta, \beta)$, $T \sim \text{Exp}(\lambda)$				
5	6	(0.5, 4, 1.5)	−0.03	0.77	−4.78	3.39	4.76	1.54
3	4	(0.5, 4, 1.5)	−0.89	1.46	−20.12	7.16	10.80	2.26
4	6	(0.5, 4, 1.5)	0.02	0.81	−4.11	3.84	4.28	1.53

Setting			$\boldsymbol{\lambda}$		$\boldsymbol{\delta}$	
τ	$\mathcal{T}_0$	(λ, δ)	**Bias (×10^2)**	**RMSE (×10)**	**Bias (×10^2)**	**RMSE (×10)**
			$X \sim \text{Exp}(\lambda)$, $T \sim \text{Exp}(\delta)$			
5	6	(0.2, 0.2)	1.47	0.63	1.57	0.63
4	5	(0.2, 0.2)	2.40	0.87	2.61	0.87
5	6	(0.5, 0.2)	0.74	0.82	0.50	0.26
3	4	(0.5, 0.2)	1.31	1.41	1.91	0.57
5	6	(0.4, 0.7)	0.41	0.37	0.68	0.72
3	4	(0.4, 0.7)	1.11	0.57	1.04	1.07

The proportional imputation method proposed by Ghosh (2010) does not allow for a warranty limit and it can only handle independent X and T. In order to compare the SEM algorithm with it, we let X and T be independent and $\tau > \mathcal{T}_0$ (i.e., no warranty consideration). The biases and RMSEs of the estimators computed from the proportional imputation approach and the SEM algorithm are presented in Table 3. The SEM estimator has much smaller biases and RMSEs. A possible explanation is that the stratified sampling scheme in the proportional imputation algorithm might introduce biases in the imputing samples. Another finding from our comparative study is that the computation time required by the SEM algorithm is much shorter compared to that of the proportional imputation algorithm. Overall, the SEM algorithm is statistically and computationally more efficient than the imputation method. More importantly, the SEM algorithm is able to handle a more general scenario with a warranty limit and dependent X and T. It also allows for construction of confidence intervals for the parameters. These advantages make the SEM algorithm attractive for the problem.

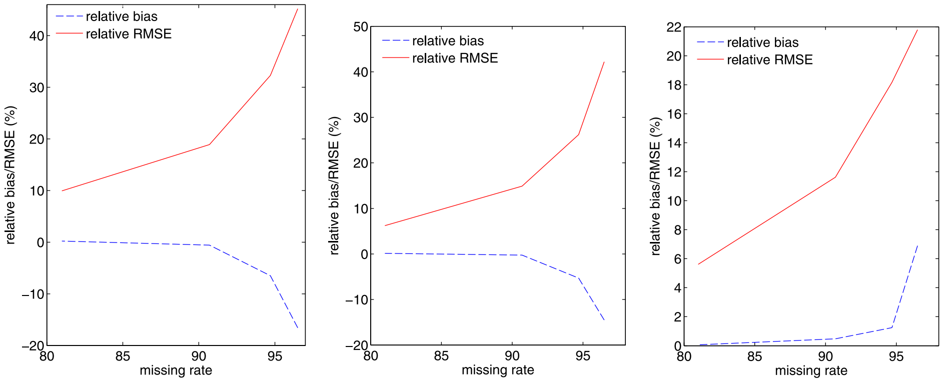
To demonstrate the advantage of the SEM algorithm over direct optimization, further simulation is conducted by assuming $X \sim \text{Exp}(\lambda = 0.7)$, $T \sim$

TABLE 3
The estimated biases and RMSEs of Ghosh's estimators [Ghosh (2010)] *and the SEM estimators*: (X, T) *are independent and* $\tau = \infty$

Setting		**Bias ($\times 10^2$)**		**RMSE ($\times 10$)**	
$(\mathcal{T}_0, \lambda, \theta, \beta)$		**Impute**	**SEM**	**Impute**	**SEM**
		$X \sim \text{Exp}(\lambda)$, $T \sim \text{Weibull}(\theta, \beta)$			
(6, 0.7, 5, 2)	λ	−4.54	−0.22	1.13	1.09
	θ	−14.95	−1.35	3.83	3.50
	β	8.46	3.30	2.16	1.97
(4, 0.7, 5, 2)	λ	−16.65	−2.29	2.54	2.33
	θ	−69.54	−9.81	10.21	8.64
	β	21.61	8.75	3.73	3.32
		$X \sim \text{Weibull}(\theta, \beta)$, $T \sim \text{Exp}(\lambda)$			
(6, 0.5, 4, 1.5)	λ	−2.93	−0.01	0.60	0.75
	θ	−21.39	−0.75	2.87	3.80
	β	2.47	2.59	1.36	1.44
(4, 0.5, 4, 1.5)	λ	−8.53	−0.94	3.13	1.50
	θ	−47.19	−3.41	26.86	8.33
	β	5.31	6.62	3.91	2.29

Setting		**Bias ($\times 10^2$)**		**RMSE ($\times 10$)**	
$(\mathcal{T}_0, \lambda, \theta, \delta)$		**Impute**	**SEM**	**Impute**	**SEM**
		$X \sim \text{Exp}(\lambda)$, $T \sim \text{Exp}(\delta)$			
(5, 0.2, 0.2)	λ	−3.45	0.58	1.59	0.70
	δ	−2.09	1.83	0.94	0.76
(4, 0.4, 0.7)	λ	−3.52	0.44	1.36	0.51
	δ	−2.64	0.42	1.19	1.01

Weibull$(\theta, \beta = 2)$, and $N = 2000$. Different missing data rates are achieved by varying θ. We find that direct maximization breaks down very quickly (i.e., fails to converge) when the missing data rate is high, say, >80%. On the other hand, the SEM algorithm performs well under much larger missing data rates. The relative biases (bias ÷ true value) and relative RMSEs (RMSE ÷ true value) of the SEM estimators are computed from 1000 MC replications, as shown in Figure 1. When the missing data rate is extremely high, say, 97% in Figure 1, the RMSE for θ is large, which can be seen as a breakdown of the SEM algorithm. For a fixed missing data rate, nevertheless, the bias and RMSE can be significantly reduced if the sample size N is increased. For illustration, given the missing data rate of 94.7%, the respective relative biases (RMSEs) for λ, θ and β decrease from −6.5% (32.2%), −5.3% (26.1%), 1.25% (18.2%) to −2.0% (15.1%), −1.8% (13.2)%, 0.15% (6.0%), respectively, when N is increased to 20,000.



(a) (b) (c)

FIG. 1. *Relative biases and relative RMSEs of the SEM estimates under the Exp–Weibull setting.* (a) *Is for* λ, (b) *is for* θ, *and* (c) *is for* β.

6. Examples. The developed algorithm is applied to three real data sets with different missing data patterns. The first example is from an industrial firm that produces residential furnace components [Ghosh (2010)]. There is an unobserved sales lag for an unreturned unit but there is no warranty limit. The second example comes from warranty claims for an automobile component with both a sales lag and a warranty limit. The third example concerns a telecommunications product [Wilson, Joyce and Lisay (2009)] where both the sales lag and report delay exist. The times in these examples are in months. These data are presented in the supplementary materials [Ye and Ng (2014)].

6.1. *Installation failure data of a furnace.* This data set is from an industrial firm producing residential furnace components during one week in May 2001. It consists of $N = 400$ furnace components and $C = 133$ returns, denoted as (x_i, t_i) for $i = 1, \dots, C$. The components are sold with life warranty, that is, $\tau = \infty$. In keeping with Ghosh (2010), suppose the sales lag is exponential, $X \sim \text{Exp}(\lambda)$, and the failure time is Weibull, $T \sim \text{Weibull}(\theta, \beta)$. Ghosh (2010) obtained estimates of the model parameters as $\hat{\lambda} = 0.57$, $\hat{\theta} = 14.47$, and $\hat{\beta} = 0.81$ by using his imputation algorithm. Here, we reanalyze the data using the SEM algorithm. We use 100 iterations for burn-in and another 900 iterations to obtain the SEM estimates. The evolution paths of the parameters are shown in Figure 2. The paths reveal no obvious trend in the simulation. The computation time for the SEM algorithm is 12.28 seconds on a laptop with an Intel® Core i5 CPU, which is faster than that required by the proportional imputation method (72.12 seconds on the same computer). We then invoke the procedure in Section 3.3 to compute the information matrix and thus the standard deviations of the estimators. To ensure an accurate approximation for the information matrix, we use $M = 100{,}000$ imputa-

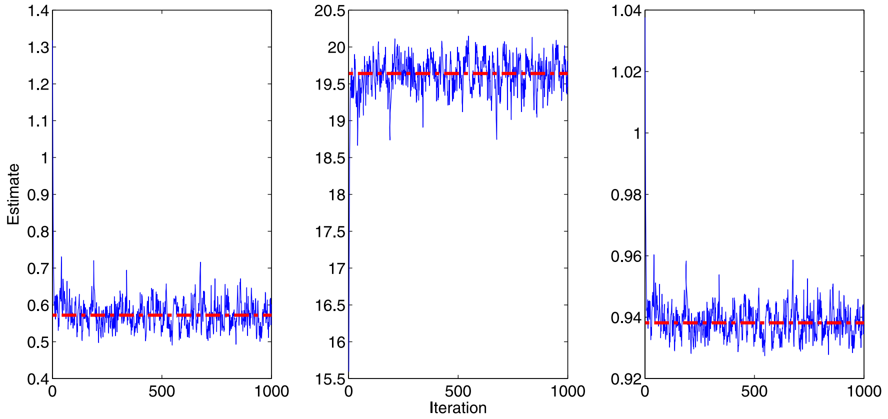


FIG. 2. *Parameter evolutions in the SEM algorithm when there is no warranty: the dashed-dotted line represents the average of the last* 900 *iterations.*

tions in (3.4). The estimates (standard errors) of the model parameters are $\hat{\lambda} = 0.57$ (0.053), $\hat{\theta} = 19.59$ (2.420), and $\hat{\beta} = 0.95$ (0.078), respectively.

6.2. *Warranty data for an automobile component.* The data analyzed here are warranty claims for a specific automobile component produced over a three-year period. The component is sold with an 18-month warranty. When a component fails within warranty and is returned as a claim, the date of manufacture, date of sale, date of claim, failure mode, and some other related information are recorded. The end-of-study date for this study is $\mathcal{T}_0 = 54$ months. We focus on the 589 components manufactured in the first month (month 0) of the production. During the observation window, 66 claims were observed.

Based on previous experience, we use a lognormal distribution for the sales lag and Weibull for the lifetime, that is, $X \sim \ln\mathcal{N}(\mu, \sigma)$ and $T \sim \text{Weibull}(\theta, \beta)$. To ensure convergence of the SEM algorithm, 100,000 iterations are used. The running time is about 10 minutes. The evolution paths of the parameter estimates versus the SEM iterations are presented in the supplementary materials [Ye and Ng (2014)]. The estimates (standard errors) of the four model parameters are $\hat{\mu} = 1.66$ (0.107), $\hat{\sigma} = 0.84$ (0.081), $\hat{\theta} = 59.5$ (10.0), and $\hat{\beta} = 1.79$ (0.224), respectively. The estimated lifetime distribution and the corresponding 95% pointwise confidence band are depicted in Figure 3. One can also obtain estimates of reliability characteristics [e.g., mean time to failure (MTTF), quantiles, etc.], which are useful in improving product reliability as well as determining the optimal warranty period.

To check the parametric model assumption, we consider different combinations of the distributions for (X, T). The log-likelihood at the estimated values of the model parameters and the AIC are presented in Table 4. A lognormal distribution for the sales lag and a Weibull distribution for the component lifetime seems reasonable.

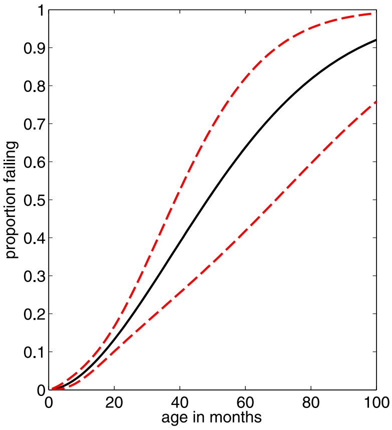


FIG. 3. *Estimated CDF and* 95% *pointwise confidence band for the failure time* T *for the automobile component.*

6.3. *Field data for a telecommunications product.* Wilson, Joyce and Lisay (2009) reported field failure data for a product installed in a telecommunications network. The data consist of 1838 units in total, out of which 26 units were returned within $\mathcal{T}_0 = 18$ months after the shipment. The failure data are grouped by month so that we only observe the number of failures for each month. All the remaining 1812 units are missing and the missing proportion is about 98.6%. In this data set the recorded time for each of the 26 returned units is the time in between the unit being shipped and being returned for repair. The recorded time includes the sales lag X, failure time T, and report delay Y. This means that a failure is recorded only when $X + T + Y < \mathcal{T}_0$. But if a failure is recorded, we only observe $X + T + Y$. In order to decouple these three random variables, Wilson, Joyce and Lisay (2009) collected additional sales-lag data and report-delay data from an old product in the same family, for which the sales lag and the return delay are assumed to be the same as the product under study. In total, there are 100 extra installation-lag data and 100 extra report-delay data records. These two data sets are also interval

TABLE 4
Values of the likelihood and Akaike's information criterion (*AIC*) *under different parametric models in Example* 6.2

Model	$X \sim$ **Exp,** $T \sim$ **Weibull**	$X \sim$ **Weibull,** $T \sim$ **Weibull**	$(X, T) \sim$ **Bivariate Logn**	$X \sim$ **Logn,** $T \sim$ **Weibull**
No. of parameter	3	4	5	4
Likelihood	−586.5	−584.2	−578.1	−578.5
AIC	1179.0	1176.4	1166.2	1165.0

censored and grouped by month. More details about the data can be found in the original paper.

Wilson, Joyce and Lisay (2009) pointed out that direct maximum likelihood estimation is difficult. They developed a Bayesian inference procedure to fit the data. The Gibbs sampling was adopted to resemble the posterior distribution. Here, we apply the SEM algorithm. Following Wilson, Joyce and Lisay (2009), we assume $X \sim \text{Gamma}(k_1, \lambda_1)$, where k_1 is the shape parameter and λ_1 is the scale parameter. The failure time is assumed to be Weibull, that is, $T \sim \text{Weibull}(\theta, \beta)$, and the report delay is gamma, that is, $Y \sim \text{Gamma}(k_2, \lambda_2)$. We fit the additional 100 installation-lag data and the additional 100 report-delay data to obtain an initial estimate of k_1, λ_1 and k_2, λ_2. These values are used as initial values for the SEM algorithm. In the SEM iterations, we impute the missing X, T, and Y based on the fact that X and Y in the additional data sets and $X + Y + T$ in the original data set are interval censored or right censored. The imputation can be done by the acceptance-rejection method with acceptance only when the imputed value falls inside the desired interval. Since the missing proportion is high, we use 100,000 iterations in the SEM algorithm. The first 10,000 iterations are discarded for burn-in purposes and the remaining 90,000 iterations are averaged for estimation. The parameter estimates are $\hat{k}_1 = 2.264$ (0.40), $\hat{\lambda}_1 = 1.714$ (0.35), $\hat{\theta} = 720.7$ (683), $\hat{\beta} = 1.153$ (0.37), $\hat{k}_2 = 2.779$ (0.49), and $\hat{\lambda}_2 = 0.080$ (0.015). The evolution paths of the six parameters are presented in the supplementary materials [Ye and Ng (2014)]. The traces for the parameters related to the lifetime T are very unstable. This can be viewed as an indication of large bias/variance in the estimation or an indicator of insufficient information for T, which might lead to the breakdown of SEM. With these results, one can decide whether a longer observation window is needed. In summary, the SEM algorithm serves well as a tool for checking whether there is sufficient information for inference.

7. Conclusions. The common problem of unknown sales dates in field failure data has posed a challenge. Direct maximization of the likelihood is difficult due to the excessive flatness of the likelihood and numerical error when evaluating the function. We have proposed an SEM framework for parametric inference. The algorithm allows for a warranty limit and possible dependence between the sales lag and the product lifetime. It is easy to implement and computationally efficient. Our examples with different missing data patterns demonstrate the flexibility of the proposed framework.

Acknowledgments. The authors are grateful to the Editor and three anonymous reviewers for constructive comments which have led to a substantial improvement to an earlier version of the paper. They would like to thank Dr. S. Wilson and Dr. T. Joyce for sharing the telecommunications product data. They are also grateful to Dr. R. Karim for the discussion of sales-lag problems in warranty data analysis and Dr. W. A. Woodward for comments on the paper.

SUPPLEMENTARY MATERIAL

Additional discussions, graphs, Matlab code, and data (DOI: 10.1214/14-AOAS752SUPP; .pdf). We provide additional discussions on the effect of model misspecification and evolution paths of parameter estimates in SEM for Sections 6.2 and 6.3. We also provide the Matlab code for simulation and the data used in the examples.

REFERENCES

AKBAROV, A. and WU, S. (2013). Warranty claims data analysis considering sales delay. *Qual. Reliab. Eng. Int.* **29** 113–123.

BLISCHKE, W., KARIM, M. and MURTHY, D. (2011). *Warranty Data Collection and Analysis*. Springer, Berlin.

BORDES, L., CHAUVEAU, D. and VANDEKERKHOVE, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Comput. Statist. Data Anal.* **51** 5429–5443. MR2370882

CARIOU, C. and CHEHDI, K. (2008). Unsupervised texture segmentation/classification using 2-d autoregressive modeling and the stochastic expectation–maximization algorithm. *Pattern Recogn. Lett.* **29** 905–917.

CELEUX, G. and DIEBOLT, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* **2** 73–82.

CHAUVEAU, D. (1995). A stochastic EM algorithm for mixtures with censored data. *J. Statist. Plann. Inference* **46** 1–25. MR1342674

COIT, D. W. and JIN, T. (2000). Gamma distribution parameter estimation for field reliability data with missing failure times. *IIE Trans.* **32** 1161–1166.

DIEBOLT, J. and CELEUX, G. (1993). Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Comm. Statist. Stochastic Models* **9** 599–613. MR1249140

GHOSH, S. (2010). An imputation-based approach for parameter estimation in the presence of ambiguous censoring with application in industrial supply chain. *Ann. Appl. Stat.* **4** 1976–1999. MR2829943

HU, X. J. and LAWLESS, J. F. (1996). Estimation of rate and mean functions from truncated recurrent event data. *J. Amer. Statist. Assoc.* **91** 300–310. MR1394085

ION, R. A., PETKOVA, V. T., PEETERS, B. H. and SANDER, P. C. (2007). Field reliability prediction in consumer electronics using warranty data. *Qual. Reliab. Eng. Int.* **23** 401–414.

KALBFLEISCH, J. D., LAWLESS, J. F. and ROBINSON, J. A. (1991). Methods for the analysis and prediction of warranty claims. *Technometrics* **33** 273–285.

KARIM, M. R. (2008). Modelling sales lag and reliability of an automobile component from warranty database. *Int. J. Reliab. Qual. Saf. Eng.* **2** 234–247.

LAWLESS, J. F. (1998). Statistical analysis of product warranty data. *Int. Stat. Rev.* **66** 41–60.

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** 226–233. MR0676213

MARSCHNER, I. C. (2001). On stochastic versions of the EM algorithm. *Biometrika* **88** 281–286. MR1841275

MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. Wiley, Hoboken, NJ. MR2392878

NIELSEN, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli* **6** 457–489. MR1762556

RAI, B. and SINGH, N. (2006). Customer-rush near warranty expiration limit, and nonparametric hazard rate estimation from known mileage accumulation rates. *IEEE Trans. Reliab.* **55** 480–489.

SVENSSON, I. and SJÖSTEDT-DE LUNA, S. (2010). Asymptotic properties of a stochastic EM algorithm for mixtures with censored data. *J. Statist. Plann. Inference* **140** 111–127. MR2568126

WEI, G. C. G. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

WILSON, S., JOYCE, T. and LISAY, E. (2009). Reliability estimation from field return data. *Lifetime Data Anal.* **15** 397–410. MR2519721

WU, S. (2012). Warranty data analysis: A review. *Qual. Reliab. Eng. Int.* **28** 795–805.

WU, H. and MEEKER, W. Q. (2002). Early detection of reliability problems using information from warranty databases. *Technometrics* **44** 120–133. MR1951722

XIE, W. and LIAO, H. (2013). Some aspects in estimating warranty and post-warranty repair demands. *Naval Res. Logist.* **60** 499–511. MR3100810

YE, Z.-S., HONG, Y. and XIE, Y. (2013). How do heterogeneities in operating environments affect field failure predictions and test planning? *Ann. Appl. Stat.* **7** 2249–2271. MR3161721

YE, Z.-S. and NG, H. K. T. (2014). Supplement to "On analysis of incomplete field failure data." DOI:10.1214/14-AOAS752SUPP.

DEPARTMENT OF INDUSTRIAL
AND SYSTEM ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE
1 ENGINEERING DRIVE 2
SINGAPORE 117576
REPUBLIC OF SINGAPORE
E-MAIL: yez@nus.edu.sg

DEPARTMENT OF STATISTICAL SCIENCE
SOUTHERN METHODIST UNIVERSITY
DALLAS, TEXAS 75275-0332
USA
E-MAIL: ngh@mail.smu.edu