An expectation-maximization scheme for measurement error models

Anindya Bhadra

Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN
47907-2066
bhadra@purdue.edu

Abstract

We treat the problem of maximum likelihood estimation in measurement error models. Direct maximization of the analytically intractable likelihood in such models is difficult. We derive an efficient expectation-maximization scheme for truncated polynomial spline models of degree one. Simulation results confirm the effectiveness of the proposed method.

Key Words: Expectation-Maximization; Maximum likelihood; Measurement error; Splines.

1 Introduction

There is a considerable literature on nonparametric regression problems where the covariates are measured with error. A semi-parametric framework is described by Carroll et al. (1999) who use the SIMEX method of Cook and Stefanski (1994). Alternatively, a fully Bayesian framework is described by Berry et al. (2002) who find marked performance improvement in simulations over the SIMEX-based method of Carroll et al. (1999) and the iterative conditional modes (ICM) approach of Besag (1986). The Bayesian approach of Berry et al. (2002) relies on a Metropolis-Hastings algorithm, which can severely restrict its applicability due to poor mixing and the requirement to carefully tune the proposal distribution. Bhadra and Carroll (2016) show that for the special case of degree one truncated polynomial and B-spline models, it is actually possible to design a Gibbs sampler, thereby eliminating the problems associated with the Metropolis-Hastings step in Berry et al. (2002).

Despite these prior works in semi-parametric and Bayesian estimation procedures, a maximum likelihood estimation (mle) technique for these models have not been studied carefully. This is in part due to the intractability of the likelihood that does not have a closed form and does not lend itself to easy maximization. Using the recent results by

Bhadra and Carroll (2016), we show that for the case of degree one truncated polynomial splines, it is possible to formulate an efficient expectation-maximization (E-M) algorithm (Dempster et al., 1977) that permits finding the mle in these models.

2 The likelihood for measurement error models

Following Carroll et al. (1999), consider the measurement error model:

$$Y_i = g(X_i) + \epsilon_i, \quad (1)$$

$$W_{ij} = X_i + U_{ij}, \quad j = 1, \dots, m_i, \quad i = 1, \dots, n, \quad (2)$$

where n is the sample size and $m_i > 1$ is the number of replicates for sample i for $i = 1, \dots, n$. Let $(Y_i, W_{i1}, \dots, W_{im_i})$ for $i = 1, \dots, n$ be the observed data and let X_i be latent Gaussian random variables with mean μ_x and variance σ_x^2 . Let ϵ_i be independent and identically distributed normal random variables with mean 0 and variance σ_ϵ^2 ; U_{ij} be independent and identically distributed normal random variables with mean 0 and variance σ_u^2 ; U_{ij} and ϵ_i be uncorrelated. Let $g(\cdot)$ denote the true, unknown mean function, which is assumed to be smooth. A useful technique for modeling $g(\cdot)$ to approximate it to a desired degree of smoothness is via penalized truncated polynomial splines (P-splines) bases (Eilers and Marx, 1996). Thus, consider the following hierarchical model for $k = 1, \dots, K$:

$$Y_i = \beta_0 + \beta_1 X_i + \sum_{k=1}^K B_k(X_i) \theta_k + \varepsilon_i, \quad (3)$$

$$W_{ij} = X_i + U_{ij}, \quad (4)$$

$$\varepsilon_i \sim \text{Normal}(0, \sigma_\varepsilon^2), U_{ij} \sim \text{Normal}(0, \sigma_u^2), X_i \sim \text{Normal}(\mu_x, \sigma_x^2). \quad (5)$$

Our goal is to find the mle of the parameters $\Psi = (\beta_0, \beta_1, \theta_1, \dots, \theta_K, \sigma_\varepsilon^2, \sigma_u^2, \mu_x, \sigma_x^2)$. Define $M = \sum_{i=1}^n m_i$, $W_{i\bullet} = \sum_{j=1}^{m_i} W_{ij}$, $\bar{W}_{i\bullet} = W_{i\bullet}/m_i$. If X_i were observed, this is a problem of finding the mle in a model with normal likelihood, which is simple. In an E-M scheme, the difficulty lies in computing the conditional expectation of the complete data log likelihood given the observed data. We show this has an analytically tractable form for degree 1

truncated polynomial splines, for which

$$B_k(X_i) = (X_i - \omega_k)_+,$$

where $\omega_1, \dots, \omega_K$ are the known knot points. Denote by $\boldsymbol{\Theta}$ the $(K + 2)$ parameter vector $(\beta_0, \beta_1, \theta_1, \dots, \theta_K)^T$ and define the $n \times (K + 2)$ matrix $\mathbf{Z}$ as

$$\mathbf{Z} = \begin{bmatrix} 1 & X_1 & B_1(X_1) & \dots & B_K(X_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & B_1(X_n) & \dots & B_K(X_n) \end{bmatrix}.$$

The log likelihood of the complete data in Equations (3-5) is given by

$$\begin{aligned} \mathcal{L} = & -(2\sigma_\varepsilon^2)^{-1}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Theta})^T(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Theta}) - (n/2)\log(\sigma_\varepsilon^2) \\ & -(2\sigma_u^2)^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} (W_{ij} - X_i)^2 - (M/2)\log(\sigma_u^2) - (2\sigma_x^2)^{-1} \sum_{i=1}^n (X_i - \mu_x)^2 - (n/2)\log(\sigma_x^2) + \text{constant}, \end{aligned}$$

where “constant” is a collection of all terms that are independent of data as well as model parameters. We need to impute the missing data in the complete data log likelihood. Expanding the log likelihood, except for an irrelevant constant,

$$\begin{aligned} \mathcal{L} = & -(2\sigma_\varepsilon^2)^{-1}(\mathbf{Y}^T\mathbf{Y} - \boldsymbol{\Theta}^T\mathbf{Z}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{Z}\boldsymbol{\Theta} + \boldsymbol{\Theta}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Theta}) - (n/2)\log(\sigma_\varepsilon^2) \\ & -(2\sigma_u^2)^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} (W_{ij} - X_i)^2 - (M/2)\log(\sigma_u^2) - (2\sigma_x^2)^{-1} \sum_{i=1}^n (X_i - \mu_x)^2 - (n/2)\log(\sigma_x^2). \end{aligned}$$

We have $\mathbf{Y}^T\mathbf{Y} = \sum_{i=1}^n Y_i^2$. We also have

$$\boldsymbol{\Theta}^T\mathbf{Z}^T\mathbf{Y} = \mathbf{Y}^T\mathbf{Z}\boldsymbol{\Theta} = \beta_0 \sum_{i=1}^n Y_i + \beta_1 \sum_{i=1}^n X_i Y_i + \sum_{i=1}^n \sum_{k=1}^K \theta_k B_k(X_i) Y_i,$$

and

$$\Theta^T \mathbf{Z}^T \mathbf{Z} \Theta = \sum_{i=1}^n \left\{ \beta_0 + \beta_1 X_i + \sum_{k=1}^K \theta_k B_k(X_i) \right\}^2.$$

Thus, the log likelihood is

$$\begin{aligned} \mathcal{L} = & -(2\sigma_\varepsilon^2)^{-1} \sum_{i=1}^n \left[Y_i^2 - 2\beta_0 Y_i - 2\beta_1 X_i Y_i - 2Y_i \sum_{k=1}^K \theta_k B_k(X_i) + \left\{ \beta_0 + \beta_1 X_i + \sum_{k=1}^K \theta_k B_k(X_i) \right\}^2 \right] \\ & -(n/2) \log(\sigma_\varepsilon^2) - (2\sigma_u^2)^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} (W_{ij} - X_i)^2 - (M/2) \log(\sigma_u^2) \\ & -(2\sigma_x^2)^{-1} \sum_{i=1}^n (X_i - \mu_x)^2 - (n/2) \log(\sigma_x^2). \end{aligned}$$

Expanding further, we get

$$\begin{aligned} \mathcal{L} = & -(2\sigma_\varepsilon^2)^{-1} \sum_{i=1}^n \left[Y_i^2 - 2\beta_0 Y_i - 2\beta_1 X_i Y_i - 2Y_i \sum_{k=1}^K \theta_k B_k(X_i) + \beta_0^2 + \beta_1^2 X_i^2 + \sum_{k=1}^K \theta_k^2 \{B_k(X_i)\}^2 \right. \\ & \left. + 2\beta_0 \beta_1 X_i + 2\beta_0 \sum_{k=1}^K \theta_k B_k(X_i) + 2\beta_1 \sum_{k=1}^K \theta_k X_i B_k(X_i) + 2 \sum_{k=1}^K \sum_{j=1}^{k-1} \theta_k \theta_j B_k(X_i) B_j(X_i) \right] \\ & -(n/2) \log(\sigma_\varepsilon^2) - (2\sigma_u^2)^{-1} \sum_{i=1}^n \left(\sum_{j=1}^{m_i} W_{ij}^2 + m_i X_i^2 - 2W_{i\bullet} X_i \right) - (M/2) \log(\sigma_u^2) \\ & -(2\sigma_x^2)^{-1} \left(\sum_{i=1}^n X_i^2 + n\mu_x^2 - 2\mu_x \sum_{i=1}^n X_i \right) - (n/2) \log(\sigma_x^2), \end{aligned} \tag{6}$$

as the complete data log likelihood. Clearly, in order to compute the observed data log likelihood, one needs to integrate out the latent X_i from the complete data log likelihood, which is an exercise that proves to be difficult. This renders a direct maximization of the observed data log likelihood untenable. However, in the next section, we show that it is possible to impute the latent X_i in the complete data log likelihood, making an E-M scheme for finding the mle feasible.

3 The proposed E-M algorithm

Define the convention $\omega_0 = -\infty$ and $\omega_{K+1} = \infty$. Our derivation of the E-M algorithm is based on the following result proved by Bhadra and Carroll (2016), who showed that for degree one truncated polynomial splines, the latent X_i given the observed data is distributed as a mixture of double truncated normal random variables, truncated at the knot points.

Proposition 1. (*Bhadra and Carroll, 2016*). *Define*

$$\begin{aligned}\zeta_{1iL} &= \{(\sigma_\varepsilon^2)^{-1}(\beta_1 + \sum_{\ell=1}^L \theta_\ell)^2 + m_i/(\sigma_u^2) + 1/(\sigma_x^2)\}^{-1}; \\ \zeta_{2iL} &= -(2\sigma_\varepsilon^2)^{-1}\{-2\beta_1 Y_i - 2Y_i \sum_{\ell=1}^L \theta_\ell + 2(\beta_0 - \sum_{\ell=1}^L \theta_\ell \omega_\ell)(\beta_1 + \sum_{\ell=1}^L \theta_\ell)\} \\ &\quad + (W_{i\bullet}/\sigma_u^2 + \mu_x/\sigma_x^2); \\ \zeta_{3iL} &= -(2\sigma_\varepsilon^2)^{-1}\{2Y_i \sum_{\ell=1}^L \theta_\ell \omega_\ell + (\beta_0 - \sum_{\ell=1}^L \theta_\ell \omega_\ell)^2\}.\end{aligned}$$

Further define

$$\begin{aligned}b_{iL} &= (\omega_{L+1} - \zeta_{1iL}\zeta_{2iL})/\sqrt{\zeta_{1iL}}; \\ a_{iL} &= (\omega_L - \zeta_{1iL}\zeta_{2iL})/\sqrt{\zeta_{1iL}}; \\ D_i &= \left[\sum_{L=0}^K \{\Phi(b_{iL}) - \Phi(a_{iL})\} \exp \{(1/2)\log(\zeta_{1iL}) + \zeta_{3iL} + \zeta_{1iL}(\zeta_{2iL})^2/2\}\right]^{-1}; \\ p_{iL} &= D_i \{\Phi(b_{iL}) - \Phi(a_{iL})\} \exp \{(1/2)\log(\zeta_{1iL}) + \zeta_{3iL} + \zeta_{1iL}(\zeta_{2iL})^2/2\} \\ &= D_i \sqrt{\zeta_{1iL}} \{\Phi(b_{iL}) - \Phi(a_{iL})\} \exp \{\zeta_{3iL} + \zeta_{1iL}(\zeta_{2iL})^2/2\}.\end{aligned}$$

Then, the complete conditional distribution of X_i is a mixture of truncated normal random variables on the intervals $[\omega_L, \omega_{L+1}]$ for $L = 0, \dots, K$, with means $\mu_{iL} = \zeta_{1iL}\zeta_{2iL}$, variances $\sigma_{iL}^2 = \zeta_{1iL}$ and with mixing probabilities p_{iL} .

The intuitive explanation for the above result to hold is that if a degree one spline is used then the only terms that enter the log likelihood are linear and quadratic in the latent X_i ; but not higher order terms in X_i . Proposition 1 then follows, essentially via a “completing the squares” argument. Using the above result, Bhadra and Carroll (2016) went on to develop a Gibbs sampler to perform Bayesian analysis for the model specified in Equations (3-5). However, it is possible to use the same result to develop an E-M algorithm to find the mle as

well. Since the complete conditional distribution of X_i is available from Proposition 1, the E-step reduces to imputing the required conditional expectations in the complete data log likelihood of Equation (6). Since the complete data likelihood is just a normal likelihood, maximizing it is considerably easier than maximizing the observed data log likelihood, which has no closed form. We detail the E and M steps in Algorithm 1 below. Steps (E.1-E.2) and (M.1-M.2) are repeated until convergence. Detailed derivation is provided in Appendix A.

Algorithm 1: E-M algorithm for measurement error models

Data: $\{Y_i, W_{ij}\}$, for $j = 1, \dots, m_i$ and $i = 1, \dots, n$.

Parameters: Knot points $\omega_1, \dots, \omega_K$; number of iterations T ; initial parameters $\hat{\Psi}^{(0)}$.

Result: Maximum likelihood estimate $\hat{\Psi}$ of $\Psi = (\beta_0, \beta_1, \theta_1, \dots, \theta_K, \sigma_\varepsilon^2, \sigma_u^2, \mu_x, \sigma_x^2)$.

```
// E-M iteration
for  $t = 0$  to  $T - 1$  do
    //  $(t + 1)$ th E-step
    (E.1) Compute  $p_{iL}, \mu_{iL}, \sigma_{iL}^2$  for  $L = 0, \dots, K$  by Proposition 1 with  $\Psi = \hat{\Psi}^{(t)}$ .
    (E.2) For  $i = 1, \dots, n$  and  $j, k = 1, \dots, K$ ; Set
        
$$\mathbb{E}[X_i] = \sum_{L=0}^K p_{iL} \mu_{iL},$$


$$\mathbb{E}[X_i^2] = \sum_{L=0}^K p_{iL} (\mu_{iL}^2 + \sigma_{iL}^2),$$


$$\mathbb{E}[B_k(X_i)] = \sum_{L=k}^K p_{iL} \mu_{iL} - \omega_k \sum_{L=k}^K p_{iL},$$


$$\mathbb{E}[X_i B_k(X_i)] = \sum_{L=k}^K p_{iL} (\mu_{iL}^2 + \sigma_{iL}^2) - \omega_k \sum_{L=k}^K p_{iL} \mu_{iL},$$


$$\mathbb{E}[\{B_k(X_i)\}^2] = \sum_{L=k}^K p_{iL} (\mu_{iL}^2 + \sigma_{iL}^2) - 2\omega_k \sum_{L=k}^K p_{iL} \mu_{iL} + \omega_k^2 \sum_{L=k}^K p_{iL},$$


$$\mathbb{E}[B_k(X_i) B_j(X_i)] =$$


$$\sum_{L=\max\{k,j\}}^K p_{iL} (\mu_{iL}^2 + \sigma_{iL}^2) - (\omega_k + \omega_j) \sum_{L=\max\{k,j\}}^K p_{iL} \mu_{iL} + \omega_k \omega_j \sum_{L=\max\{k,j\}}^K p_{iL}.$$

    //  $(t + 1)$ th M-step
    (M.1) Set
        
$$\hat{\beta}_0^{(t+1)} = \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_1^{(t)} \mathbb{E}[X_i] - \sum_{k=1}^K \hat{\theta}_k^{(t)} \mathbb{E}[B_k(X_i)])}{\sum_{i=1}^n 1},$$


$$\hat{\beta}_1^{(t+1)} = \frac{\sum_{i=1}^n (Y_i \mathbb{E}[X_i] - \hat{\beta}_0^{(t+1)} \mathbb{E}[X_i] - \sum_{k=1}^K \hat{\theta}_k^{(t)} \mathbb{E}[X_i B_k(X_i)])}{\sum_{i=1}^n \mathbb{E}[X_i^2]},$$


$$\hat{\theta}_k^{(t+1)} = \frac{\sum_{i=1}^n (Y_i \mathbb{E}[B_k(X_i)] - \hat{\beta}_0^{(t+1)} \mathbb{E}[B_k(X_i)] - \hat{\beta}_1^{(t+1)} \mathbb{E}[X_i B_k(X_i)] - \sum_{j=1}^{k-1} \hat{\theta}_j^{(t+1)} \mathbb{E}[B_j(X_i) B_k(X_i)])}{\sum_{i=1}^n \mathbb{E}[\{B_k(X_i)\}^2]}; \forall k,$$


$$\{\hat{\sigma}_\varepsilon^2\}^{(t+1)} = \frac{A_{\sigma_\varepsilon^2}^{(t+1)}}{n}; \text{ where } A_{\sigma_\varepsilon^2}^{(t+1)} \text{ is as defined in Appendix A.2,}$$


$$\{\hat{\sigma}_u^2\}^{(t+1)} = \frac{A_{\sigma_u^2}^{(t+1)}}{M}; \text{ where } A_{\sigma_u^2}^{(t+1)} = \sum_{i=1}^n \left( \sum_{j=1}^{m_i} W_{ij}^2 + m_i \mathbb{E}[X_i^2] - 2W_{i\bullet} \mathbb{E}[X_i] \right),$$


$$\{\hat{\sigma}_x^2\}^{(t+1)} = \frac{A_{\sigma_x^2}^{(t+1)}}{n}; \text{ where } A_{\sigma_x^2}^{(t+1)} = \sum_{i=1}^n \left( \mathbb{E}[X_i^2] + \{\hat{\mu}_x^{(t)}\}^2 - 2\hat{\mu}_x^{(t)} \mathbb{E}[X_i] \right),$$


$$\hat{\mu}_x^{(t+1)} = \frac{\sum_{i=1}^n \mathbb{E}[X_i]}{n}.$$

    (M.2) Set  $\hat{\Psi} = \hat{\Psi}^{(t+1)} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\theta}_1, \dots, \hat{\theta}_K, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_u^2, \hat{\mu}_x, \hat{\sigma}_x^2\}^{(t+1)}.$ 
end
```

4 Numerical examples

Our simulation example is similar to one of the examples of Bhadra and Carroll (2016). Let the true, nonlinear mean function be

$$g(x) = \frac{3 \sin(\pi x/2)}{1 + 2x^2\{I(x > 0) + 1\}},$$

where I denotes the indicator function. We generate X_i from $\text{Normal}(\mu_x, \sigma_x^2)$. To simulate data according to Equations (1-2), we use $n = 500, m_i = 2$ for all $i, \mu_x = 0, \sigma_x^2 = 1, \sigma_\epsilon^2 = 0.09$ and $\sigma_u^2 = 0.33$. Table 1 shows the true parameter values and the maximum likelihood

Parameter	True value	mle
μ_x	0	-0.01
σ_x^2	1	1.04
σ_u^2	0.33	0.28
σ_ϵ^2	0.09	0.11

Table 1: Parameter estimates by the E-M algorithm.

estimates found by the proposed E-M algorithm (Algorithm 1). We use 25 knots and place them evenly on the sample quantiles of $W_{i\bullet}$. We now provide some guidelines for choosing the initial values in the E-M algorithm. We use $\hat{\mu}_x^{(0)} = \bar{W}_{\bullet\bullet} = \sum_{i=1}^n W_{i\bullet} / \sum_{i=1}^n m_i$. Since $\{W_{ij}\}$ can be thought of as observations from a random one-factor ANOVA with n levels and total number of observations $m_\bullet = \sum_{i=1}^n m_i$, quite straightforward initial estimates of σ_u^2 and σ_x^2 can be obtained. If we define

$$MSE = \sum_{i=1}^n \sum_{j=1}^{m_i} (W_{ij} - \bar{W}_{i\bullet})^2 / (m_\bullet - n), \quad MSTR = \sum_{i=1}^n m_i (\bar{W}_{i\bullet} - \bar{W}_{\bullet\bullet})^2 / (n - 1),$$

then $\mathbb{E}(MSE) = \sigma_u^2$ and $\mathbb{E}(MSTR) = c\sigma_x^2 + \sigma_u^2$, where $c = (m_\bullet - \sum_{i=1}^n m_i^2 / m_\bullet) / (n - 1)$. Thus the initial estimates of $\{\hat{\sigma}_u^2\}^{(0)}$ and $\{\hat{\sigma}_x^2\}^{(0)}$ may be taken as

$$\{\hat{\sigma}_u^2\}^{(0)} = MSE, \quad \{\hat{\sigma}_x^2\}^{(0)} = \max\{(MSTR - MSE)/c, \alpha\},$$

where $\alpha > 0$ is a small constant, such as 0.01, that ensures a positive initial value for $\{\hat{\sigma}_x^2\}^{(0)}$. Our choice of initial values can be interpreted as simple, method of moments estimators of the parameters and similar guidelines were used by Berry et al. (2002). Moreover, we can obtain the initial estimates of β 's, θ 's and σ_ϵ^2 by fitting the regression of Equation (3) by ordinary least squares where we substitute $\overline{W}_{i\bullet}$ instead of the unobserved X_i . In our simulations, with these choices for the initial values, the E-M algorithm converges to the mle within 100 iterations. We also experiment with slightly different starting values, by adding mean zero Gaussian noise with variance 0.1 to the method of moments starting values listed above and the convergence of the algorithm appears to be stable. The total computer time required by our MATLAB implementation is around 2 minutes on a Macintosh with a 3.1 GHz Intel processor and 16 GB of RAM.

5 Discussion

We proposed an E-M algorithm for computing the maximum likelihood estimate in a measurement error model. Our results complement the semi-parametric technique of Carroll et al. (1999) and the Bayesian techniques of Berry et al. (2002) and Bhadra and Carroll (2016). We used a truncated polynomial spline of degree 1 as our choice of basis function. Using the results in Proposition 2 of Bhadra and Carroll (2016), it can be seen easily that a similar E-M scheme will still be a feasible if degree one B-spline bases are used instead. Its implementation should be considered future work.

Acknowledgements

The author thanks Raymond J. Carroll for providing feedback on the manuscript. The manuscript was considerably improved by the review comments. This material is based upon work supported by the National Science Foundation under Grant No. DMS-1613063.

A Derivation of the E-M algorithm

A.1 E-step

Note from Equation (6), in the $(t + 1)$ 'th E-step, it is only necessary to impute the following terms: $\mathbb{E}(X_i|\mathbf{Y}, \mathbf{W}, \Psi = \hat{\Psi}^{(t)})$, $\mathbb{E}(X_i^2|\mathbf{Y}, \mathbf{W}, \Psi = \hat{\Psi}^{(t)})$, $\mathbb{E}[B_k(X_i)|\mathbf{Y}, \mathbf{W}, \Psi = \hat{\Psi}^{(t)}]$, $\mathbb{E}[\{B_k(X_i)\}^2|\mathbf{Y}, \mathbf{W}, \Psi = \hat{\Psi}^{(t)}]$, $\mathbb{E}[X_i B_k(X_i)|\mathbf{Y}, \mathbf{W}, \Psi = \hat{\Psi}^{(t)}]$ and $\mathbb{E}[B_k(X_i)B_j(X_i)|\mathbf{Y}, \mathbf{W}, \Psi = \hat{\Psi}^{(t)}]$, where $\hat{\Psi}^{(0)}$ is an initial guess for the parameter values. Without risking ambiguity, we drop the conditioning variables from the following and let $\pi(x_i)$ denote the density of $(X_i|\mathbf{Y}, \mathbf{W}, \Psi = \hat{\Psi}^{(t)})$. Now using Proposition 1,

$$\begin{aligned}\mathbb{E}[X_i] &= \sum_{L=0}^K p_{iL} \mu_{iL}, \\ \mathbb{E}[X_i^2] &= \sum_{L=0}^K p_{iL} (\mu_{iL}^2 + \sigma_{iL}^2).\end{aligned}$$

For $\mathbb{E}[B_k(X_i)]$ note that

$$\begin{aligned}\mathbb{E}[B_k(X_i)] &= \mathbb{E}[(X_i - \omega_k)I(X_i > \omega_k)] \\ &= \int_{\omega_k}^{\infty} (x_i - \omega_k)\pi(x_i)dx_i \\ &= \int_{\omega_k}^{\infty} x_i\pi(x_i)dx_i - \omega_k \int_{\omega_k}^{\infty} \pi(x_i)dx_i \\ &= \sum_{L=k}^K p_{iL} \mu_{iL} - \omega_k \sum_{L=k}^K p_{iL}.\end{aligned}$$

Similarly, for $\mathbb{E}[X_i B_k(X_i)]$ note that

$$\begin{aligned}\mathbb{E}[X_i B_k(X_i)] &= \mathbb{E}[X_i(X_i - \omega_k)I(X_i > \omega_k)] \\ &= \int_{\omega_k}^{\infty} x_i(x_i - \omega_k)\pi(x_i)dx_i \\ &= \int_{\omega_k}^{\infty} x_i^2\pi(x_i)dx_i - \omega_k \int_{\omega_k}^{\infty} x_i\pi(x_i)dx_i \\ &= \sum_{L=k}^K p_{iL} (\mu_{iL}^2 + \sigma_{iL}^2) - \omega_k \sum_{L=k}^K p_{iL} \mu_{iL}.\end{aligned}$$

For $\mathbb{E}[\{B_k(X_i)\}^2]$ note that

$$\begin{aligned}
\mathbb{E}[\{B_k(X_i)\}^2] &= \mathbb{E}[(X_i - \omega_k)^2 I(X_i > \omega_k)] \\
&= \int_{\omega_k}^{\infty} (x_i - \omega_k)^2 \pi(x_i) dx_i \\
&= \int_{\omega_k}^{\infty} x_i^2 \pi(x_i) dx_i - 2\omega_k \int_{\omega_k}^{\infty} x_i \pi(x_i) dx_i + \omega_k^2 \int_{\omega_k}^{\infty} \pi(x_i) dx_i \\
&= \sum_{L=k}^K p_{iL} (\mu_{iL}^2 + \sigma_{iL}^2) - 2\omega_k \sum_{L=k}^K p_{iL} \mu_{iL} + \omega_k^2 \sum_{L=k}^K p_{iL}.
\end{aligned}$$

Finally, for $\mathbb{E}[B_k(X_i)B_j(X_i)]$ note that

$$\begin{aligned}
\mathbb{E}[B_k(X_i)B_j(X_i)] &= \mathbb{E}[(X_i - \omega_k)(X_i - \omega_j) I(X_i > \max\{\omega_k, \omega_j\})] \\
&= \int_{\max\{\omega_k, \omega_j\}}^{\infty} (x_i - \omega_k)(x_i - \omega_j) \pi(x_i) dx_i \\
&= \int_{\max\{\omega_k, \omega_j\}}^{\infty} x_i^2 \pi(x_i) dx_i - (\omega_k + \omega_j) \int_{\max\{\omega_k, \omega_j\}}^{\infty} x_i \pi(x_i) dx_i + \omega_k \omega_j \int_{\max\{\omega_k, \omega_j\}}^{\infty} \pi(x_i) dx_i \\
&= \sum_{L=\max\{k,j\}}^K p_{iL} (\mu_{iL}^2 + \sigma_{iL}^2) - (\omega_k + \omega_j) \sum_{L=\max\{k,j\}}^K p_{iL} \mu_{iL} + \omega_k \omega_j \sum_{L=\max\{k,j\}}^K p_{iL}.
\end{aligned}$$

Substituting these terms in Equation (6) completes the E-step.

A.2 M-step

Suppose the latent component of the complete data log likelihood has been imputed according to the conditional expectations computed in the E-step. In what follows, $f(\nu)$ denotes the terms containing parameter ν in the complete data log likelihood of Equation (6). Then we perform a conditional maximization of one parameter given the rest (i.e., a cyclic coordinate descent) for the $(t+1)$ 'th M-step as follows:

1.

$$f(\beta_0) = -(2\sigma_\varepsilon^2)^{-1} \sum_{i=1}^n \left(-2\beta_0 Y_i + \beta_0^2 + 2\beta_0 \beta_1 \mathbb{E}[X_i] + 2\beta_0 \sum_{k=1}^K \theta_k \mathbb{E}[B_k(X_i)] \right).$$

Equating $f'(\beta_0) = 0$ we get

$$\hat{\beta}_0^{(t+1)} = \frac{\sum_{i=1}^n \left(Y_i - \hat{\beta}_1^{(t)} \mathbb{E}[X_i] - \sum_{k=1}^K \hat{\theta}_k^{(t)} \mathbb{E}[B_k(X_i)] \right)}{n}.$$

2.

$$f(\beta_1) = -(2\sigma_\varepsilon^2)^{-1} \sum_{i=1}^n \left(-2\beta_1 Y_i \mathbb{E}[X_i] + \beta_1^2 \mathbb{E}[X_i^2] + 2\beta_0 \beta_1 \mathbb{E}[X_i] + 2\beta_1 \sum_{k=1}^K \theta_k \mathbb{E}[X_i B_k(X_i)] \right).$$

Equating $f'(\beta_1) = 0$ we get

$$\hat{\beta}_1^{(t+1)} = \frac{\sum_{i=1}^n \left(Y_i \mathbb{E}[X_i] - \hat{\beta}_0^{(t+1)} \mathbb{E}[X_i] - \sum_{k=1}^K \hat{\theta}_k^{(t)} \mathbb{E}[X_i B_k(X_i)] \right)}{\sum_{i=1}^n \mathbb{E}[X_i^2]}.$$

3. For $k = 1, \dots, K$

$$\begin{aligned} f(\theta_k) = & -(2\sigma_\varepsilon^2)^{-1} \sum_{i=1}^n \left(-2\theta_k Y_i \mathbb{E}[B_k(X_i)] + \theta_k^2 \mathbb{E}[\{B_k(X_i)\}^2] + 2\beta_1 \theta_k \mathbb{E}[X_i B_k(X_i)] \right. \\ & \left. + 2\beta_0 \theta_k \mathbb{E}[B_k(X_i)] + 2\theta_k \sum_{j=1}^{k-1} \theta_j \mathbb{E}[B_j(X_i) B_k(X_i)] \right). \end{aligned}$$

Equating $f'(\theta_k) = 0$ we get

$$\hat{\theta}_k^{(t+1)} = \frac{\sum_{i=1}^n \left(Y_i \mathbb{E}[B_k(X_i)] - \hat{\beta}_0^{(t+1)} \mathbb{E}[B_k(X_i)] - \hat{\beta}_1^{(t+1)} \mathbb{E}[X_i B_k(X_i)] - \sum_{j=1}^{k-1} \hat{\theta}_j^{(t+1)} \mathbb{E}[B_j(X_i) B_k(X_i)] \right)}{\sum_{i=1}^n \mathbb{E}[\{B_k(X_i)\}^2]}$$

4. Let

$$\begin{aligned} A_{\sigma_\varepsilon^2}^{(t+1)} = & \sum_{i=1}^n \left(Y_i^2 - 2\beta_0 Y_i - 2\beta_1 Y_i \mathbb{E}[X_i] - 2Y_i \sum_{k=1}^K \theta_k \mathbb{E}[B_k(X_i)] + \beta_0^2 + \beta_1^2 \mathbb{E}[X_i^2] \right. \\ & \left. + \sum_{k=1}^K \theta_k^2 \mathbb{E}[\{B_k(X_i)\}^2] + 2\beta_0 \beta_1 \mathbb{E}[X_i] + 2\beta_0 \sum_{k=1}^K \theta_k \mathbb{E}[B_k(X_i)] + 2\beta_1 \sum_{k=1}^K \theta_k \mathbb{E}[X_i B_k(X_i)] \right) \end{aligned}$$

$$+2 \sum_{k=1}^K \sum_{j=1}^{k-1} \theta_k \theta_j \mathbb{E}[B_k(X_i) B_j(X_i)] \Big),$$

where we use $\beta_0 = \hat{\beta}_0^{(t+1)}$, $\beta_1 = \hat{\beta}_1^{(t+1)}$ and $\theta_k = \hat{\theta}_k^{(t+1)}$. Then

$$f(\sigma_\varepsilon^2) = -\frac{A_{\sigma_\varepsilon^2}^{(t+1)}}{2\sigma_\varepsilon^2} - (n/2)\log(\sigma_\varepsilon^2),$$

and using $f'(\sigma_\varepsilon^2) = 0$ we have

$$\{\hat{\sigma}_\varepsilon^2\}^{(t+1)} = \frac{A_{\sigma_\varepsilon^2}^{(t+1)}}{n}.$$

5. Let

$$A_{\sigma_u^2}^{(t+1)} = \sum_{i=1}^n \left(\sum_{j=1}^{m_i} W_{ij}^2 + m_i \mathbb{E}[X_i^2] - 2W_{i\bullet} \mathbb{E}[X_i] \right).$$

Then

$$f(\sigma_u^2) = -\frac{A_{\sigma_u^2}^{(t+1)}}{2\sigma_u^2} - (M/2)\log(\sigma_u^2).$$

and using $f'(\sigma_u^2) = 0$ we have

$$\{\hat{\sigma}_u^2\}^{(t+1)} = \frac{A_{\sigma_u^2}^{(t+1)}}{M}.$$

6. Let

$$A_{\sigma_x^2}^{(t+1)} = \sum_{i=1}^n \left(\mathbb{E}[X_i^2] + \{\hat{\mu}_x^{(t)}\}^2 - 2\hat{\mu}_x^{(t)} \mathbb{E}[X_i] \right).$$

Then

$$f(\sigma_x^2) = -\frac{A_{\sigma_x^2}^{(t+1)}}{2\sigma_x^2} - (n/2)\log(\sigma_x^2),$$

and using $f'(\sigma_x^2) = 0$ we have

$$\{\hat{\sigma}_x^2\}^{(t+1)} = \frac{A_{\sigma_x^2}^{(t+1)}}{n}.$$

7. Finally, we have

$$\hat{\mu}_x^{(t+1)} = \frac{\sum_{i=1}^n \mathbb{E}[X_i]}{n}.$$

The E and M steps are repeated until the convergence of the algorithm.

References

- Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association* **97**, 160–169.
- Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B* **48**, 259–302.
- Bhadra, A. and Carroll, R. J. (2016). Exact sampling of the unobserved covariates in Bayesian spline models for measurement error problems. *Statistics and Computing* **26**, 827–840.
- Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86**, 541–554.
- Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* **89**, 1314–1328.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* **39**, 1–38.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.