Quality Engineering

ISSN: 0898-2112 (Print) 1532-4222 (Online) Journal homepage: https://www.tandfonline.com/loi/lqen20

Rejoinder

Sumin Shen, Huiying Mao & Xinwei Deng

To cite this article: Sumin Shen, Huiying Mao & Xinwei Deng (2019) Rejoinder, Quality Engineering, 31:3, 516-521, DOI: 10.1080/08982112.2019.1610775

To link to this article: https://doi.org/10.1080/08982112.2019.1610775

QUALITY ENGINEERING
2019, VOL. 31, NO. 3, 516–521
https://doi.org/10.1080/08982112.2019.1610775

Rejoinder

Sumin Shen, Huiying Mao, and Xinwei Deng

Department of Statistics, Virginia Tech, Blacksburg, Virginia, USA

Background

In the article concerning our recent publication (Shen, Mao, and Deng 2018), Steiner and Mackay (2019) suggested that our EM-algorithm approach to the open challenges (Jensen 2018) may fail in two simulation cases, that is, when the sample size is larger than 1000 and when the correlation between the predictor (intermediate measurement) X and the response (final measurement) Y is small. They attempted to show that the challenge problem is impossible to solve and suggested that some paired intermediate and final measurements are needed. We greatly appreciate the interest and comments from Steiner and Mackay (2019). Here we would like to provide a rejoinder to their responses.

Assume that there is an available data set with m samples measured at the intermediate stage, $\boldsymbol{x}_{obs} = (x_1, x_2, ..., x_m)$, and $n-m$ samples measured at the final stage, $\boldsymbol{y}_{obs} = (y_{m+1}, y_{m+2}, ..., y_n)$. In Steiner and Mackay (2019), since intermediate and final measurements are collected from *different sets of units*, they obtained the *marginal* log-likelihood function as

$$\begin{aligned} l_{marg}(\alpha, \delta, \mu_y, \sigma_y) &= l_x(\alpha, \delta) + l_y(\mu_y, \sigma_y) \\ &= -m\ln(\delta) - \frac{\sum_{i=1}^{m}(x_i-\alpha)^2}{2\delta^2} \\ &\quad -(n-m)\ln(\sigma_y) - \frac{\sum_{i=m+1}^{n}(y_i-\mu_y)^2}{2\sigma_y^2} \end{aligned} \tag{1}$$

In their setting, they have considered the relationship between the predictor (intermediate measurement) X and the response (final measurement) Y as $Y = \beta_0 + \beta_1 X + \epsilon$ with $X \sim N(\alpha, \delta^2)$ and $\epsilon \sim N(0, \sigma^2)$. Thus, the pair (X, Y) can be viewed from a bivariate normal distribution with parameters $(\mu_x, \sigma_x, \mu_y, \sigma_y, \rho)$, where ρ is the correlation between the intermediate measurement X and the final measurement Y.

To study the relationship between the intermediate measurement X and the final measurement Y, we (Shen, Mao, and Deng 2018) believe that the consideration of missing intermediate and final measurements is essential in fully understanding the open challenge problem since intermediate and final measurements are collected from *different sets of units*. With the missing measurements, the paired complete data set $(\boldsymbol{x}, \boldsymbol{y})$ can be expressed as

$$\boldsymbol{x} = (\boldsymbol{x}_{obs}, \boldsymbol{x}_{mis}) = (x_1, ..., x_m, x^*_{m+1}, ..., x^*_n), \tag{2}$$

$$\boldsymbol{y} = (\boldsymbol{y}_{mis}, \boldsymbol{y}_{obs}) = (y^*_1, ..., y^*_m, y_{m+1}, ..., y_n). \tag{3}$$

Then the likelihood function based on *joint distribution* can be established to investigate the dependency between X and Y. While using marginal likelihood function as in Steiner and Mackay (2019) for such an investigation is not adequate and not well defined.

This article is organized in the following way. First, we provide our opinions on the Steiner and Mackay's (2019) on the challenge problem and point out the flaws in their statement. Then, we justify our EM-algorithm approach in the simulations used in Steiner and Mackay (2019). Lastly, we conclude with a short discussion on the suggestion that some paired observations are needed in order to solve the challenge problem.

Is the challenge problem impossible to solve?

The goal of the open challenge problem (Jensen 2018) is to quantify the relationship between the intermediate and the final measurements. Consequently, the tolerance of the intermediate measurements given the specification on the final measurements is obtained from the relationship. However, when the measurements of products are destructive, it is difficult to quantify the relationship because it is not possible to test the same product twice. It is worth remarking that the focus is on the relationship between the

CONTACT Xinwei Deng xdeng@vt.edu Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA.

Table 1. Simulated mean and standard deviation of the estimates for parameters α, β_0, $|\beta_1|$, δ, and σ at different sample sizes.

Parameter	Method		$m=20$	$m=50$	$m=100$	$m=1000$	$m=5000$
α	EM	mean	5.11	5.11	5.09	5.05	4.91
		sd	0.18	0.12	0.07	0.09	0.08
	LR	mean	5.00	5.00	5.00	5.00	5.00
		sd	0.08	0.05	0.04	0.01	0.01
β_0	EM	mean	27.16	26.29	26.13	26.47	27.23
		sd	8.04	5.07	2.79	1.02	0.65
	LR	mean	29.95	29.95	29.98	30.01	30.00
		sd	2.31	1.44	1.01	0.34	0.14
$\lvert\beta_1\rvert$	EM	mean	5.35	5.51	5.57	5.60	5.76
		sd	1.62	1.05	0.57	0.19	0.12
	LR	mean	5.01	5.01	5.00	5.00	5.00
		sd	0.46	0.29	0.20	0.07	0.03
δ	EM	mean	0.52	0.52	0.51	0.51	0.51
		sd	0.12	0.08	0.04	0.01	0.01
	LR	mean	0.50	0.50	0.50	0.50	0.50
		sd	0.05	0.04	0.03	0.01	0.00
σ	EM	mean	0.67	0.61	0.58	0.59	0.58
		sd	0.37	0.24	0.11	0.04	0.02
	LR	mean	1.43	1.44	1.44	1.44	1.44
		sd	0.16	0.11	0.07	0.02	0.01

intermediate measurement X and the final measurement Y, not on the statistical correlation between X and Y.

Steiner and Mackay (2019) proposed to view the challenge problem by deriving the marginal log-likelihood function in Eq. [1] for the observed measurements $\boldsymbol{x}_{obs}$ and $\boldsymbol{y}_{obs}$. Based on the linear model assumption, they then reparametrize their log-likelihood with

$$\mu_x = \alpha,\ \ \mu_y = \beta_0 + \beta_1\alpha,\ \ \sigma_y = \sqrt{\beta_1^2\delta^2 + \sigma^2}. \tag{4}$$

Thus, their Eq. [1] can be rewritten as

$$\begin{aligned} l_{marg}(\alpha,\delta,\mu_y,\sigma_y) = &-m\ln(\delta) - \frac{\sum_{i=1}^{m}(x_i-\alpha)^2}{2\delta^2} \\ &-\frac{n-m}{2}\ln\left(\beta_1^2\delta^2+\sigma^2\right) \\ &-\frac{\sum_{i=m+1}^{n}(y_i-\beta_0-\beta_1\alpha)^2}{2\left(\beta_1^2\delta^2+\sigma^2\right)}. \end{aligned} \tag{5}$$

We humbly point out that Eq. [5], which is also equation 5 in Steiner and Mackay (2019), is a marginal log-likelihood function with "over-parametrization" problem. Since the relationship between the intermediate measurement (X) and the final measurement (Y) is assumed as $Y = \beta_0 + \beta_1 X + \epsilon$, the joint likelihood function based on the joint distribution $p(y|x)p(x)$ should be considered for analysis rather than the marginal likelihood based on $p(x)p(y)$, where $p(\cdot)$ is the corresponding probability density function. Consequently, a linear relationship between the paired intermediate and final measurements can be established with the consideration of missing observations.

In addition, we would like to point out two false statements in the section of "Why the Challenge as Stated is Impossible to Solve" (Steiner and Mackay 2019). First, the sentence "We have Equation (5). The function (5) is then maximized using the EM algorithm." is not correct. We have the joint log-likelihood function clearly expressed in the section "Expectation-Maximization algorithm" in Shen, Mao, and Deng (2018). Specifically, the log-likelihood function after replacing the missing values in $\boldsymbol{x}$ and $\boldsymbol{y}$ with the respective conditional expectation given the observed data, $\mathrm{E}(x_i|y_i;\boldsymbol{\theta})$ and $\mathrm{E}(y_i|x_i;\boldsymbol{\theta})$, is

$$\begin{aligned} E[l_c(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y})] = &-n\log(2\pi) - n\log(\delta\sigma) \\ &-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left[E(y_i^2|x_i;\boldsymbol{\theta}) - 2(\beta_0+\beta_1x_i)E(y_i|x_i;\boldsymbol{\theta}) + (\beta_0+\beta_1x_i)^2\right] \\ &-\frac{1}{2\sigma^2}\sum_{i=m+1}^{n}\left[(y_i-\beta_0)^2 + \beta_1^2E(x_i^2|y_i;\boldsymbol{\theta}) - 2\beta_1(y_i-\beta_0)E(x_i|y_i;\boldsymbol{\theta})\right] \\ &-\frac{1}{2\delta^2}\sum_{i=1}^{m}(x_i-\alpha)^2 - \frac{1}{2\delta^2}\sum_{i=m+1}^{n}\left[E(x_i^2|y_i;\boldsymbol{\theta}) - 2\alpha E(x_i|y_i;\boldsymbol{\theta}) + \alpha^2\right], \end{aligned} \tag{6}$$

where the conditional expectation is given by

$$\begin{aligned} E(y_i|x_i) &= \beta_0 + \beta_1x_i,\ \ i = 1,\ldots,m. \\ E(y_i^2|x_i) &= \sigma^2 + (\beta_0+\beta_1x_i)^2,\ \ i = 1,\ldots,m. \\ E(x_i^2|y_i) &= \frac{\sigma^2\delta^2}{\sigma^2+\beta_1^2\delta^2} + \left(E(x_i|y_i)\right)^2,\ \ i = m+1,\ldots,n. \end{aligned}$$

Clearly, the jointly log-likelihood function $E[l_c(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y})]$ in Eq. [6] is different from the marginal log-likelihood function $l_{marg}(\alpha,\delta,\mu_y,\sigma_y)$ in Eq. [5].

Second, the sentence "Using the real data in Shen et al., we have $\hat{\mu}_x = 41.778$, $\hat{\sigma}_x^2 = 30.957$ that do not match the estimates for $\hat{\alpha}, \hat{\delta}_x^2$ for any starting value of β_1 as given in their Table 1." is not correct. In Table 1 of our paper (Shen, Mao, and Deng 2018), it is clearly written that $\hat{\alpha} \approx 41$ and $\hat{\delta}^2 \approx 28$ under the defined notation $X \sim N(\alpha, \delta^2)$, which are close to the estimates from Steiner and Mackay's numerical evaluation. These two false statements make their conclusions unreliable.

Justification of the EM-algorithm approach in numerical cases

Steiner and Mackay (2019) considered two simulation cases where in both cases $n = 2m$ and each simulation case is repeated for 1000 runs. We have re-run the two simulation studies to justify the EM-algorithm

approach in "Simulation study 1 - effect of sample size" and "Simulation study 2 - effect of correlation" sections. We modified the R code provided by Steiner and Mackay (2019) to run the simulation studies and made the code available in Bitbucket (https://bitbucket.org/vtshen/rpackages/src/master/). The simulation settings are almost identical, except that we used the original data generation mechanism in Shen, Mao, and Deng (2018), instead of the one mentioned in Steiner and Mackay (2019). Our data generation approach is intended to reflect the underlying data generation mechanism in the challenge problem. It allows us to conduct comparison with the benchmark method: using linear regression (LR) on the full data $(\boldsymbol{x}, \boldsymbol{y})$ for evaluating the effect of missing observations on parameter estimation.

Note that there are typos found in the simulation section in Shen, Mao, and Deng (2018). The true values for the parameters including the intercept in the LR model, β_0, and the mean of the normal distribution of variable X, α, should be 30 and 5, respectively.

Moreover, in the "Simulation study 3 - effect of signal-to-noise ratio" section, we will investigate the effect of signal-to-noise ratio (SNR) on the performance of the proposed EM-algorithm approach, where the SNR value varies within {0.5, 1, 3, 6, 10} given parameters $(\alpha, \beta_0, \beta_1, \delta, m)$.

Simulation study 1 - effect of sample size

In the first simulation study, we follow the settings in Steiner and Mackay (2019) where the parameters $(\alpha, \beta_0, \beta_1, \delta, \text{SNR}) = (5, 30, 5, 0.5, 3)$ and the sample size m is within {20, 50, 100, 1000, 5000}. The study is intended to evaluate the effect of sample size on the performance of the EM-algorithm approach. Steiner and Mackay (2019) reported that the EM-algorithm approach works reasonably well when the sample size is less than 1000, while its performance gets worse, especially in terms of the parameter estimation of β_1, when the sample size is larger than 1000. It is reported previously from Steiner and Mackay that the distribution of estimates of β_1 appears to be bimodal with one mode near the true value and another mode at another value. We have re-run the simulation and also found that the provided original R code would deteriorates the performance of the EM-algorithm when the sample size is greater than 1000. The reason is due to encountering singularity errors in the matrix inversion in the **M-step**. Now we have fixed the bugs in the R code regarding the β_0 and β_1 update in the **M-step** so that the updated version is robust for large sample sizes. In the updated R code, when singularity error occurs, we use the explicit estimates of β_0 and β_1 in **M-step** in our paper (Shen, Mao, and Deng 2018). That is,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \hat{\alpha},$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{m} (x_i - \hat{\alpha}) \left[E\left(y_i | x_i; \boldsymbol{\theta}\right) - \bar{y}\right] + \sum_{i=m+1}^{n} \left[E\left(x_i | y_i; \boldsymbol{\theta}\right) - \hat{\alpha}\right] (y_i - \bar{y})}{\sum_{i=1}^{m} (x_i - \hat{\alpha})^2 + \sum_{i=m+1}^{n} \left[E\left(x_i | y_i; \boldsymbol{\theta}\right) - \hat{\alpha}\right]^2}.$$

After these changes in the R code, we found that the performance of the EM-algorithm approach is consistently good at both small and large sample sizes.

The simulation results are provided in Table 1, which reports the estimates for parameters α, β_0, $|\beta_1|$, δ, and σ. The $|\beta_1|$ instead of β_1 is reported because the sign of the estimated β_1 is determined by the sign of the initial β_1 in the proposed EM-algorithm approach. However, the absolute value of the estimated β_1 is not related to the sign of the initial β_1. In addition, we show that the tolerance of x does not depend on the sign of the estimated β_1 in Appendix (Shen, Mao, and Deng 2018). Note that the parameters μ_x, μ_y, σ_x, and σ_y are not of interest in the open challenge problem. The goal of the open challenge problem is to estimate the tolerance of the intermediate measurements given the specification on the final measurements, which is obtained in the framework of the LR relationship in our approach (Shen, Mao, and Deng 2018). It is seen in Table 1 that the EM-algorithm approach provides reasonably good estimators at various sample sizes. Note that the EM-algorithm does not provide as accurate estimators as the LR method even though they are comparable to some extent. This is expected because the LR method utilizes all the observations assuming that the missing intermediate and final measurements are recovered.

Steiner and Mackay (2019) also reported that the estimates for the correlation ρ are always close to 1 or -1. Thus, they claimed that the EM-algorithm

approach could not possibly work because of the poor estimation of ρ, regardless of the fact that the EM-algorithm approach provides reasonably good estimates for other parameters such as $|\beta_1|$. We would like to reiterate that the goal of the open challenge problem is to estimate the tolerance of the intermediate measurements given the specification on the final measurements. Thus, quantifying the relationship between the intermediate and final measurements (i.e., estimate of β_1) is more important, while the estimate of ρ is not of major interest. Note that in Fig. 1 of the paper Shen, Mao, and Deng (2018), it is visually evident that the estimated ρ values from $\boldsymbol{x}$ and $\boldsymbol{y}$ are close to 1 or -1. The explanation unfolds with the following justification. In the EM-algorithm approach, we update the parameters β_0, β_1, α, σ^2, and δ^2 until their convergence. The ρ is thus estimated based on $\boldsymbol{x}$ and $\boldsymbol{y}$, including the imputed missing intermediate and final measurements. Because the missing parts in observations are imputed based on the estimated LR model, the estimated value of σ^2 is small compared to that of $\beta_1\delta$. According to the equation $\rho = \frac{\beta_1\delta}{\sqrt{\beta_1^2\delta^2+\sigma^2}}$, the estimates of ρ would be close to 1 or -1.

Simulation study 2 - effect of correlation

Steiner and Mackay (2019) explored the effect of correlation (between intermediate measurement X and final measurement Y) on the method performance, but they did not present the effect of changing correlation on the SNR. According to the equations in Steiner and Mackay (2019), $\beta_1 = 2(2.77\rho)$ and $\sigma = \sqrt{(2.77^2 - 0.25\beta_1^2)}$, changing the correlation would affect both β_1 and σ. Note that SNR is defined (Wu and Hamada 2009) as

$$\text{SNR} = \frac{\text{Variance}(\beta_0 + \beta_1 X)}{\sigma^2} = \frac{\beta_1^2\delta^2}{\sigma^2} = \frac{\rho^2}{1-\rho^2}.$$

Therefore, in their simulation study 2, they actually investigated two effects, β_1 and SNR, at the same time. Specifically, when the setting ρ is within {1, 0.9, 0.75, 0.5, 0.25}, the corresponding β_1 is {5.54, 4.99, 4.16, 2.77, 1.39} and SNR is {∞, 4.26, 1.29, 0.33, 0.07}. These actions complicated the simulation processes and we believe it would be better to investigate the effects of β_1 and SNR separately. However, the purpose of this letter is to give a rejoinder to their responses (Steiner and Mackay, 2019). Thus, we first follow their approach but add more details to provide our response in the simulation study 2. Furthermore, we investigate the effect of SNR in Section "Simulation study 3 - effect of signal-to-noise ratio".

Table 2. Simulated mean and standard deviation of the estimates for parameters μ_x, μ_y, ρ, σ_x, and σ_y at different correlation values.

Parameter	Method		$\rho = 0.99$	$\rho = 0.9$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.25$
μ_x	EM	mean	5.12	5.15	5.20	5.11	5.59
		sd	0.13	0.17	0.21	0.18	0.73
	naiveLR	mean	5.00	5.00	5.00	5.00	5.00
		sd	0.07	0.07	0.07	0.07	0.07
	LR	mean	5.00	5.00	5.00	5.00	5.00
		sd	0.05	0.05	0.05	0.05	0.05
	marginal	mean	5.00	5.00	5.00	5.00	5.00
		sd	0.05	0.05	0.05	0.05	0.05
μ_y	EM	mean	76.31	74.25	70.80	65.72	59.87
		sd	0.54	0.64	0.74	0.54	1.22
	naiveLR	mean	76.93	74.93	71.60	66.09	60.53
		sd	0.39	0.40	0.36	0.26	0.14
	LR	mean	76.93	74.94	71.62	66.08	60.54
		sd	0.27	0.28	0.26	0.18	0.10
	marginal	mean	76.93	74.93	71.60	66.09	60.53
		sd	0.27	0.28	0.26	0.18	0.10
ρ	EM	mean	0.97	0.98	0.97	0.97	0.99
		sd	0.09	0.08	0.12	0.03	0.01
	naiveLR	mean	0.12	0.11	0.11	0.11	0.12
		sd	0.09	0.08	0.09	0.08	0.08
	LR	mean	0.99	0.92	0.83	0.75	0.72
		sd	0.00	0.01	0.02	0.03	0.04
	marginal	mean	–	–	–	–	–
		sd	–	–	–	–	–
σ_x	EM	mean	0.53	0.54	0.57	0.53	1.00
		sd	0.08	0.10	0.13	0.13	0.47
	naiveLR	mean	0.50	0.50	0.50	0.50	0.50
		sd	0.05	0.05	0.05	0.05	0.05
	LR	mean	0.50	0.50	0.50	0.50	0.50
		sd	0.04	0.04	0.04	0.04	0.04
	marginal	mean	0.50	0.50	0.50	0.50	0.50
		sd	0.04	0.04	0.04	0.04	0.04
σ_y	EM	mean	2.80	2.75	2.59	1.85	1.49
		sd	0.25	0.27	0.32	0.31	0.51
	naiveLR	mean	2.79	2.72	2.50	1.84	0.97
		sd	0.28	0.29	0.28	0.20	0.11
	LR	mean	2.77	2.71	2.49	1.82	0.96
		sd	0.20	0.22	0.21	0.15	0.08
	marginal	mean	2.77	2.69	2.48	1.82	0.96
		sd	0.20	0.22	0.21	0.15	0.08

We follow the settings in the simulation study 2 (Steiner and Mackay 2019) where the correlation ρ varies. That is, $(\mu_x, \sigma_x, \mu_y, \sigma_y) = (5, 0.5, 55, 2.77)$, which corresponds to the settings $(\alpha, \delta) = (5, 0.5)$, $\beta_1 = 2(2.77\rho)$, $\beta_0 = 55 - \beta_1$, and $\sigma = \sqrt{(2.77^2 - 0.25\beta_1^2)}$. Note we consider $\rho = \{0.99, 0.9, 0.75, 0.5, 0.25\}$. We do not consider the case where $\rho = 1$ because when $\rho = 1$, $\sigma = 0$ and it is not practical.

It is seen in Table 2 that the reproduced results for parameters include μ_x (i.e. α), σ_x (i.e. δ), μ_y, σ_y, and ρ at various ρ values from the proposed EM-algorithm method (Shen, Mao, and Deng 2018) and the so-called marginal method (Steiner and Mackay 2019). Furthermore, we add two more approaches in comparison: the naive LR (naiveLR) method and the LR

Table 3. Simulated mean and standard deviation of the estimates for parameters α, β_0, $|\beta_1|$, δ, and σ at different correlation values.

Parameter	Method		$\rho = 0.99$	$\rho = 0.9$	$\rho = 0.75$	$\rho = 0.5$	$\rho = 0.25$
α	EM	mean	5.12	5.15	5.20	5.11	5.59
		sd	0.13	0.17	0.21	0.18	0.73
	naiveLR	mean	5.00	5.00	5.00	5.00	5.00
		sd	0.07	0.07	0.07	0.07	0.07
	LR	mean	5.00	5.00	5.00	5.00	5.00
		sd	0.05	0.05	0.05	0.05	0.05
β_0	EM	mean	49.39	48.22	47.33	48.06	50.91
		sd	4.80	4.40	4.64	2.49	1.16
	naiveLR	mean	77.10	74.87	71.65	66.02	60.60
		sd	4.15	3.88	3.63	2.61	1.44
	LR	mean	49.51	50.07	50.82	52.26	53.59
		sd	0.40	1.09	1.41	1.20	0.68
$\|\beta_1\|$	EM	mean	5.26	5.07	4.53	3.46	1.66
		sd	0.99	0.93	0.97	0.51	0.49
	naiveLR	mean	0.65	0.61	0.57	0.41	0.23
		sd	0.50	0.47	0.44	0.32	0.17
	LR	mean	5.49	4.97	4.16	2.76	1.39
		sd	0.08	0.22	0.28	0.24	0.14
δ	EM	mean	0.53	0.54	0.57	0.53	1.00
		sd	0.08	0.10	0.13	0.13	0.47
	naiveLR	mean	0.50	0.50	0.50	0.50	0.50
		sd	0.05	0.05	0.05	0.05	0.05
	LR	mean	0.50	0.50	0.50	0.50	0.50
		sd	0.04	0.04	0.04	0.04	0.04
σ	EM	mean	0.56	0.53	0.49	0.46	0.15
		sd	0.22	0.19	0.25	0.12	0.02
	naiveLR	mean	2.76	2.69	2.48	1.82	0.96
		sd	0.28	0.29	0.28	0.20	0.11
	LR	mean	0.39	1.09	1.37	1.19	0.67
		sd	0.03	0.08	0.10	0.08	0.05

Table 4. Simulated mean and standard deviation of the estimates for parameters α, β_0, $|\beta_1|$, δ, and σ at different SNR.

Parameter	Method		SNR = 0.5	SNR = 1	SNR = 3	SNR = 6	SNR = 10
α	EM	mean	5.07	5.07	5.11	5.14	5.16
		sd	0.08	0.08	0.12	0.15	0.18
	LR	mean	5.00	5.00	5.00	5.00	5.00
		sd	0.05	0.05	0.05	0.05	0.05
β_0	EM	mean	12.75	20.13	26.26	28.27	29.25
		sd	6.04	4.74	4.95	4.89	4.66
	LR	mean	29.98	29.96	29.95	29.94	30.00
		sd	3.63	2.55	1.50	1.03	0.81
$\|\beta_1\|$	EM	mean	8.22	6.79	5.51	5.08	4.86
		sd	1.14	0.95	1.00	1.01	0.98
	LR	mean	5.00	5.01	5.01	5.01	5.00
		sd	0.73	0.51	0.30	0.20	0.16
δ	EM	mean	0.50	0.50	0.51	0.54	0.55
		sd	0.05	0.05	0.07	0.09	0.11
	LR	mean	0.50	0.50	0.50	0.50	0.50
		sd	0.03	0.04	0.03	0.04	0.03
σ	EM	mean	1.15	0.93	0.60	0.54	0.50
		sd	0.33	0.24	0.23	0.23	0.23
	LR	mean	3.52	2.50	1.43	1.02	0.79
		sd	0.24	0.17	0.10	0.07	0.06

method. The naiveLR method is to apply the LR method on the observations $\boldsymbol{x}_{obs}$ and $\boldsymbol{y}_{obs}$, given the number of intermediate and final measurements is equal in the simulation settings. The LR method is to apply the LR method on the complete paired measurements, $\boldsymbol{x}$ and $\boldsymbol{y}$, which include both missing and observed measurements. Note that in the simulation case we can have the complete paired measurement.

For a fair comparison, we include the estimates for parameters α, δ, β_0, $|\beta_1|$, and σ in Table 3. When $\rho > 0.75$, the EM-algorithm method provides reasonably good performance in terms of β_1. Note that the EM-algorithm method obtains the estimates considering the paired measurements $\boldsymbol{x}$ and $\boldsymbol{y}$ with missing observations, while the so-called marginal method is built on the $\boldsymbol{x}_{obs}$ and $\boldsymbol{y}_{obs}$. The marginal method has comparable performance with the naiveLR and the LR method in terms of μ_x (i.e., α), μ_y, σ_x (i.e., δ), σ_y, but these parameters are not of main interest. The naiveLR has the worst performance in β_0, β_1, and σ estimation compared to the EM and the LR methods. The poor performance of the naiveLR method emphasizes the importance of considering paired $\boldsymbol{x}$ and $\boldsymbol{y}$, including the missing and observed measurements, for quantifying the relationship between intermediate and final measurements.

Again, Steiner and Mackay (2019) mentioned the estimation of ρ is close to 1 or -1, but it can be explained by the same reasoning we have stated in Section "Simulation study 1 - effect of sample size".

Simulation study 3 - effect of signal-to-noise ratio

In this section, the objective is to investigate the effect of SNR on the performance of the EM-algorithm method. Such a simulation study is not performed in Steiner and Mackay (2019). In this simulation, we set $(\alpha, \beta_0, \beta_1, \delta, m) = (5, 30, 5, 0.5, 50)$ and vary SNR within {0.5, 1, 3, 6, 10}. The simulation results are provided in Table 4.

Table 4 shows that when the SNR is as high as 6 or 10, the EM-algorithm approach performs reasonably well. While when the SNR is as low as 0.5 or 1, the performance of the EM-algorithm method is not that good, especially in terms of the estimation of β_1. The performance of the LR method also deteriorates when the SNR is low as its standard deviation of estimator β_1 gets larger. To better understand the effect of SNR on the method performance, recall the proportion of variance explained (PVE, Hastie, Tibshirani, and Tibshirani 2017) as

$$\text{PVE} = 1 - \frac{\text{Variance}(\epsilon)}{\text{Variance}(\beta_0 + \beta_1 x + \epsilon)} = \frac{\text{SNR}}{1 + \text{SNR}}.$$

When the SNR decreases to small values, the PVE gets close to zero. For example, when SNR is 0.5 or 1,

then the corresponding PVE is only 33.3% or 50%, respectively. The small value of SNR implies that the data is very noisy and the LR method generally does not work well. Therefore, it is expected that the performance of methods based on the linear regression, including the EM-algorithm method and the LR method, is getting worse with SNR being smaller.

Summary

Overall, the performance of our proposed EM-algorithm method (Shen, Mao, and Deng 2018) in the numerical studies are properly justified. First, the EM-algorithm method performs reasonably well and consistently at various sample sizes. Second, the estimation of the correlation ρ is based on $\boldsymbol{x}$ and $\boldsymbol{y}$, including the imputed missing intermediate and final measurements in the EM-algorithm approach. The missing parts in observations are imputed based on the estimated LR model. Thus, the estimated value of σ^2 is small compared to that of $\beta_1\delta$. According to the equation $\rho = \frac{\beta_1\delta}{\sqrt{\beta_1^2\delta^2+\sigma^2}}$, the estimates of ρ could be close to 1 or -1, but the major focus of the open challenge problem is not on estimation of ρ. Third, the performance of the proposed EM-algorithm is reasonably good when SNR is high but deteriorates when SNR is as low as 0.5 or 1. The reason is that when SNR is low, the data is highly noisy, and the LR methods do not work well in such situations.

We would like to emphasize that the performance of the proposed EM-algorithm method can be affected by the choice of initial values. In the associated R code, we provide a strategy to choose appropriate initial values for parameters, especially β_1. The initial values are chosen such that the imputed $\boldsymbol{x}_{mis}$ has similar distribution as $\boldsymbol{x}_{obs}$. This is not the only strategy to select appropriate initial values. Other reasonable strategies for searching initial values are also possible.

Discussion

Steiner and Mackay (2019) suggested obtaining some paired data in combination with the unpaired intermediate and final measurements. If some paired data were available, the proposed EM-algorithm method can easily accommodate such situations. What we need to update is the complete log-likelihood function in the E-step in Shen, Mao, and Deng (2018). After that, the M-step, which is maximizing the expected complete log-likelihood function, is straightforward. We also would like to point out that, besides the proposed EM-algorithm method in Shen, Mao, and Deng (2018) to address the open challenge problem, a more general method can be the Bayesian method (Kang et al. 2018) to construct the joint likelihood function $p(y|x)p(x)$ for quantifying the relationship between intermediate and final measurements. Results on this direction will be reported in the future.

Acknowledgments

We would like to sincerely thank the Editor for this great opportunity to provide a rejoinder to Steiner and Mackay's comments. Their comments and the reviewer's comments are really insightful to enhance our understanding of the open challenge problem.

ORCID

Xinwei Deng http://orcid.org/0000-0002-1560-2405

References

Hastie, T., R. Tibshirani, and R. J. Tibshirani. 2017. Extended comparisons of best subset selection, forward stepwise selection, and the Lasso. arXiv preprint arXiv: 1707.08692.

Jensen, W. A. 2018. Open challenges: Correlation of intermediate and final measurements. *Quality Engineering* 30 (1):167–8. doi:10.1080/08982112.2017.1402936.

Kang, L., X. Kang, X. Deng, and R. Jin. 2018. Bayesian hierarchical models for quantitative and qualitative responses. *Journal of Quality Technology* 50(3):290–308. doi:10.1080/00224065.2018.1489042.

Shen, S., H. Mao, and X. Deng. 2018. An EM-algorithm approach to open challenges on correlation of intermediate and final measurements. *Quality Engineering*. Advance online publication. doi:10.1080/08982112.2018.1497181.

Steiner, S. H., and R. J. MacKay. 2019. Response to open challenges on correlation of intermediate and final Measurements - Solving the impossible problem? Quality Engineering.

Wu, C. F. J., and M. S. Hamada. 2009. *Experiments: Planning, analysis, and optimization*. 2nd Edition, Hoboken, NJ: Wiley.