ISSN 1973-2201 STATISTICA, anno LXXXV, n. 1, 2025
DOI: https://doi.org/10.60923/issn.1973-2201/21497

AN ALGORITHM FOR ESTIMATING MEASUREMENT ERROR MODELS EMPLOYING SPLINE APPROXIMATION

Yashi Srivastava [1]
Decision Sciences Area, Indian Institute of Management Lucknow, Lucknow, India
Gaurav Garg
Decision Sciences Area, Indian Institute of Management Lucknow, Lucknow, India

SUMMARY

An algorithm has been derived for finding maximum likelihood estimates of a measurement error model when the B-spline basis function of degree one is used to model the true regression function. In a measurement error model, the true variable is not observed. Hence the likelihood function is not completely known. This in turn induces difficulty in the maximization of the likelihood function. To address this problem, the Expectation-Maximization algorithm has been used to derive the results. We also illustrate the algorithm using simulation and a real life application.

Keywords: B-spline basis; Expectation-Maximization; Maximum likelihood; Measurement error.

1. INTRODUCTION

Measurement errors are ubiquitous in nearly all applications of finance, economics, engineering, medicine, agriculture and others. It is important to be wary of measurement errors, as their negligence could instigate bias while estimation (Carroll *et al.*, 2006). In the case where the true values of independent variables are not observed, the likelihood function of the model is not entirely known. This leads to the intractability of the maximum likelihood approach in estimating measurement error models. This complexity has resulted in a relatively limited exploration of the maximum likelihood estimation procedure in these models.

In this work, we propose an algorithm to find maximum likelihood estimates of a measurement error model when the true regression function is modeled using B-spline basis of degree one. The basis has an advantage over other basis functions due to its appealing properties, which makes it useful both computationally and theoretically (see Lyche *et al.*, 2018, for details).

[1] Corresponding Author. E-mail: phd21008@iiml.ac.in

There is substantial work on measurement error models in parametric and nonparametric paradigms. In the context of a parametric measurement error model, the estimation procedure by Cook and Stefanski (1994) known as SIMEX (Simulation-Extrapolation) integrates the parametric bootstrap with the method of moments. Carroll *et al.* (1999) illustrate the implementation of SIMEX in nonparametric regression and derive its asymptotic properties for kernel regression. Ayub *et al.* (2022) suggest an alternative procedure based on the conditional expectation of the target variable, leading to an estimation equation. Contributing to the nonparametric framework, Li and Vuong (1998) recommend a method based on the empirical characteristic distribution to estimate the distribution of latent variables and errors.

In classical measurement error models, identifiability is ensured if either the measurement error distribution is completely known or replicated measurements are available, which allow the measurement error variance to be estimated (Carroll *et al.*, 2006). However, in semiparametric and nonparametric paradigms, models may be identifiable even when replicated data are not available. Butucea and Matias (2005) discussed the estimation of the signal density in a semiparametric convolution model with normally distributed measurement errors of unknown variance. They proposed an estimation procedure under the assumption that the signal density is ordinary smooth (e.g., laplace distibution). Similarly Meister (2006) argued that, under the smoothness assumption of the density of the unobserved variable, the model is identifiable. They also noted that, if both the measurement error density and the true signal are super smooth (e.g., normal distribution), the problem becomes unidentifiable, as it is impossible to recover the component variances.

Bhadra (2017) gives Expectation-Maximization (EM) (Dempster *et al.*, 1977) based algorithm to estimate parameters of a measurement error model when the true regression function is approximated by a degree one truncated polynomial spline basis.

Bhadra and Carroll (2016) demonstrate that, when considering a B-spline of degree one, the complete conditional distribution of the covariate with error can be modeled as a mixture of truncated normal distributions. This allows the formation of Gibbs sampler for the unobserved variable. Motivated by the work of Bhadra (2017) and using the distributional result of Bhadra and Carroll (2016) for the unobserved covariate, we propose an EM algorithm based estimate for the parameters associated with degree one B-spline basis. The simulation study and real life application confirm the applicability of the proposed algorithm over the estimation method that naively assumes the observed covariate as error free.

The structure of the paper is as follows. In Section 2, we provide the model and the corresponding likelihood function. Section 3 defines the basis function and states a key proposition that serves as a fundamental component of our algorithm. The proposed algorithm is detailed in Section 4. A simulation study is presented in Section 5, followed by an application in Section 6. Conclusions and final remarks are provided in Section 7. Detailed derivations are included in the Appendix.

2. Model and the likelihood function

Let us consider the following measurement error model, as discussed by Bhadra and Carroll (2016):

$$\begin{aligned} Y_i &= g(X_i) + \epsilon_i, \\ W_{ij} &= X_i + U_{ij}, \quad j = 1, \ldots, m_i,\ i = 1, \ldots, n, \end{aligned} \tag{1}$$

where n denotes sample size and m_i is the number of replications corresponding to the i^{th} sample.

Let $\mathscr{Y}_i = (Y_i, W_{i1}, W_{i2}, \ldots, W_{im_i})$ be the observed data corresponding to the i^{th} sample. The variable X_i is unobserved and, therefore, latent variable.
We assume that X_i is a normal random variable with mean μ_x and variance σ_x^2. The errors ϵ_i's are independent and identically distributed normal random variables with mean 0 and variance σ_ϵ^2. Additionally, U_{ij} is assumed to be normally distributed with mean 0 and variance σ_u^2. It is further assumed that U_{ij} and ϵ_i are uncorrelated.

The unknown function $g(\cdot)$ represents the true relationship between Y_i and X_i. Assuming that $g(\cdot)$ is smooth, we can model it using the B-spline basis function of degree one, as suggested by Eilers and Marx (1996).

Suppose that $\{\widetilde{B}_k(X_i);\ k = 1, 2, \ldots, (K+1)\}$ are the $(K+1)$ basis functions. This leads us to the following model:

$$Y_i = \beta_0 + \beta_1 X_i + \sum_{k=1}^{k=K+1} \widetilde{B}_k(X_i)\theta_k + \epsilon_i, \tag{2}$$

$$W_{ij} = X_i + U_{ij}, \quad j = 1, \ldots, m_i,\ i = 1, \ldots, n, \tag{3}$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, $U_{ij} \sim N(0, \sigma_u^2)$, $X_i \sim N(\mu_x, \sigma_x^2)$. The set of parameters is $\psi = (\beta_0, \beta_1, \theta_1, \theta_2, \ldots, \theta_{K+1}, \sigma_\epsilon^2, \sigma_u^2, \mu_x, \sigma_x^2)$.

Define $M = \sum_{i=1}^{n} m_i$, $W_{i\cdot} = \sum_{j=1}^{m_i} W_{ij}$, $\overline{W}_{i\cdot} = W_{i\cdot}/m_i$. Our objective is to find the maximum likelihood estimates of ψ when the basis function is approximated using a degree one B-spline basis function. Similar to the approach adopted by Bhadra (2017),

the complete data log likelihood for the model in Equations (2) and (3) is given as

$$\begin{aligned}\mathscr{L}(\psi|\mathscr{Y}) &= -(2\sigma_\epsilon^2)^{-1}\sum_{i=1}^{n}\Big\{Y_i^2-2\beta_0 Y_i-2\beta_1 X_i Y_i-2Y_i\sum_{k=1}^{K+1}\theta_k\widetilde{B}_k(X_i)+\beta_0^2\\ &\quad+\beta_1^2X_i^2+\sum_{k=1}^{K+1}\theta_k^2\big[\widetilde{B}_k(X_i)\big]^2+2\beta_0\beta_1X_i+2\beta_0\sum_{k=1}^{K+1}\theta_k\widetilde{B}_k(X_i)\\ &\quad+2\beta_1\sum_{k=1}^{K+1}\theta_kX_i\widetilde{B}_k(X_i)+2\sum_{k=1}^{K+1}\sum_{j=1}^{k-1}\theta_k\theta_j\widetilde{B}_k(X_i)\widetilde{B}_j(X_i)\Big\}-(n/2)\log(\sigma_\epsilon^2)\\ &\quad-(2\sigma_u^2)^{-1}\sum_{i=1}^{n}\Big(\sum_{j=1}^{m_i}W_{ij}^2+m_iX_i^2-2W_{i.}X_i\Big)-(M/2)\log(\sigma_u^2)\\ &\quad-(2\sigma_x^2)^{-1}\Big(\sum_{i=1}^{n}X_i^2+n\mu_x^2-2\mu_x\sum_{i=1}^{n}X_i\Big)-(n/2)\log(\sigma_x^2). \qquad (4)\end{aligned}$$

If X_is were observed or $\mathscr{L}$ were completely known, the problem of estimating the model in Equations (2) and (3) would straightforwardly involve the maximization of $\mathscr{L}$. Since $\mathscr{L}$ is not fully known, we use the EM algorithm to obtain the model estimates, as explained in Section 4.

3. B-SPLINE BASIS FUNCTION

In this section, we introduce the B-spline basis functions and state a key proposition that will be utilized in the development of our proposed algorithm in the next section.

The basis function is free from the intercept and X_i and hence, $\beta_0=\beta_1=0$ in the model defined in Eq. (2).

Let $(\xi_0,\xi_1,\xi_2,\ldots,\xi_{K-1},\xi_K)$ be the knots on which the B-spline basis is defined. The set of $(K+1)$ B-spline basis functions of degree one (or the linear B-spline), $\widetilde{B}(x)=\{\widetilde{B}_1(x),\ldots,\widetilde{B}_{K+1}(x)\}$ is defined as

$$\begin{aligned}\widetilde{B}_1(x) &= \frac{\xi_1-x}{\xi_1-\xi_0}I(\xi_0\le x<\xi_1),\\ \widetilde{B}_{K+1}(x) &= \frac{x-\xi_{K-1}}{\xi_K-\xi_{K-1}}I(\xi_{K-1}\le x<\xi_K),\\ \widetilde{B}_k(x) &= \frac{x-\xi_{k-2}}{\xi_{k-1}-\xi_{k-2}}I(\xi_{k-2}\le x<\xi_{k-1})+\frac{\xi_k-x}{\xi_k-\xi_{k-1}}I(\xi_{k-1}\le x<\xi_k),\end{aligned}$$

for $k=2,3,\ldots,K$. Here $I(\xi_{k-1}\le x<\xi_k)$ is the indicator function which takes value 1 for all $\xi_{k-1}\le x<\xi_k$ and 0 otherwise.

Now we state a result, given by Bhadra and Carroll (2016), which is instrumental in computing the expected values required in the Exp step of the EM algorithm explained in Section 4.

PROPOSITION 1. *Suppose the true regression function $g(\cdot)$ in Eq. (1) is modeled using the degree one B-spline basis. Assume that $\xi_{-1} \to -\infty$, $\xi_{K+1} \to \infty$ and let $\theta_o = \theta_{K+2} = 0$. For each interval (ξ_L, ξ_{L+1}), we define the following quantities:*

$$\zeta_{1iL} = \left\{ (\sigma_\epsilon^2)^{-1} \left(\frac{\theta_{L+2} - \theta_{L+1}}{\xi_{L+1} - \xi_L} \right)^2 + m_i/(\sigma_u^2) + 1/(\sigma_x^2) \right\}^{-1},$$

$$\zeta_{2iL} = -(2\sigma_\epsilon^2)^{-1} \left\{ -2 \left(\frac{\theta_{L+2} - \theta_{L+1}}{\xi_{L+1} - \xi_L} \right) Y_i + 2 \left(\frac{\xi_{L+1}\theta_{L+1} - \xi_L \theta_{L+2}}{\xi_{L+1} - \xi_L} \right) \left(\frac{\theta_{L+2} - \theta_{L+1}}{\xi_{L+1} - \xi_L} \right) \right\} + (W_{i\cdot}/\sigma_u^2 + \mu_x/\sigma_x^2),$$

$$\zeta_{3iL} = -(2\sigma_\epsilon^2)^{-1} \left\{ -2Y_i \frac{\xi_{L+1}\theta_{L+1} - \xi_L \theta_{L+2}}{\xi_{L+1} - \xi_L} + \left(\frac{\xi_{L+1}\theta_{L+1} - \xi_L \theta_{L+2}}{\xi_{L+1} - \xi_L} \right)^2 \right\}.$$

Now we define the standardized boundaries b_{iL} and a_{iL} as

$$b_{iL} = \frac{(\xi_{L+1} - \zeta_{1iL}\zeta_{2iL})}{\sqrt{\zeta_{1iL}}}, \quad a_{iL} = \frac{(\xi_L - \zeta_{1iL}\zeta_{2iL})}{\sqrt{\zeta_{1iL}}}.$$

Let $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal distribution. We define the normalizing constant

$$D_i = \left[\sum_{L=-1}^{K} \{\Phi(b_{iL}) - \Phi(a_{iL})\} \exp\{(1/2)\log(\zeta_{1iL}) + \zeta_{3iL} + \zeta_{1iL}(\zeta_{2iL})^2/2\} \right]^{-1}$$

and the mixing probabilities

$$p_{iL} = D_i \{\Phi(b_{iL}) - \Phi(a_{iL})\} \exp\{(1/2)\log(\zeta_{1iL}) + \zeta_{3iL} + \zeta_{1iL}(\zeta_{2iL})^2/2\}.$$

Then, the complete conditional distribution of X_i is a mixture of truncated normal distributions on the intervals $[\xi_L, \xi_{L+1}]$ for $L = -1, \ldots, K$ with means $\mu_{iL} = \zeta_{1iL}\zeta_{2iL}$, variances $\sigma_{iL}^2 = \zeta_{1iL}$ and mixing probabilities p_{iL}.

4. The proposed algorithm

Algorithm 2.

Data: $\mathscr{Y}_i = (Y_i, W_{i1}, W_{i2}, \ldots, W_{im_i})$. *Parameters: Knot points* $\xi_0, \xi_1, \xi_2, \ldots, \xi_K, \xi_{K+1}$. *Initial parameters* $\hat{\psi}^{(0)}$. *Number of iterations* $\boldsymbol{T}$.

For $t = 0$ *to* $t = T-1$ *do*

Exp step:

Compute p_{iL}, μ_{iL}, σ^2_{iL} *for* $L = -1, 0, \ldots, K$ *using Proposition 1.*
(Note: $k = 2, 3, \ldots, K$ *and* $j = 1, 2, \ldots, K$ *in Exp step.)*

$$
\begin{aligned}
E(X_i) &= \sum_{L=-1}^{K} p_{iL}\mu_{iL}, \quad E(X_i^2) = \sum_{L=-1}^{K} p_{iL}(\mu_{iL}^2 + \sigma_{iL}^2),\\
E\big[\widetilde{B}_1(X_i)\big] &= \frac{p_{i0}(\xi_1 - \mu_{i0})}{(\xi_1 - \xi_0)},\\
E\big[\widetilde{B}_k(X_i)\big] &= \frac{p_{i(k-2)}}{(\xi_{k-1} - \xi_{k-2})}[\mu_{i(k-2)} - \xi_{k-2}] + \frac{p_{i(k-1)}}{(\xi_k - \xi_{k-1})}[\xi_k - \mu_{i(k-1)}],\\
E\big[\widetilde{B}_{K+1}(X_i)\big] &= \frac{p_{i(K-1)}\big[\mu_{i(K-1)} - \xi_{K-1}\big]}{(\xi_K - \xi_{K-1})},\\
E\big\{\big[\widetilde{B}_1(X_i)\big]^2\big\} &= \frac{p_{i0}}{(\xi_1 - \xi_0)^2}(\mu_{i0}^2 + \sigma_{i0}^2 - 2\xi_1\mu_{i0} + \xi_1^2),\\
E\big\{\big[\widetilde{B}_k(X_i)\big]^2\big\} &= \frac{p_{i(k-2)}}{(\xi_{k-1} - \xi_{k-2})^2}[\mu_{i(k-2)}^2 + \sigma_{i(k-2)}^2 - 2\xi_{k-2}\mu_{i(k-2)} + \xi_{k-2}^2]\\
&\quad + \frac{p_{i(k-1)}}{(\xi_k - \xi_{k-1})^2}[\mu_{i(k-1)}^2 + \sigma_{i(k-1)}^2 - 2\xi_k\mu_{i(k-1)} + \xi_k^2],\\
E\big\{\big[\widetilde{B}_{K+1}(X_i)\big]^2\big\} &= \frac{p_{i(K-1)}}{(\xi_K - \xi_{K-1})^2}\big[\mu_{i(K-1)}^2 + \sigma_{i(K-1)}^2 - 2\xi_{K-1}\mu_{i(K-1)} + \xi_{k-1}^2\big],\\
E\big[X_i\widetilde{B}_1(X_i)\big] &= \frac{p_{i0}}{(\xi_1 - \xi_0)}(\mu_{i0}\xi_1 - \mu_{i0}^2 - \sigma_{i0}^2),\\
E\big[X_i\widetilde{B}_k(X_i)\big] &= \frac{p_{i(k-2)}}{(\xi_{k-1} - \xi_{k-2})}[\mu_{i(k-2)}^2 + \sigma_{i(k-2)}^2 - \xi_{k-2}\mu_{i(k-2)}]\\
&\quad + \frac{p_{i(k-1)}}{(\xi_k - \xi_{k-1})}[-\mu_{i(k-1)}^2 - \sigma_{i(k-1)}^2 + \xi_k\mu_{i(k-1)}],\\
E\big[X_i\widetilde{B}_{K+1}(X_i)\big] &= \frac{p_{i(K-1)}}{(\xi_K - \xi_{K-1})}\big[\mu_{i(K-1)}^2 + \sigma_{i(K-1)}^2 - \xi_{K-1}\mu_{i(K-1)}\big],\\
E\big[\widetilde{B}_j(X_i)\widetilde{B}_{j+1}(X_i)\big] &= \frac{p_{i(j-1)}}{(\xi_j - \xi_{j-1})^2}\big[\mu_{i(j-1)}(\xi_j + \xi_{j-1}) - \xi_{j-1}\,\xi_j - (\mu_{i(j-1)}^2 + \sigma_{i(j-1)}^2)\big].
\end{aligned}
$$

Max step

(Note: $k = 1, 2, 3, \ldots, (K+1)$ in Max step.)

Set

$$
\begin{aligned}
\hat{\theta}_k^{(t+1)} &= \frac{\sum_{i=1}^{n}\{Y_i E[\widetilde{B}_k(X_i)] - \sum_{j=1}^{k-1}\hat{\theta}_j^{(t+1)} E[\widetilde{B}_j(X_i)\widetilde{B}_k(X_i)]\}}{\sum_{i=1}^{n} E\{[\widetilde{B}_k(X_i)]^2\}}, \\
\hat{\mu}_x^{(t+1)} &= \frac{\sum_{i=1}^{n} E(X_i)}{n}, \\
(\hat{\sigma}_\epsilon^2)^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i^2 - 2Y_i\sum_{k=1}^{K+1}\theta_k E[\widetilde{B}_k(X_i)] + \sum_{k=1}^{K+1}\theta_k^2 E\{[\widetilde{B}_k(X_i)]^2\} \\
&\quad + 2\sum_{k=1}^{K+1}\sum_{j=1}^{k-1}\theta_k\theta_j E[\widetilde{B}_k(X_i)\widetilde{B}_j(X_i)]\Big), \\
(\hat{\sigma}_u^2)^{(t+1)} &= \frac{\sum_{i=1}^{n}[\sum_{j=1}^{m_i} W_{ij}^2 + m_i E(X_i^2) - 2W_{i\cdot}E(X_i)]}{M}, \\
(\hat{\sigma}_x^2)^{(t+1)} &= \frac{\sum_{i=1}^{n}[E(X_i^2) + \{\hat{\mu}_x^{(t)}\}^2 - 2\hat{\mu}_x^{(t)}E(X_i)]}{n}.
\end{aligned}
$$

end.

The **Exp step** involves finding expectation of the log likelihood for the whole data, given the observed data and the current parameter estimates. Taking the expected value of the complete data log likelihood in Eq. (4), given the observed data $\mathcal{Y}_i = (Y_i, W_{i1}, \ldots, W_{im_i})$, since $\beta_0 = \beta_1 = 0$ for the linear B-spline basis, we get

$$
\begin{aligned}
E[\mathcal{L}(\psi)|\mathcal{Y}_i] = -(2\sigma_\epsilon^2)^{-1}\sum_{i=1}^{n} E\Big[\Big(\Big\{Y_i^2 - 2Y_i\sum_{k=1}^{K+1}\theta_k\widetilde{B}_k(X_i) + \sum_{k=1}^{K+1}\theta_k^2[\widetilde{B}_k(X_i)]^2 \\
+ 2\sum_{k=1}^{K+1}\sum_{j=1}^{k-1}\theta_k\theta_j\widetilde{B}_k(X_i)\widetilde{B}_j(X_i)\Big\} - (n/2)\log(\sigma_\epsilon^2) \\
-(2\sigma_u^2)^{-1}\sum_{i=1}^{n}\Big(\sum_{j=1}^{m_i} W_{ij}^2 + m_i X_i^2 - 2W_{i\cdot}X_i\Big) - (M/2)\log(\sigma_u^2) \\
-(2\sigma_x^2)^{-1}\Big(\sum_{i=1}^{n} X_i^2 + n\mu_x^2 - 2\mu_x\sum_{i=1}^{n} X_i\Big) - (n/2)\log(\sigma_x^2)\Big)\Big|\mathcal{Y}_i\Big].
\end{aligned}
$$

Using Proposition 1, for the B-spline basis of degree one, the Exp step reduces to computing the conditional expectations of X_i, X_i^2, $\widetilde{B}_k(X_i)$, $\widetilde{B}_k(X_i)^2$, $X_i\widetilde{B}_k(X_i)$ and $\widetilde{B}_k(X_i)\widetilde{B}_j(X_i)$. These are outlined in Algorithm 2 and the detailed derivations are provided in Appendix A.

After computing the unknown terms in the complete data log likelihood, we proceed to the **Max step** by maximizing the likelihood function with respect to the parameter

of interest, conditional on the remaining parameters. This completes one full iteration of the EM algorithm. Further details on the Max step can be found in Bhadra (2017).

It is important to note that the algorithm requires a judicious choice of initial parameter estimates at $t = 0$. These can be conveniently obtained through the method of moments. In particular, the overall mean of W_{ij} serves as the initial estimate of μ_x. The variability of the replicate measurements of the observed covariate around the subject means provides an initial estimate for σ_u^2, while the variability of the subject means of the observed covariate around the overall mean gives the estimate of σ_x^2. The regression parameters are initialized using ordinary least square (OLS) regression of the response on the basis function, where the subject-level mean of the observed covariate is used as the explanatory variable. The detailed implementation of these initializations is provided in the simulation section.

At iteration $t = 1$, the parameter estimates obtained at $t = 0$ are used as initial values to perform the Exp step followed by the Max step. The resulting estimates at $t = 1$ serve as the starting values for iteration $t = 2$, and so on. The iterative procedure continues until convergence of the parameter estimates is achieved.

Note: It is important to discuss the convergence of the proposed algorithm. By applying Jensen's inequality to the complete data log likelihood and using the maximization step of the proposed algorithm, the observed data log likelihood is a non decreasing function of the parameters. Moreover, since the likelihood function is bounded above for fixed sample size and spline basis, the algorithm converges to a finite limit. For further details, see Wu (1983).

5. Simulation

We illustrate the proposed algorithm and compare its performance with the algorithm introduced by Bhadra (2017) (hereafter referred to as Bhadra's), which employs truncated polynomial spline of degree one as the basis function, as well as with another estimator that does not account for the presence of measurement error. Since the latter treats the observed explanatory variable as the true one without correcting for the error, we term it as the 'naive' estimator. It is also worth noting that, since our proposed algorithm and Bhadra's estimator use a B-spline basis function of degree one and a truncated polynomial spline of degree one, respectively, we require separate naive estimators corresponding to each of these approaches.

We consider the following true mean function, as in Bhadra and Carroll (2016)

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

where I denotes the indicator function. We generate $n = 50$ observations with $m_i = 2$ using the following parameter values: $\mu_x = 2$, $\sigma_x^2 = 1$, $\sigma_\epsilon^2 = 0.09$, $\sigma_u^2 = 0.03$.

Table 1 presents the average values of the parameter estimates corresponding to 10000 replications, computed using the proposed algorithm (EM based), Bhadra's and

their corresponding naive estimators, EM_{naive} and $\text{Bhadra}_{\text{naive}}$. The corresponding empirical mean square errors (MSEs) are also reported. It should be noted that, since the naive estimator attributes all variability in the observed explanatory variable to variability in the true covariate, it is not possible to define a naive estimator for σ_u^2.

Notably, the empirical MSEs for the EM based estimates and Bhadra's estimator are generally lower than their naive counterparts across all parameters. This suggests that the proposed algorithm and Bhadra's estimator more effectively capture the underlying model structure than the naive estimates.

TABLE 1
Parameter estimates and their empirical MSEs for the simulation study.

Parameter	True Value	Parameter Estimates				Empirical MSE			
		EM based	EM_{naive}	Bhadra's	$\text{Bhadra}_{\text{naive}}$	EM based	EM_{naive}	Bhadra's	$\text{Bhadra}_{\text{naive}}$
μ_x	2	1.9995	1.9901	1.9676	1.8538	0.020	0.021	0.012	0.022
σ_x^2	1	1.0131	0.7316	1.0473	0.9556	0.041	0.114	0.031	0.217
σ_ϵ^2	0.09	0.1367	0.0007	0.1802	0.1660	0.003	0.008	0.018	0.031
σ_u^2	0.03	0.0803	–	0.0326	–	0.003	–	0.000	–

We use 10 knots placed evenly on the sample quantiles of $W_{i\cdot}$. The procedure for the choice of initial values of the parameter estimates follows that of Bhadra (2017) and is described below.

Let $\overline{W}_{\cdot\cdot} = \sum_{i=1}^n W_{i\cdot}/\sum_{i=1}^n m_i$ denote the overall mean of the observed covariate, $m_\cdot = \sum_{i=1}^n m_i$ be the total number of observations.

Define

$$A = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(W_{ij} - \overline{W}_{i\cdot})^2/(m_\cdot - n), \quad B = \sum_{i=1}^{n} m_i(\overline{W}_{i\cdot} - \overline{W}_{\cdot\cdot})^2/(n-1),$$

the initial estimates are given as: $\hat{\mu}_x^{(0)} = \overline{W}_{\cdot\cdot}$ and $(\hat{\sigma}_u^2)^{(0)} = A$. Also for $\alpha > 0$, $(\hat{\sigma}_x^2)^{(0)} = max[(B-A)/c, \alpha]$. Here α ensures a positive initial value for $(\hat{\sigma}_x^2)^{(0)}$ and $c = (m_\cdot - \sum_{i=1}^n m_i^2/m_\cdot)/(n-1)$. Further, the initial estimates of the B-spline coefficients θ's and error variance σ_ϵ^2, denoted by $\hat{\theta}^{(0)}$ and $(\hat{\sigma}_\epsilon^2)^{(0)}$ respectively, are obtained by fitting an OLS regression of Y on $\widetilde{B}_1(\cdot), \widetilde{B}_2(\cdot), \ldots, \widetilde{B}_{K+1}(\cdot)$, using $\overline{W}_{i\cdot}$ in place of X_i in Eq. (2).

The naive estimates are computed by ignoring the measurement error. Specifically, for EM_{naive}, the estimates for μ_x and σ_x^2 are given by $\overline{W}_{\cdot\cdot}$ and $Var(\overline{W}_{i\cdot})$, respectively. The naive estimate for σ_ϵ^2 is obtained using the OLS regression of Y on the basis function evaluated at $\overline{W}_{i\cdot}$, replacing X_i in Eq. (2). For further details on Bhadra's estimator, please see Bhadra (2017).

6. Real life application

We utilize the bloodpressure dataset available in mecor package of `R`. This dataset contains information on creatinine levels and four replicated measurements of blood pres-

sure for 450 pregnant women. Since measurement errors are inherent in biomedical data, this dataset provides a relevant testing ground for evaluating the performance of our algorithm.

We focus on modeling the relationship between the blood pressure (as the covariate subject to measurement error) and creatinine level (as the response variable) using model in Equations (2) and (3). The systolic blood pressure variables in the dataset, sbp60 and sbp90, are used as the first and second replicates, respectively. We use these two variables to find the estimate of the variance of measurement error which is given by

$$\hat{\sigma}_u^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(W_{ij} - \overline{W}_{i\cdot})^2.$$

For details, see Carroll *et al.* (2006).

To ensure that normality assumptions are reasonably satisfied, Box-Cox transformations are applied prior to model fitting on the response variable. Following the approach used in the simulation study, the performance of our EM based spline estimation procedure is compared against that of Bhadra's estimator and the naive estimator, which ignores measurement errors. This comparison highlights the advantages of properly accounting for error in variables in real-world settings, as achieved by our proposed algorithm.

Table 2 reports the parameter estimates and their empirical MSEs. The empirical MSEs for the EM based estimates and Bhadra's estimates are lower than their corresponding naive estimates, suggesting that the EM based and Bhadra's approach outperform their corresponding naive method.

TABLE 2
Parameter estimates and their empirical MSEs for the bloodpressure dataset.

Parameter	True Value	Parameter Estimates				Empirical MSE			
		EM based	EM_{naive}	Bhadra's	$\text{Bhadra}_{\text{naive}}$	EM based	EM_{naive}	Bhadra's	$\text{Bhadra}_{\text{naive}}$
μ_x	–	124.8501	123.7123	120.4418	120.4378	–	–	–	–
σ_x^2	48.2196	50.7423	62.3649	53.8691	68.8837	40.085	340.742	31.917	427.004
σ_ϵ^2	91.5013	90.3106	136.5904	89.7997	90.5875	3.085	10.491	0.835	2.895
σ_u^2	28.2902	20.3380	–	30.1785	–	10.865	–	3.565	–

It is important to note that, unlike in simulation studies, the true model parameters are unknown in real world applications. Consequently, it is not possible to directly compute empirical MSEs. Instead, we determine 'reference values' for the model parameters of Equations (2) and (3) against which the empirical MSEs are calculated. Specifically, the variance of the transformed response variable is used as a reference value for σ_ϵ^2, while the reference value for σ_x^2 is given by the method of moments estimator, $\text{Var}(\overline{W}_{i\cdot}) - \frac{\hat{\sigma}_u^2}{2}$ (Carroll *et al.*, 2006). It should be noted that, since the X_is are unobserved and the naive estimator of μ_x is $\overline{W}_{\cdot\cdot}$ itself, we are not in a position to suggest a reference value for μ_x.

When comparing the empirical MSEs of the EM based estimator and Bhadra's estimator, we observe that the MSEs for Bhadra's are lower for most parameters in both the simulation study and in the real life application. However, it is not straightforward to designate one estimator as superior, since their performance ultimately depends on the objectives and characteristics of the problem under consideration.

7. Conclusion

This work contributes to the literature on measurement error models through the use of spline approximation, specifically targeting the problem of estimation. Through the simulation study and real life application, we showcase that the proposed algorithm significantly improves the estimation accuracy compared to the naive estimates that ignore measurement error. A key strength of the spline method is its flexibility in capturing complex underlying relationships between variables without imposing restrictive parametric assumptions.

Some promising directions for future work include extending the current methodology to handle heteroscedastic measurement errors and exploring the use of alternative basis functions, provided that the necessary expectations can be appropriately derived.

Acknowledgements

We are deeply grateful to the editor and reviewers for their invaluable comments and suggestions that have significantly enhanced the quality of this manuscript.

Appendix

A. Proofs of the Exp step

Proof. At the t^{th} iteration, the known quantities are $\mathbf{Y}^{(t)}$, $\mathbf{W}^{(t)}$ and $\psi^{(t-1)}$. Without loss of generality, let us omit the conditioning variables for the ease of notation; let $\pi(x_i)$ be the conditional density of X_i.

Using Proposition 1:

$$E(X_i) = \sum_{L=-1}^{K} p_{iL}\mu_{iL},$$

$$E(X_i^2) = \sum_{L=-1}^{K} p_{iL}(\mu_{iL}^2 + \sigma_{iL}^2).$$

From Section 3, the degree one B-spline basis is defined separately for $k = 1$, $k = (K+1)$ and $k = 2,3,\ldots K$, hence, we calculate $E[\widetilde{B}_k(X_i)]$, $E\{[\widetilde{B}_k(X_i)]^2\}$, $E[X_i\widetilde{B}_k(X_i)]$, $E[\widetilde{B}_j(X_i)\widetilde{B}_k(X_i)]$ accordingly.

- Derivation of $E[\widetilde{B}_k(X_i)]$.

For $k=1$

$$\begin{aligned}
E\big[\widetilde{B}_1(X_i)\big] &= E\left[\left(\frac{\xi_1-X_i}{\xi_1-\xi_0}\right)I(\xi_0\le X_i<\xi_1)\right]\\
&= \int_{\xi_0}^{\xi_1}\left(\frac{\xi_1-x_i}{\xi_1-\xi_0}\right)\pi(x_i)\,dx_i\\
&= \frac{\xi_1}{\xi_1-\xi_0}\int_{\xi_0}^{\xi_1}\pi(x_i)\,dx_i-\frac{1}{\xi_1-\xi_0}\int_{\xi_0}^{\xi_1}x_i\pi(x_i)\,dx_i\\
&= \frac{\xi_1 p_{i0}}{\xi_1-\xi_0}-\frac{p_{i0}\mu_{i0}}{\xi_1-\xi_0}\\
&= \frac{p_{i0}(\xi_1-\mu_{i0})}{(\xi_1-\xi_0)}.
\end{aligned}$$

For $k=2,3,\ldots,K$

$$\begin{aligned}
E\big[\widetilde{B}_k(X_i)\big] &= E\left[\frac{X_i-\xi_{k-2}}{\xi_{k-1}-\xi_{k-2}}I(\xi_{k-2}\le X_i<\xi_{k-1})+\frac{\xi_k-X_i}{\xi_k-\xi_{k-1}}I(\xi_{k-1}\le X_i<\xi_k)\right]\\
&= \int_{\xi_{k-2}}^{\xi_{k-1}}\left(\frac{x_i-\xi_{k-2}}{\xi_{k-1}-\xi_{k-2}}\right)\pi(x_i)\,dx_i+\int_{\xi_{k-1}}^{\xi_k}\left(\frac{\xi_k-x_i}{\xi_k-\xi_{k-1}}\right)\pi(x_i)\,dx_i\\
&= \frac{1}{(\xi_{k-1}-\xi_{k-2})}\left[\int_{\xi_{k-2}}^{\xi_{k-1}}x_i\pi(x_i)\,dx_i-\int_{\xi_{k-2}}^{\xi_{k-1}}\xi_{k-2}\pi(x_i)\,dx_i\right]\\
&\quad+\frac{1}{(\xi_k-\xi_{k-1})}\left[\int_{\xi_{k-1}}^{\xi_k}\xi_k\pi(x_i)\,dx_i-\int_{\xi_{k-1}}^{\xi_k}x_i\pi(x_i)\,dx_i\right]\\
&= \frac{\mu_{i(k-2)}\,p_{i(k-2)}}{(\xi_{k-1}-\xi_{k-2})}-\frac{\xi_{k-2}\,p_{i(k-2)}}{(\xi_{k-1}-\xi_{k-2})}+\frac{\xi_k\,p_{i(k-1)}}{(\xi_k-\xi_{k-1})}-\frac{\mu_{i(k-1)}\,p_{i(k-1)}}{(\xi_k-\xi_{k-1})}\\
&= \frac{p_{i(k-2)}}{(\xi_{k-1}-\xi_{k-2})}\Big[\mu_{i(k-2)}-\xi_{k-2}\Big]+\frac{p_{i(k-1)}}{(\xi_k-\xi_{k-1})}\Big[\xi_k-\mu_{i(k-1)}\Big].
\end{aligned}$$

Similarly for $k=(K+1)$, we can show that

$$E\big[\widetilde{B}_{K+1}(X_i)\big]=\frac{p_{i(K-1)}(\mu_{i(K-1)}-\xi_{K-1})}{(\xi_K-\xi_{K-1})}.$$

- Derivation of $E\big[\{\widetilde{B}_k(X_i)\}^2\big]$.

  For $k=1$

$$\begin{aligned}
E\big[\{\widetilde{B}_1(X_i)\}^2\big] &= E\Big[\Big(\frac{\xi_1-X_i}{\xi_1-\xi_0}\Big)^2 I(\xi_0\le X_i<\xi_1)\Big]\\
&= \int_{\xi_0}^{\xi_1}\Big(\frac{\xi_1-x_i}{\xi_1-\xi_0}\Big)^2\pi(x_i)\,dx_i\\
&= \frac{1}{(\xi_1-\xi_0)^2}\Big[\int_{\xi_0}^{\xi_1}\xi_1^2\,\pi(x_i)\,dx_i-\int_{\xi_0}^{\xi_1}2\xi_1x_i\pi(x_i)\,dx_i+\int_{\xi_0}^{\xi_1}x_i^2\pi(x_i)dx_i\Big]\\
&= \frac{1}{(\xi_1-\xi_0)^2}\Big[\xi_1^2\,p_{i0}-2\xi_1\,p_{i0}\,\mu_{i0}+p_{i0}(\mu_{i0}^2+\sigma_{i0}^2)\Big]\\
&= \frac{p_{i0}}{(\xi_1-\xi_0)^2}\Big[\mu_{i0}^2+\sigma_{i0}^2-2\xi_1\,\mu_{i0}+\xi_1^2\Big].
\end{aligned}$$

  For $k=2,3,\ldots,K$

$$\begin{aligned}
E\big[\{\widetilde{B}_k(X_i)\}^2\big] &= E\Big[\Big(\frac{X_i-\xi_{k-2}}{\xi_{k-1}-\xi_{k-2}}\Big)^2 I(\xi_{k-2}\le X_i<\xi_{k-1})\\
&\quad+\Big(\frac{\xi_k-X_i}{\xi_k-\xi_{k-1}}\Big)^2 I(\xi_{k-1}\le X_i<\xi_k)\Big]\\
&= \frac{1}{(\xi_{k-1}-\xi_{k-2})^2}\int_{\xi_{k-2}}^{\xi_{k-1}}(x_i^2-2x_i\xi_{k-2}+\xi_{k-2}^2)\,\pi(x_i)dx_i\\
&\quad+\frac{1}{(\xi_k-\xi_{k-1})^2}\int_{\xi_{k-1}}^{\xi_k}(\xi_k^2-2\xi_kx_i+x_i^2)\,\pi(x_i)dx_i\\
&= \frac{1}{(\xi_{k-1}-\xi_{k-2})^2}\Big[p_{i(k-2)}(\mu_{i(k-2)}^2+\sigma_{i(k-2)}^2)-2\xi_{k-2}p_{i(k-2)}\mu_{i(k-2)}\\
&\quad+\xi_{k-2}^2p_{i(k-2)}\Big]+\frac{1}{(\xi_k-\xi_{k-1})^2}\Big[\xi_k^2\,p_{i(k-1)}-2\xi_k\,p_{i(k-1)}\mu_{i(k-1)}\\
&\quad+p_{i(k-1)}(\mu_{i(k-1)}^2+\sigma_{i(k-1)}^2)\Big]\\
&= \frac{p_{i(k-2)}}{(\xi_{k-1}-\xi_{k-2})^2}\big[\mu_{i(k-2)}^2+\sigma_{i(k-2)}^2-2\xi_{k-2}\mu_{i(k-2)}+\xi_{k-2}^2\big]\\
&\quad+\frac{p_{i(k-1)}}{(\xi_k-\xi_{k-1})^2}\big[\mu_{i(k-1)}^2+\sigma_{i(k-1)}^2-2\xi_k\mu_{i(k-1)}+\xi_k^2\big].
\end{aligned}$$

  Similarly for $k=(K+1)$, we can show that

$$E\big[\{\widetilde{B}_{K+1}(X_i)\}^2\big]=\frac{p_{i(K-1)}}{(\xi_K-\xi_{K-1})^2}\Big[\mu_{i(K-1)}^2+\sigma_{i(K-1)}^2-2\xi_{K-1}\mu_{i(K-1)}+\xi_{K-1}^2\Big].$$

- Derivation of $E[X_i\widetilde{B}_k(X_i)]$.

For $k=1$,

$$\begin{aligned} E[X_i\widetilde{B}_1(X_i)] &= E\Big[X_i\left(\frac{\xi_1-X_i}{\xi_1-\xi_0}\right)I(\xi_0\le X_i<\xi_1)\Big] \\ &= \frac{1}{(\xi_1-\xi_0)}\int_{\xi_0}^{\xi_1}\left(x_i\xi_1-x_i^2\right)\pi(x_i)dx_i \\ &= \frac{\xi_1}{(\xi_1-\xi_0)}p_{i0}\mu_{i0}-\frac{p_{i0}}{(\xi_1-\xi_0)}(\mu_{i0}^2+\sigma_{i0}^2) \\ &= \frac{p_{i0}}{(\xi_1-\xi_0)}\big[\mu_{i0}\xi_1-\mu_{i0}^2-\sigma_{i0}^2\big]. \end{aligned}$$

For $k=2,3,\ldots,K$,

$$\begin{aligned} E[X_i\widetilde{B}_k(X_i)] &= E\Big[X_i\left(\frac{X_i-\xi_{k-2}}{\xi_{k-1}-\xi_{k-2}}\right)I(\xi_{k-2}\le X_i<\xi_{k-1}) \\ &\quad +X_i\left(\frac{\xi_k-X_i}{\xi_k-\xi_{k-1}}\right)I(\xi_{k-1}\le X_i<\xi_k)\Big] \\ &= \frac{1}{(\xi_{k-1}-\xi_{k-2})}\Big[\int_{\xi_{k-2}}^{\xi_{k-1}}x_i^2\pi(x_i)dx_i-\int_{\xi_{k-2}}^{\xi_{k-1}}x_i\xi_{k-2}\pi(x_i)dx_i\Big] \\ &\quad +\frac{1}{(\xi_k-\xi_{k-1})}\Big[\int_{\xi_{k-1}}^{\xi_k}x_i\xi_k\pi(x_i)dx_i-\int_{\xi_{k-1}}^{\xi_k}x_i^2\pi(x_i)dx_i\Big] \\ &= \frac{p_{i(k-2)}}{(\xi_{k-1}-\xi_{k-2})}\big[\mu_{i(k-2)}^2+\sigma_{i(k-2)}^2-\xi_{k-2}\mu_{i(k-2)}\big] \\ &\quad +\frac{p_{i(k-1)}}{(\xi_k-\xi_{k-1})}\big[-\mu_{i(k-1)}^2-\sigma_{i(k-1)}^2+\xi_k\mu_{i(k-1)}\big]. \end{aligned}$$

Similarly for $k=(K+1)$, we can show that

$$E[X_i\widetilde{B}_{K+1}(X_i)]=\frac{p_{i(K-1)}}{(\xi_K-\xi_{K-1})}\big[\mu_{i(K-1)}^2+\sigma_{i(K-1)}^2-\xi_{K-1}\mu_{i(K-1)}\big].$$

- Derivation of $E[\widetilde{B}_j(X_i)\widetilde{B}_k(X_i)]$.

Note from Section 3 that $E[\widetilde{B}_j(X_i)\widetilde{B}_k(X_i)]$ is non-zero when $|j-k|=1$.

For $j=1$ and $k=2$

$$\begin{aligned}
E[\widetilde{B}_1(X_i)\widetilde{B}_2(X_i)] &= E\Big\{\Big[\Big(\frac{\xi_1-X_i}{\xi_1-\xi_0}\Big)I(\xi_0\le X_i<\xi_1)\Big] \\
&\quad\times\Big[\Big(\frac{X_i-\xi_0}{\xi_1-\xi_0}\Big)I(\xi_0\le X_i<\xi_1)+\frac{\xi_2-X_i}{\xi_2-\xi_1}I(\xi_1\le X_i<\xi_2)\Big]\Big\} \\
&= E\Big[\Big(\frac{\xi_1-X_i}{\xi_1-\xi_0}\Big)I(\xi_0\le X_i<\xi_1)\Big(\frac{X_i-\xi_0}{\xi_1-\xi_0}\Big)I(\xi_0\le X_i<\xi_1)\Big] \\
&= \frac{1}{(\xi_1-\xi_0)^2}E\big[(\xi_1X_i-\xi_0\xi_1-X_i^2+X_i\xi_0)I(\xi_0\le X_i<\xi_1)\big] \\
&= \frac{1}{(\xi_1-\xi_0)^2}\Big[\int_{\xi_0}^{\xi_1}(\xi_1x_i-\xi_0\xi_1-x_i^2+x_i\xi_0)\,\pi(x_i)dx_i\Big] \\
&= \frac{1}{(\xi_1-\xi_0)^2}\big[\xi_1\,p_{i0}\,\mu_{i0}-\xi_0\,\xi_1\,p_{i0}-p_{i0}\,(\mu_{i0}^2+\sigma_{i0}^2)+\xi_0\,p_{i0}\,\mu_{i0}\big] \\
&= \frac{p_{i0}}{(\xi_1-\xi_0)^2}\big[\mu_{i0}(\xi_1+\xi_0)-\xi_0\,\xi_1-(\mu_{i0}^2+\sigma_{i0}^2)\big].
\end{aligned}$$

Similarly deriving $E[\widetilde{B}_j(X_i)\widetilde{B}_k(X_i)]$ for $j=2,3,\ldots,K$, $k=(j+1)$, for $j=1,2,\ldots,K$ we get

$$E[\widetilde{B}_j(X_i)\widetilde{B}_{j+1}(X_i)]=\frac{p_{i(j-1)}}{(\xi_j-\xi_{j-1})^2}\big[\mu_{i(j-1)}(\xi_j+\xi_{j-1})-\xi_{j-1}\,\xi_j-(\mu_{i(j-1)}^2+\sigma_{i(j-1)}^2)\big].$$

□

References

K. Ayub, W. Song, J. Shi (2022). *Extrapolation estimation in parametric regression models with measurement error*. Computational Statistics and Data Analysis, 172, p. 107478.

A. Bhadra (2017). *An expectation–maximization scheme for measurement error models*. Statistics & Probability Letters, 120, pp. 61–68.

A. Bhadra, R. J. Carroll (2016). *Exact sampling of the unobserved covariates in Bayesian spline models for measurement error problems*. Statistics and Computing, 26, no. 4, pp. 827–840.

C. Butucea, C. Matias (2005). *Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model*. Bernoulli, 11, no. 2, pp. 309–340.

R. J. Carroll, J. D. Maca, D. Ruppert (1999). *Nonparametric regression in the presence of measurement error*. Biometrika, 86, no. 3, pp. 541–554.

R. J. CARROLL, D. RUPPERT, L. A. STEFANSKI, C. M. CRAINICEANU (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, Boca Raton, 2nd ed.

J. COOK, L. STEFANSKI (1994). *Simulation-extrapolation estimation in parametric measurement error models*. Journal of the American Statistical Association, 89, no. 428, pp. 1314–1328.

A. P. DEMPSTER, N. M. LAIRD, D. B. RUBIN (1977). *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society: Series B (Methodological), 39, no. 1, pp. 1–22.

P. H. C. EILERS, B. D. MARX (1996). *Flexible smoothing with B-splines and penalties*. Statistical Science, 11, no. 2, pp. 89 – 121.

T. LI, Q. VUONG (1998). *Nonparametric estimation of the measurement error model using multiple indicators*. Journal of Multivariate Analysis, 65, no. 2, pp. 139–165.

T. LYCHE, C. MANNI, H. SPELEERS (2018). *Foundations of Spline Theory: B-Splines, Spline Approximation, and Hierarchical Refinement*. Springer, Cham.

A. MEISTER (2006). *Density estimation with normal measurement error with unknown variance*. Statistica Sinica, 16, no. 1, pp. 195–211.

C. F. J. WU (1983). *On the convergence properties of the EM algorithm*. The Annals of Statistics, 11, no. 1, pp. 95–103.