A Mixed Model Approach to Measurement Error in Semiparametric Regression

Mohammad W. Hattab

School of Medicine, Johns Hopkins University

and

David Ruppert

Department of Statistics and Data Science and School of Operations Research and Information Engineering, Cornell University

Abstract

An essential assumption in traditional regression techniques is that predictors are measured without errors. Failing to take into account measurement error in predictors may result in severely biased inferences. Correcting measurement-error bias is an extremely difficult problem when estimating a regression function nonparametrically. We propose an approach to deal with measurement errors in predictors when modelling flexible regression functions. This approach depends on directly modelling the mean and the variance of the response variable after integrating out the true unobserved predictors in a penalized splines model. We demonstrate through simulation studies that our approach provides satisfactory prediction accuracy largely outperforming previously suggested local polynomial estimators even when the model is incorrectly specified and is competitive with the Bayesian estimator.

Keywords: Nonparametric regression; Penalized splines; Variance function estimation.

1 Introduction

A key assumption of parametric and nonparametric regression models is that predictors are measured without errors. Regression techniques that ignore measurement errors in predictors may produce highly biased estimates leading to erroneous inferences. The magnitude of the bias resulting from ignoring measurement errors in predictors decreases only slightly as the sample size increases and does not converge to 0. For instance, in simple linear regression the least squares slope estimate from naively fitting the response variable on a distorted version of the true predictor is biased by a factor that depends on the ratio of the variance of the measurement error to the variance of the true unobserved predictor and is independent of the sample size. Discussion of measurement error models in parametric regression can be found in Fuller (1987), Carroll (1989), Cook and Stefanski (1994), Spiegelman, Rosner and Logan (2000) and Carroll et al. (2006).

It is extremely difficult to estimate regression functions nonparametrically with mismeasured predictors. To handle this problem, Carroll, Maca, and Ruppert (1999) extend the simulation–extrapolation method (Cook and Stefanski 1994) to nonparametric regression. Their method outperforms the kernel based estimator proposed by Fan and Truong (1993). However, the comparison with Fan and Truong's (1993) method is considered incomplete since the latter method was introduced without an optimal bandwidth selector. In Section 4, we will contrast the prediction accuracy of our method against Fan and Truong's (1993) method that uses the 2-stage plug-in bandwidth estimator presented in Delaigle and Gijbels (2002). Staudenmayer and Ruppert (2004) presents local polynomial simulation–extrapolation estimator that improves upon Carroll et al. (1999) method by utilizing a bandwidth selector that aims to minimize the estimated mean squared error of the final estimate. Staudenmayer and Ruppert (2004) method is also included in the comparison study in Section 4 as well as the local polynomial estimators proposed by Delaigle, Fan and Carroll (2009) and Huang and Zhou (2017). Delaigle et al. (2009) generalizes Fan and Truong's (1993) method to local polynomial estimators using a complex transformation of the kernel function. Huang and Zhou (2017) adjusts the estimator in Delaigle et al. (2009) to work directly with transformations of the naive estimator that ignores measurement errors in predictors.

Berry, Ruppert and Carroll (2002) developed a fully Bayesian approach utilizing penalized splines (Eilers and Marx 1996). They indicated through simulations that their method showed superior performance in terms of the mean squared error as compared to Carroll's et al. (1999) methods.

Ruppert, Wand and Carroll (2003) and Ganguli, Staudenmayer and Wand (2005) noted that the observed data likelihood function in Berry et al. (2002) cannot be computed analytically since it contains intractable high-dimensional integrals in the unobserved predictors. Since it is not possible to integrate out the unobserved predictors, Ganguli et al. (2005) resorted to a Monte Carlo expectation maximization (EM) algorithm to compute the maximum likelihood estimates. This algorithm is computationally demanding, as much as the Bayesian approach. It requires at every iteration a Metropolis-Hastings step to draw hundreds of samples from the conditional distribution of the unobserved predictors, which is evidently a time consuming step when n is large.

Another interesting Bayesian approach is developed by Sarkar, Mallick, and Carroll (2014). By using mixtures of Dirichlet processes, it generalizes the Berry et al. (2002) approach by modeling the distribution of the unobserved predictor and allowing both the measurement error and the regression error to be non-normal and heteroscedastic.

In this article, we present a frequentist method to the problem of measurement error in nonparametric regression that adopts the same set of assumptions as Berry et al. (2002) and Ganguli et al. (2005).

While the exact distribution of the response variable given the observed predictors cannot be obtained, exact expressions of the mean and the covariance matrix of the response variable given the observed predictors conditionally on the regression coefficients can be found as will be demonstrated later. These results are novel and may be of independent interest. We propose fitting iteratively a heteroscedastic mixed model derived from the conditional moments. The focus of this method is on the first and the second moments without specifying a particular distribution.

Section 2 reviews the measurement error problem in semiparametric regression in greater detail. Specifically, penalized-spline regression is given in the form of linear mixed models along with the measurement error mechanism that controls the relationship between the observed predictors and the unobserved predictors. The difficulty of obtaining the observed data likelihood function is highlighted before introducing estimation methods and the Bayesian approach. In Section 3 we derive the basis of our method and describe an algorithm to execute it. Through simulations, Section 4 studies the prediction power of the proposed method under a variety of cases and sample sizes. We conclude that this method offers superior performance over a set of local polynomial estimators even when the model is misspecified and is competitive with the Bayesian approach. It is very fast to compute allowing one to adjust for measurement error in nonparametric regression models for very large samples. An application to real data is provided in Section 5 and Section 6 summarizes the findings. For simplicity and to conserve space, the following discussion is restricted to one predictor, but its extension to additive models is straightforward.

2 Penalized splines with measurement error

Consider the linear mixed model representation of the linear penalized splines model given by Ruppert and Carroll (2000) and Ruppert et al. (2003)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \mathrm{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}) \quad \text{and} \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I}) \tag{1}$$

where $\boldsymbol{Y} = [y_i]$ is an $n \times 1$ vector of observable response variables, $\boldsymbol{\beta} = [\beta_0 \quad \beta_1]^\top$ is an unknown vector of regression parameters, $\boldsymbol{e} = [e_i]$ is an $n \times 1$ vector of unobservable errors and $\boldsymbol{u} = [u_j]$ is a $k \times 1$ vector of unobservable random effects. The random vector $\boldsymbol{e}$ is independent of $\boldsymbol{u}$. Here $\boldsymbol{X} = [\boldsymbol{1} \quad \boldsymbol{x}]$ an $n \times 2$ matrix with $\boldsymbol{x} = [x_1 \ldots x_n]^\top$ and $\boldsymbol{Z} = [(x_i - k_j)_+]_{i=1,\ldots n}^{j=1,\ldots k}$ an $n \times k$ matrix where $\{k_1 \ldots k_k\}$ is a fixed set of knots. The representation in $\boldsymbol{Z}$ uses the truncated lines basis, but B-splines could be used instead. Simple linear regression corresponds to $\sigma_u^2 = 0$ (overly smoothed) whereas $\sigma_u^2 \to \infty$ implies ordinary multiple linear regression in $\boldsymbol{X}$ and $\boldsymbol{Z}$ with $\boldsymbol{u}$ now being treated as a fixed effect vector resulting in no smoothing. Maximum likelihood or restricted maximum likelihood (REML) estimation is often used to estimate the parameters.

Let $\boldsymbol{w} = [w_i]$ be a $n \times 1$ vector of the observed values of the predictor x. Typically, a measurement error model assumes that

$$\boldsymbol{w}|\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{x}, \sigma^2_{w|x}\boldsymbol{I}), \quad \text{and} \quad \boldsymbol{x} \sim \mathrm{N}(\mu_x \mathbf{1}, \sigma^2_x \boldsymbol{I}) \tag{2}$$

The observed predictor w is a distorted version of x, where the level of distortion is controlled by $\sigma^2_{w|x}$. Equations 1 and 2 result in a measurement error model with $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2_e, \sigma^2_u, \sigma^2_{w|x}, \mu_x, \sigma^2_x)$ being the vector of the unknown parameters to be estimated from the data. Define

$$\begin{aligned} l(\boldsymbol{\theta}) &\propto \frac{\parallel \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u} \parallel^2}{\sigma^2_e} + \frac{\parallel \boldsymbol{w} - \boldsymbol{x} \parallel^2}{\sigma^2_{w|x}} + \frac{\parallel \boldsymbol{x} - \mu_x \mathbf{1} \parallel^2}{\sigma^2_x} + \frac{\parallel \boldsymbol{u} \parallel^2}{\sigma^2_u} \\ &+ n \log \sigma^2_e + n \log \sigma^2_{w|x} + n \log \sigma^2_x + k \log \sigma^2_u \end{aligned} \tag{3}$$

Ganguli et al. (2005) denoted $l(\boldsymbol{\theta})$ as the complete data log-likelihood that corresponds to the complete data $\boldsymbol{D}_{\text{comp}} = \left[\boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{w}\right]$, whereas $l_{\text{obs}}(\boldsymbol{\theta})$ corresponds to the observed data $\boldsymbol{D}_{obs} = \left[\boldsymbol{Y}, \boldsymbol{w}\right]$. Then, $l_{\text{obs}}(\boldsymbol{\theta}) = \log \left[\int \exp l(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{u}\, \mathrm{d}\boldsymbol{x}\right]$. Ruppert et al. (2003) and Ganguli et al. (2005) noted that this high dimensional integral is intractable and therefore $l_{\text{obs}}(\boldsymbol{\theta})$ cannot be obtained. Specifically, integrating $\boldsymbol{u}$ out will result in an n-dimensional integral with an integrand that contains a very complicated function of $\boldsymbol{x}$. Based on the fact that the density of $\boldsymbol{x}|\boldsymbol{u}, \boldsymbol{Y}, \boldsymbol{w}$ is proportional to the exponential of the first three terms in (3), Ganguli et al. (2005) developed a nested Monte Carlo expectation maximization (EM) algorithm to estimate $\boldsymbol{\theta}$. It requires at every iteration a Metropolis-Hastings algorithm to draw m samples from the distribution of $\boldsymbol{x}|\boldsymbol{u}, \boldsymbol{Y}, \boldsymbol{w}$ which is time consuming, especially when n is large. Also, Ruppert et al. (2003) indicated that a reasonable choice of m is unclear.

Carroll et al. (1999) suggest estimating $\sigma^2_{w|x}$ via either external data sets or using the pooled sample variance of $\boldsymbol{w}$ if replicate measurements are available. They showed that the asymptotic effect of these estimation methods on the nonparametric regression function is negligible. The Monte Carlo EM approach and a partial Bayesian approach suggested by Berry et al. (2002) adopted these methods as well. They also estimated μ_x and σ_x^2 by the sample mean and sample variance of $\boldsymbol{w}$ minus the estimate of $\sigma_{w|x}^2$, respectively.

Obviously, an immediate advantage of a fully Bayesian approach is its ability to take into account the uncertainty in estimating the regression function by placing prior distributions on $\boldsymbol{\beta}$, μ_x and all variance parameters. Diffuse priors are typically assumed. The joint posterior distribution cannot be analytically found. Instead, Berry et al. (2002) developed an MCMC scheme to sample from the posterior distribution. A Metropolis-Hastings step is needed to sample from the conditional distribution of $\boldsymbol{x}$. Once the posterior draws have been generated, one can easily obtain posterior means and credible intervals for all parameters including $\boldsymbol{x}$ and $\boldsymbol{u}$. Berry et al. (2002) demonstrated that their method provided satisfactory prediction accuracy and performed better in terms of the mean squared error than previously suggested methods.

Next, we develop competitive frequentist approach that models the mean and the variance of $\boldsymbol{Y}$ given $\boldsymbol{w}$ and $\boldsymbol{u}$.

3 Heterosedastic mixed model approach

Ruppert et al. (2003) and Ganguli et al. (2005) noted that the density of $\boldsymbol{Y}|\boldsymbol{w}$ is proportional to the intractable integral $\int \exp l(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{u}\, \mathrm{d}\boldsymbol{x}$. It is straightforward to show that $\boldsymbol{Y}|\boldsymbol{x} \sim \mathrm{N}\big(\boldsymbol{X\beta}, \sigma_u^2 \boldsymbol{ZZ}^\top + \sigma_e^2 \boldsymbol{I}\big)$ but integrating $\boldsymbol{x}$ out to obtain the density of $\boldsymbol{Y}|\boldsymbol{w}$ is not possible due to the presence of $\boldsymbol{x}$ in $\mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{x})$. Therefore, we will not seek the density of $\boldsymbol{Y}|\boldsymbol{w}$ but rather we will derive the exact expressions of $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u})$ and $\mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u})$ and suggest an algorithm based on an iterative-heteroscedastic mixed model with the random effects contained in $\boldsymbol{u}$; see (1).

First, the conditional mean of $\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u}$ is given by

$$\begin{aligned}
\mathrm{E}(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u}) &= \mathrm{E}\big(\mathrm{E}(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u},\boldsymbol{x})|\boldsymbol{w},\boldsymbol{u}\big) \\
&= \mathrm{E}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}|\boldsymbol{w},\boldsymbol{u}) \\
&= \mathrm{E}(\boldsymbol{X}|\boldsymbol{w})\boldsymbol{\beta}+\mathrm{E}(\boldsymbol{Z}|\boldsymbol{w})\boldsymbol{u} \qquad (4)
\end{aligned}$$

where the expectation operator is computed element-wise. Next, the conditional covariance matrix of $\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u}$ is given by

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u}) &= \mathrm{E}\big(\mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u},\boldsymbol{x})|\boldsymbol{w},\boldsymbol{u}\big)+\mathrm{Cov}\big(\mathrm{E}(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u},\boldsymbol{x})|\boldsymbol{w},\boldsymbol{u}\big) \\
&= \sigma_e^2\boldsymbol{I}+\mathrm{Cov}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}|\boldsymbol{w},\boldsymbol{u}) \\
&= \sigma_e^2\boldsymbol{I}+\mathrm{Cov}(\boldsymbol{X}\boldsymbol{\beta}|\boldsymbol{w})+\mathrm{Cov}(\boldsymbol{Z}\boldsymbol{u}|\boldsymbol{w},\boldsymbol{u}) \\
&+\mathrm{Cov}(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{Z}\boldsymbol{u}|\boldsymbol{w},\boldsymbol{u})+\mathrm{Cov}(\boldsymbol{Z}\boldsymbol{u},\boldsymbol{X}\boldsymbol{\beta}|\boldsymbol{w},\boldsymbol{u})
\end{aligned} \qquad (5)$$

The second and the third equations follow from the facts that $\boldsymbol{Y}$ and $\boldsymbol{w}$ are conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{u}$, and $\boldsymbol{x}$ and $\boldsymbol{u}$ are marginally and conditionally independent given $\boldsymbol{w}$.

To compute (4) and (5), we need the following theorems. The first establishes the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{w}$. It is a direct application of Bayes theorem and the joint Gaussian distribution of $(\boldsymbol{x},\boldsymbol{w})$, and will be given without a proof. The second gives integral identities related to the expressions (4) and (5). The proof is provided in the Appendix.

Theorem 1. The assumptions in (2) are equivalent to

$$\boldsymbol{x}|\boldsymbol{w} \sim \mathrm{N}\big(\frac{\sigma_x^2\boldsymbol{w}+\mu_x\sigma_{w|x}^2\mathbf{1}}{\sigma_x^2+\sigma_{w|x}^2}, \frac{\sigma_x^2\sigma_{w|x}^2}{\sigma_x^2+\sigma_{w|x}^2}\boldsymbol{I}\big) \quad \text{and} \quad \boldsymbol{w} \sim \mathrm{N}\big(\mu_x\mathbf{1}, (\sigma_x^2+\sigma_{w|x}^2)\boldsymbol{I}\big) \tag{6}$$

Theorem 2. Let f and F be the density function and the cumulative distribution function for a normal distribution with mean μ and variance s^2 and a and b any two real numbers with $a < b$. Then,

$$\int_{-\infty}^{+\infty} (x-a)_+ f(x)\,\mathrm{d}x = s^2 f(a) + (\mu - a)(1 - F(a)) \tag{7}$$

$$\int_{-\infty}^{+\infty} (x-a)_+^2 f(x)\,\mathrm{d}x = s^2 f(a)(\mu - a) + \big((\mu-a)^2 + s^2\big)\big(1 - F(a)\big) \tag{8}$$

$$\int_{-\infty}^{+\infty} x(x-a)_+ f(x)\,\mathrm{d}x = s^2 f(a)\mu + (s^2 + \mu(\mu - a))(1 - F(a)) \tag{9}$$

$$\int_{-\infty}^{+\infty} (x-a)_+(x-b)_+ f(x)\,\mathrm{d}x = s^2 f(b)(\mu - a) + (s^2 + (\mu-a)(\mu-b))(1 - F(b)) \tag{10}$$

By using (6), the ith row of the matrix $\mathrm{E}(\boldsymbol{X}|\boldsymbol{w})$ is

$$\begin{bmatrix}1 & \mathrm{E}(x_i|w_i)\end{bmatrix} = \begin{bmatrix}1 & \dfrac{\sigma_x^2 w_i + \mu_x\sigma_{w|x}^2}{\sigma_x^2+\sigma_{w|x}^2}\end{bmatrix} \tag{11}$$

and by (7) the (i,j)-entry of the matrix $\mathrm{E}(\boldsymbol{Z}|\boldsymbol{w})$ is given by

$$\mathrm{E}\big((x_i - k_j)_+|w_i\big) = \frac{\sigma_x^2\sigma_{w|x}^2}{\sigma_x^2+\sigma_{w|x}^2} f_i(k_j) + \left(\frac{\sigma_x^2 w_i + \mu_x\sigma_{w|x}^2}{\sigma_x^2+\sigma_{w|x}^2} - k_j\right)(1 - F_i(k_j)) \tag{12}$$

where now f_i and F_i are the pdf and the cdf of $x_i|w_i$. As expected, nothing can be learned about the regression function through $\boldsymbol{w}$ when $\sigma^2_{w|x} \to \infty$ as the contributions of $\boldsymbol{w}$ will disappear in both (11) and (12). Notice that when $\sigma^2_{w|x} = 0$, $F_i(k_j) = 1$ if $x_i \leq k_j$ and 0 otherwise and in this case (12) is simply a spline function with knot k_j.

Finding the conditional variance in (5) is more involved. The details are in the Appendix. There it is shown that $\text{Cov}(\boldsymbol{Y}|\boldsymbol{w}, \boldsymbol{u})$ is a diagonal matrix with the ith diagonal element given by

$$v_i = \sigma^2_e + \beta^2_1 \frac{\sigma^2_x \sigma^2_{w|x}}{\sigma^2_x + \sigma^2_{w|x}} + 2\beta_1 \text{Cov}(x_i, \boldsymbol{z}_i|w_i)\boldsymbol{u} + \boldsymbol{u}^\top \text{Cov}(\boldsymbol{z}_i|w_i)\boldsymbol{u} \tag{13}$$

The parameters $\sigma^2_{w|x}$, μ_x and σ^2_x are estimated as described in Section 2. Once the estimates of $\sigma^2_{w|x}$, μ_x and σ^2_x have been obtained, they are plugged in (4) and (5). The problem can be put in a form of a heteroscedastic mixed model. Specifically,

$$\boldsymbol{Y} = \text{E}(\boldsymbol{X}|\boldsymbol{w})\boldsymbol{\beta} + \text{E}(\boldsymbol{Z}|\boldsymbol{w})\boldsymbol{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \psi(\boldsymbol{0}, \boldsymbol{V}) \quad \text{and} \quad \boldsymbol{u} \sim \text{N}(\boldsymbol{0}, \sigma^2_u \boldsymbol{I}) \tag{14}$$

where $\psi(\boldsymbol{0}, \boldsymbol{V})$ is an unknown distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{V} = \text{diag}(v_i)$. Evidently, ψ does not correspond to a normal distribution. The matrices $\text{E}(\boldsymbol{X}|\boldsymbol{w})$ and $\text{E}(\boldsymbol{Z}|\boldsymbol{w})$ represent known predictors. Despite its apparent similarity with model (1), model (14) is different in three main aspects. First, the response variable does not follow a normal distribution, second, the variance is not constant, third, and more importantly the error vector $\boldsymbol{\epsilon}$ depends on the random effect $\boldsymbol{u}$ and β_1 through the covariance matrix $\boldsymbol{V}$. The term v_i is a quadratic function in β_1 and $\boldsymbol{u}$. However, $\text{E}(\boldsymbol{\epsilon}|\boldsymbol{u}) = \boldsymbol{0}$, which is critical.

Even if $\boldsymbol{Y}|\boldsymbol{w}, \boldsymbol{u}$ follows a normal distribution, the dependence between $\boldsymbol{\epsilon}$ and $\boldsymbol{u}$ means that the likelihood function cannot be directly optimized (Davidian and Giltinan 1995). The following iterative algorithm aims to estimate the parameters $\boldsymbol{\beta}$, σ_e^2 and σ_u^2 and predict $\boldsymbol{u}$:

1. Fit a homoscedastic linear mixed model to (14), that is, as a working model take $\boldsymbol{V} = \sigma_\epsilon^2 \boldsymbol{I}$, $\sigma_\epsilon > 0$.

2. From the fitted model, obtain the estimate $\hat{\beta}_1$ and the prediction $\hat{\boldsymbol{u}}$ and plug-in them in (13) so $\hat{v}_i = \sigma_e^2 + c_i$ where c_i's are known constants.

3. Fit the heteroscedastic mixed model in (14) with v_i being set at $\hat{v}_i$.

4. Repeat steps 2–3 until stabilization.

5. Using the estimates from the previous step, obtain the fitted curve over a grid of values $\boldsymbol{d} = (d_1, \ldots, d_t)$

$$\begin{bmatrix}1 & d_i\end{bmatrix}_{i=1,\ldots t} \hat{\boldsymbol{\beta}} + [(d_i - k_j)_+]_{i=1,\ldots t}^{j=1,\ldots k} \hat{\boldsymbol{u}}$$

This method will be called by the conditional approach since it is based on the mean and the variance of $\boldsymbol{Y}$ given the random effects $\boldsymbol{u}$. We recommend using k-fold cross validation to estimate the variance parameters. Specifically, for a certain value of σ_e^2, the weights are formed, the data are split into k folds and then predictions for each fold are computed using the model fitted in the remaining $k-1$ folds. The model is fitted using a sequence of values of σ_u^2. The process is repeated over a grid values of σ_e^2 and the pair of σ_u^2 and σ_e^2 that minimizes the mean squared error is selected. Instead of providing a grid of σ_e^2, one can use an optimization method (e.g. BFGS) aiming to find σ_e^2 that minimizes the cross validation mean squared error in σ_u^2. We adopt this option in the simulation section. We have found that cross validation is more robust than REML when the distribution of x is incorrectly specified as normal.

There are three important points to consider. First, as explained in Ruppert et al. (2003), under homoscedasticity such as the model in (1), the penalty parameter in the penalized fitting criterion is σ_e^2/σ_u^2. For the heteroscedastic model in (20), it can be shown that the penalty parameter is σ_u^{-2}. The equivalency between the two parameters allows one to work with well developed software (e.g. `glmnet` in R) that aim to find the optimal value of the penalty parameter. Second, optimal estimators cannot be obtained using this method since ψ does not correspond to a density of normal distribution. This method focuses completely on modelling the mean and the variance without specifying a particular distribution. Third, it is straightforward to generalize this method when measurement errors are heteroscedastic (e.g. the covariate for the males group is measured more accurately than the females group).

The prediction of the process, y_0 when $x = x_0$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 + [(x_0 - k_j)_+]^{j=1,\ldots k}\hat{\boldsymbol{u}} \tag{15}$$

If w_0 a distorted version of x_0 is observed rather than x_0, it can be shown that the estimated best linear unbiased predictor (EBLUP) of y_0 is given by

$$\hat{\beta}_0 + \frac{\sigma_x^2 w_0 + \mu_x \sigma_{w|x}^2}{\sigma_x^2 + \sigma_{w|x}^2}\hat{\beta}_1 + \left[\mathrm{E}((x_0 - k_j)_+|w_0)\right]^\top_{j=1,\ldots k}\hat{\boldsymbol{u}} \tag{16}$$

Finally, the measurement error model assumes that $\boldsymbol{x}$ is a latent variable similar to the random effect $\boldsymbol{u}$. It can be predicted given $\boldsymbol{Y}$ using the same theoretical tools. First, it is not too difficult to show that the covariance between x_i and y_j given $\boldsymbol{w}$ is $\beta_1 \dfrac{\sigma_x^2 \sigma_{w|x}^2}{\sigma_x^2 + \sigma_{w|x}^2}$ whenever $i = j$ and zero otherwise. Thus, the EBLUP of $\boldsymbol{x}$ is

$$\frac{\sigma_x^2\boldsymbol{w} + \mu_x\sigma_{w|x}^2\mathbf{1}}{\sigma_x^2 + \sigma_{w|x}^2} + \hat{\beta}_1\frac{\sigma_x^2\sigma_{w|x}^2}{\sigma_x^2 + \sigma_{w|x}^2}\mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{w})^{-1}(\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y}|\boldsymbol{w})\hat{\boldsymbol{\beta}}) \tag{17}$$

The details are omitted for brevity, but it can be proved that the marginal moments of $\boldsymbol{Y}|\boldsymbol{w}$ are given by

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}|\boldsymbol{w}) &= \mathrm{E}(\boldsymbol{X}|\boldsymbol{w})\boldsymbol{\beta}, \quad \text{and} && (18)\\ \mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{w}) &= \big(\sigma_e^2 + \beta_1^2\frac{\sigma_x^2\sigma_{w|x}^2}{\sigma_x^2 + \sigma_{w|x}^2}\big)\boldsymbol{I} \\ &\quad + \sigma_u^2\left(\mathrm{diag}\big(\sum_{j=1}^{k}\mathrm{Var}((x_i - k_j)_+|\boldsymbol{w})\big) + \mathrm{E}(\boldsymbol{Z}|\boldsymbol{w})\mathrm{E}(\boldsymbol{Z}|\boldsymbol{w})^\top\right) && (19) \end{aligned}$$

Next, the prediction accuracy of the conditional approach is studied through simulations and compared against the local constant estimator of Fan and Truong (1993), the local polynomial estimators in Staudenmayer and Ruppert (2004), Delaigle et al. (2009) and Huang and Zhou (2017), and the Bayesian approach. It is shown that it has superior performance over other frequentist methods and is highly competitive with the Bayesian approach.

4 Simulations

Four regression functions are studied to evaluate the prediction accuracy of the conditional approach presented in the previous section. The comparison scheme presented here is similar to the one presented in Berry et al. (2002) except that here the sample size is varied from 2^7 to 2^{14} for each case.

In terms of the mean squared error (MSE), the conditional approach is compared against the naive model that fits a linear spline model to the observed predictor w without accounting for measurement errors and against the local constant estimator of Fan and Truong (1993), the local polynomial estimators in Staudenmayer and Ruppert (2004), Delaigle et al. (2009) and Huang and Zhou (2017), and the Bayesian approach of Berry et al. (2002). The MSE is calculated over a grid of 101 points in the interval $[a, b]$ covering most of the range of x. For each case, 100 data-sets have been generated and $\sigma^2_{w|x}$ is assumed to be known for the all approaches.

The bandwidth parameter in methods of Fan and Truong (1993), Delaigle et al. (2009) and Huang and Zhou (2017) are estimated using the 2-stage plug-in estimator proposed by Delaigle and Gijbels (2002, 2004). Delaigle and Gijbels (2002, 2004) showed that this method approximately minimizes the asymptotic mean integrated squared error and performs well in finite samples. Similar to Delaigle and Hall's (2008) method, Huang and Zhou (2017) developed a bandwidth selector procedure based on SIMEX and cross validation to be used in conjunction with their method and with Delaigle et al. (2009) method as well. We recompute these estimators according to this bandwidth selector. However, it is noted that this procedure is very computationally demanding. At $n = 2^{14}$, it may take more than 5 hours to compute. For this reason, we implemented it on 20 data-sets only.

Along with Staudenmayer and Ruppert (2004) method, there are in total six local polynomial estimators. To ease visualisation, at each sample size we will only report the lowest MSE of all local polynomial estimators. This unrealistic estimator is called the best local polynomial estimator, henceforth BLPE. Consequently, the BLPE has been given an unfair advantage over the other estimators. Nonetheless, as we will see, our conditional approach outperforms BLPE.

The conditional approach and the Bayesian approach are set to use the same set of 40 equally-spaced knots. In Appendix B, we vary the number of knots from 10 to 50 and find that the number of knots has little effect on accuracy. In fact, Wang, Shen, and Ruppert (2011) demonstrated that performance is largely independent of the number of knots, because the number of knots is not a smoothing parameter of a penalized spline. For the conditional approach, the R-package `glmnet` is used to perform 5-fold cross validation. For the Bayesian approach, non-informative priors have been assumed for all parameters. Specifically, normal distributions with large variance and inverse gamma distributions with small scale and shape parameters are used as prior distributions on the mean parameters and on the variance parameters, respectively. The total number of posterior samples is 4000 including a burn-in period of 1000. We developed a customized MCMC algorithm to compute the Bayesian estimator that is much faster than Bayesian software such as JAGS or STAN.

The four cases considered are as follows:

1. The parameters are $a = -2$, $b = 2$, $\sigma_e^2 = 0.3^2$, $\sigma_{w|x}^2 = 0.8^2$, $\mu_x = 0$ and $\sigma_x^2 = 1$. The regression function is given by

$$m(x) = \frac{\sin(\pi x/2)}{1 + 2x^2(\mathrm{sign(x)} + 1)}$$

2. $a = 0.1$, $b = 0.9$, $\sigma_e^2 = 0.0015^2$, $\sigma_{w|x}^2 = (3/7)\sigma_x^2$, $\mu_x = 0.5$, $\sigma_x^2 = 0.25^2$, and the regression function

$$m(x) = 1000(x)_+^3(1-x)_+^3$$

3. $a = 0.1$, $b = 0.9$, $\sigma_e^2 = 0.05^2$, $\sigma_{w|x}^2 = 0.141^2$, $\mu_x = 0.5$, $\sigma_x^2 = 0.25^2$, and the regression function

$$m(x) = 10\sin(4\pi x)$$

4. $a = 0.1$, $b = 0.9$, $\sigma_e^2 = 0.35^2$, $\sigma_{w|x}^2 = 0.1^2$, $\mu_x = 0.5$, $\sigma_x^2 = 1/6^2$, and the regression function

$$m(x) = 3\exp(-78(x-0.38)^2) + \exp(-200(x-0.75)^2) - x$$

The four functions described above are displayed in Figure 1 along with one simulation example of the observed data, (y, w). Note that the features of the true curve may not be easily detected through the scatterplot of the observed data.

The results of the simulation study are summarized in Figure 2. Figure 2 presents the median MSE for all approaches and the ratio of squared bias to the MSE for the naive and the conditional approaches.

To demonstrate the importance and the necessity of taking measurement error into account regardless of the sample size, notice that the naive approach has unsatisfactory performance in all samples sizes and cases. Its performance improves only slightly as n increases. In case-1, when $n = 2^{14}$ the MSE is reduced only by a factor of 15.6% and 4% when compared to $n = 2^7$ and $n = 2^9$ respectively. Similar numbers have been observed for the other cases. This slight decrease in MSE is mainly due to the reduction of the prediction variance as the bias remains high regardless of the sample size. In case-1, the bias at $n = 2^{14}$ is only smaller by 8.8% and 2.9% when compared against $n = 2^7$ and $n = 2^9$. The bias contributes about $97.52\% - 99.89\%$ to the MSE for $n \geq 2^9$. This range slightly changes to $98.1\% - 99.99\%$ for the other cases.

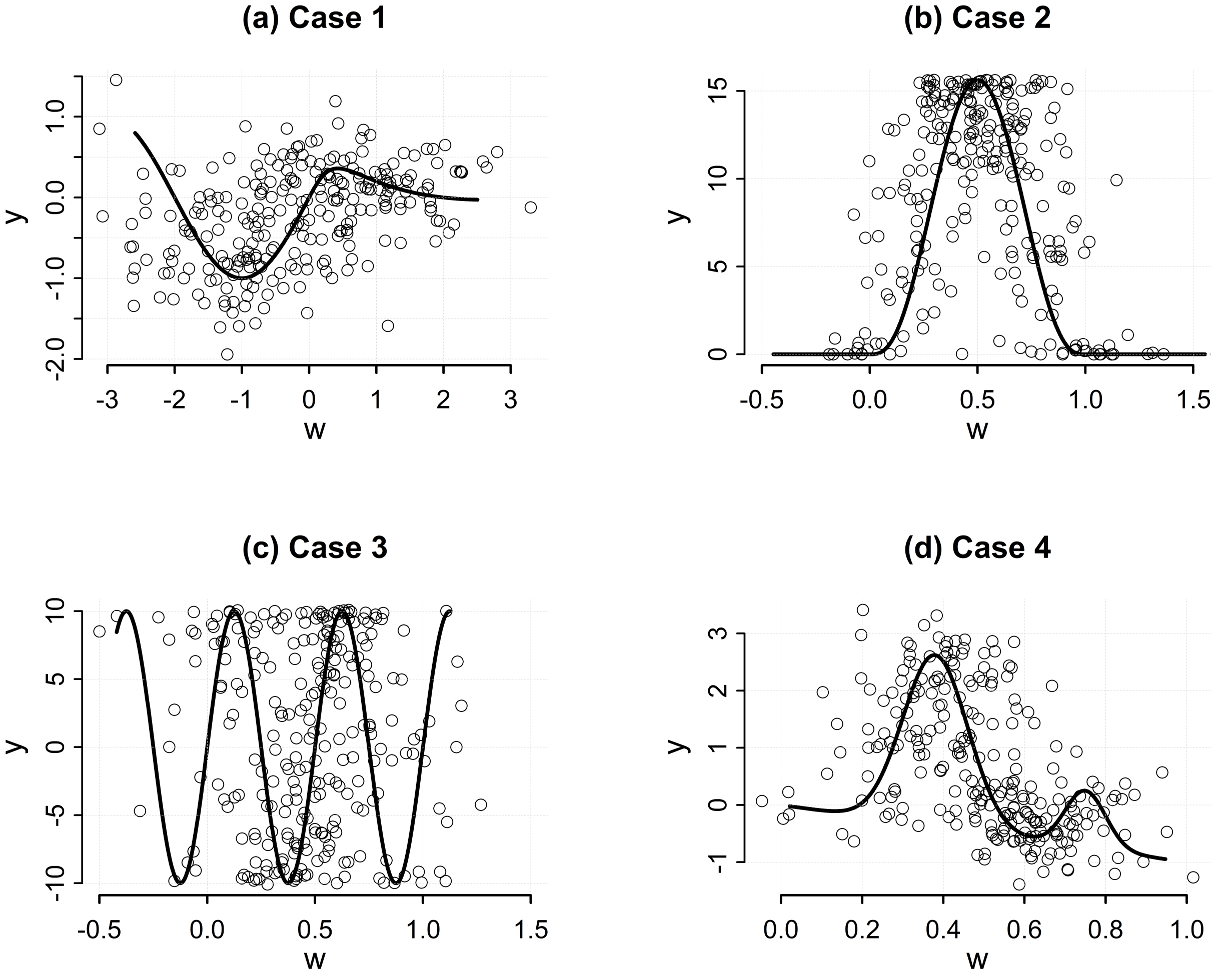

Figure 1. The Regression functions with a simulation example of the observed data for cases 1–4.

As shown in Figure 2, the performance of the conditional approach largely improves as n increases or in other words the empirical rates of convergence improves with n as specified in Section 3. It corrects for the bias even for small samples.

The conditional approach has a far superior performance than the naive estimators and all local polynomial estimators in all cases and sample sizes including the small ones. Also, as compared to the Bayesian approach, it has a better performance in Cases 2 and 3 and showed a competitive performance in Cases 1 and 4. A pointwise MSE assessment for the boundary, critical

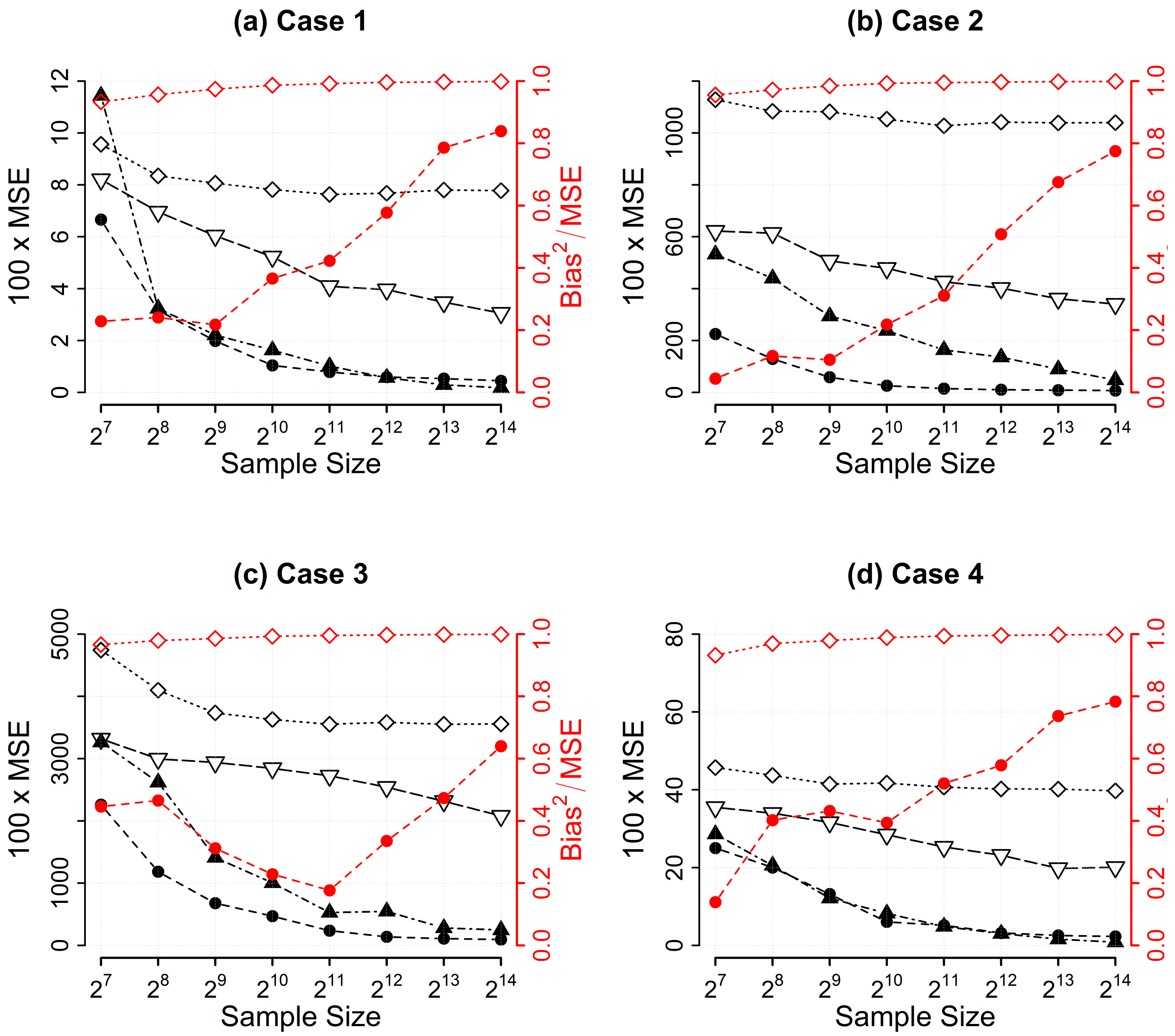

Figure 2. MSE (black) for cases 1–4 computed for the naive approach (open diamonds), the BLPE (open triangles), the Bayesian approach (solid triangles), and the conditional approach (solid circles). The squared bias/MSE is shown in red for the naive and the conditional estimators.

and inflection points for Case 1 is given in Appendix C.

The nonparametric convergence rate without measurement error is MSE $= Kn^{-\alpha}$ for some $\alpha > 0$. Then log(MSE) should be approximately linear in $\log n$. With normally-distributed measurement error, the optimal convergence rate is $K\log(n)^{-\alpha}$. Ideally, log(MSE) should be

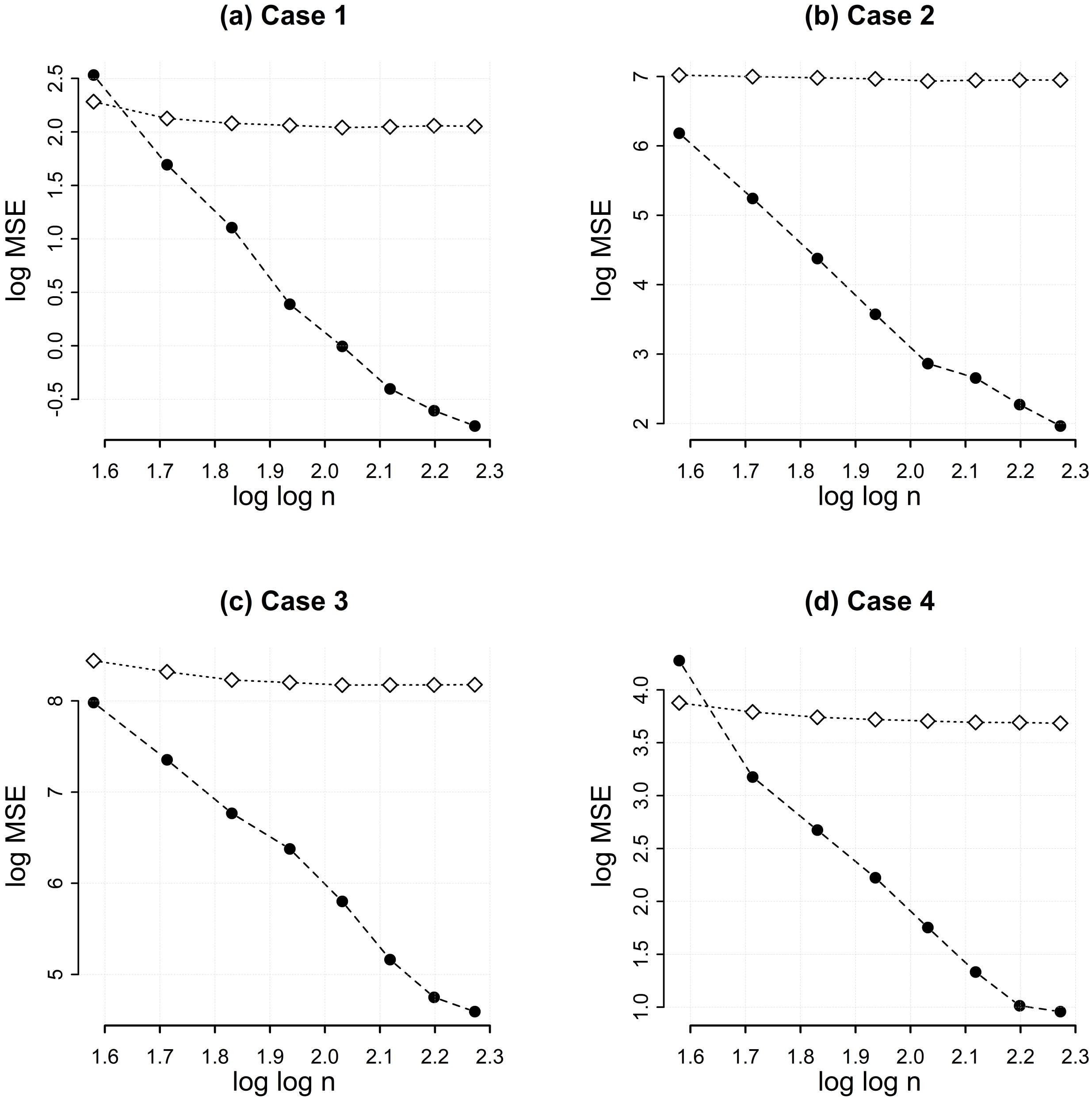

Figure 3. $\log\log n$ versus $\log(\text{MSE})$ for cases 1–4. The conditional approach is presented by the solid circles whereas the naive approach are shown by the open diamonds.

approximately linear in $\log\log n$. Figure 3 displays the log(MSE) for the conditional approach versus $\log\log n$ for all cases. Note that the curves for the conditional approach largely follow linear patterns with a weak evidence of curvature.

To test the robustness to model specification, the simulation study is repeated with x sampled from a skew normal distribution (Azzalini 1985 and Azzalini 2013) with shape parameter $\alpha = 6$ (long right-tailed distribution) and is repeated again with x sampled from a uniform distribution (short tailed distribution). In both scenarios, the mean and variance of x are the same as specified in cases 1–4. Notice that for local polynomial estimators no assumptions have been made about the distribution of x. The conditional approach and the Bayesian approach are fitted mistakenly assuming normality of x. The results for are shown in Figure 4 (skew normal) and Figure 5 (uniform).

The MSE for both the conditional approach and Bayesian approach have evidently increased. The structure of the mean function in the conditional approach is computed assuming normality of x. Mistakenly assuming normality induced mean misspecification in $E(\boldsymbol{Y}|\boldsymbol{w}, \boldsymbol{u})$ causing additional bias to incur. The performance of local polynomial estimators has deteriorated as well although there is no assumption has been made about the distribution of x in these methods.

Despite the serious departure from the normality assumption of x, the conditional approach remained superior to the naive approach and to the local polynomial estimators across all cases and sample sizes under both distributions. Also, the conditional approach maintained its dominance over the Bayesian approach in Cases 2 and 3. In addition to that, it showed better performance in Case 4 when sampling from skew normal. The Bayesian approach performed better in Case 1 under both sampling distributions.

As mentioned before, the BLPE may report a different estimator at each sample size n. In total there are 96 combinations of sample sizes, cases and sampling distributions. For example in Case 1 under normality, at $n = 2^7$ BLPE is the Staudenmayer and Ruppert (2004) estimator whereas

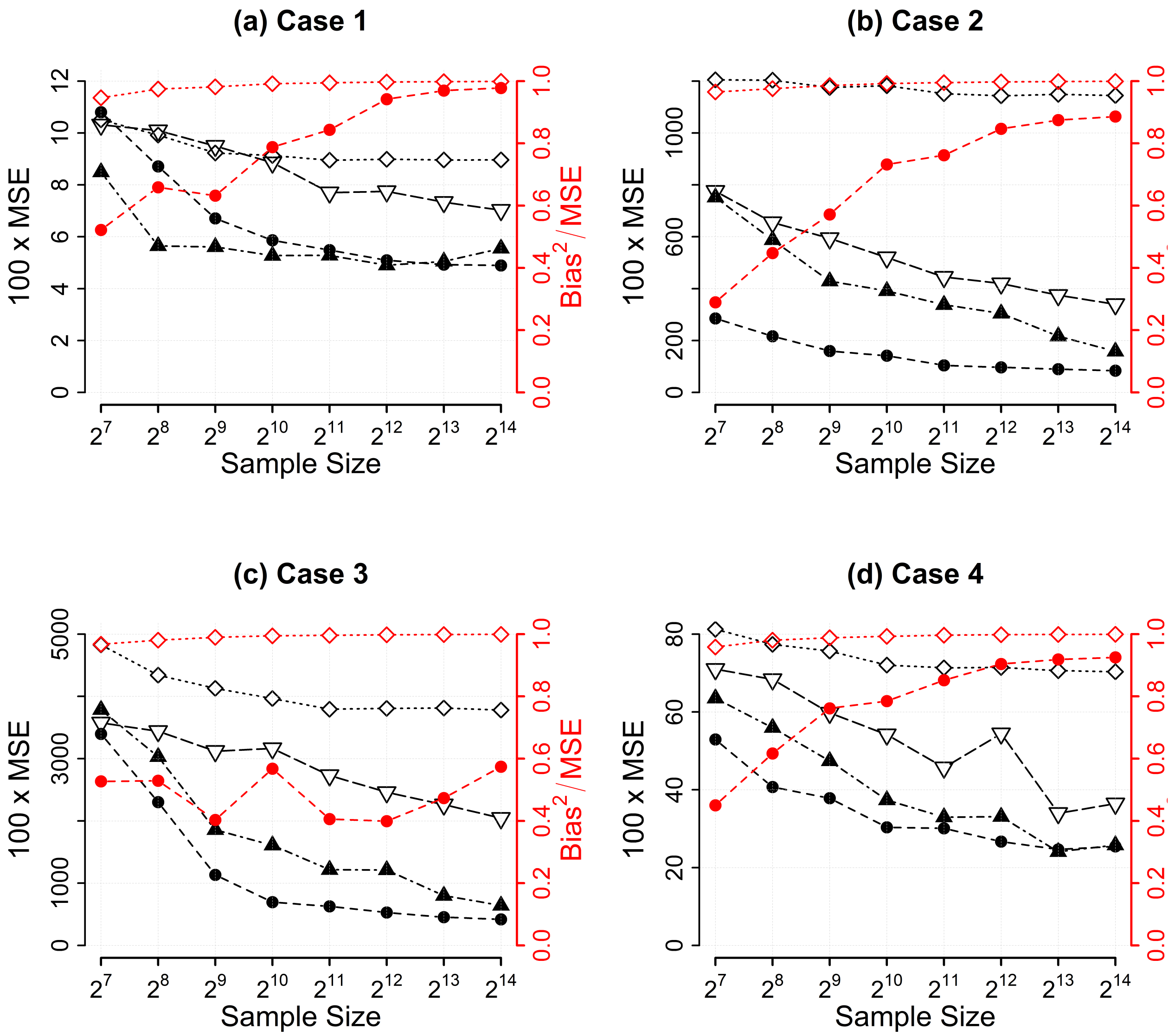

Figure 4. MSE results when x is sampled from skew normal distribution. MSE (black) for cases 1–4 computed for the naive approach (open diamonds), the BLPE (open triangles), the Bayesian approach (solid triangles), and the conditional approach (solid circles). The squared bias/MSE is shown in red for the naive and the conditional estimators.

at $n = 2^8$ and $n = 2^9$ is Delaigle et al. (2009) method with plug-in bandwidth selector and for $n \geq 2^{10}$, Delaigle et al. (2009) with Huang and Zhou (2017) bandwidth selector is reported. Again in Case 1 but now when sampling x from a skew normal distribution, the minimum is achieved at Fan and Truong (1993) method at all sample sizes. Generally speaking, despite its complexity,

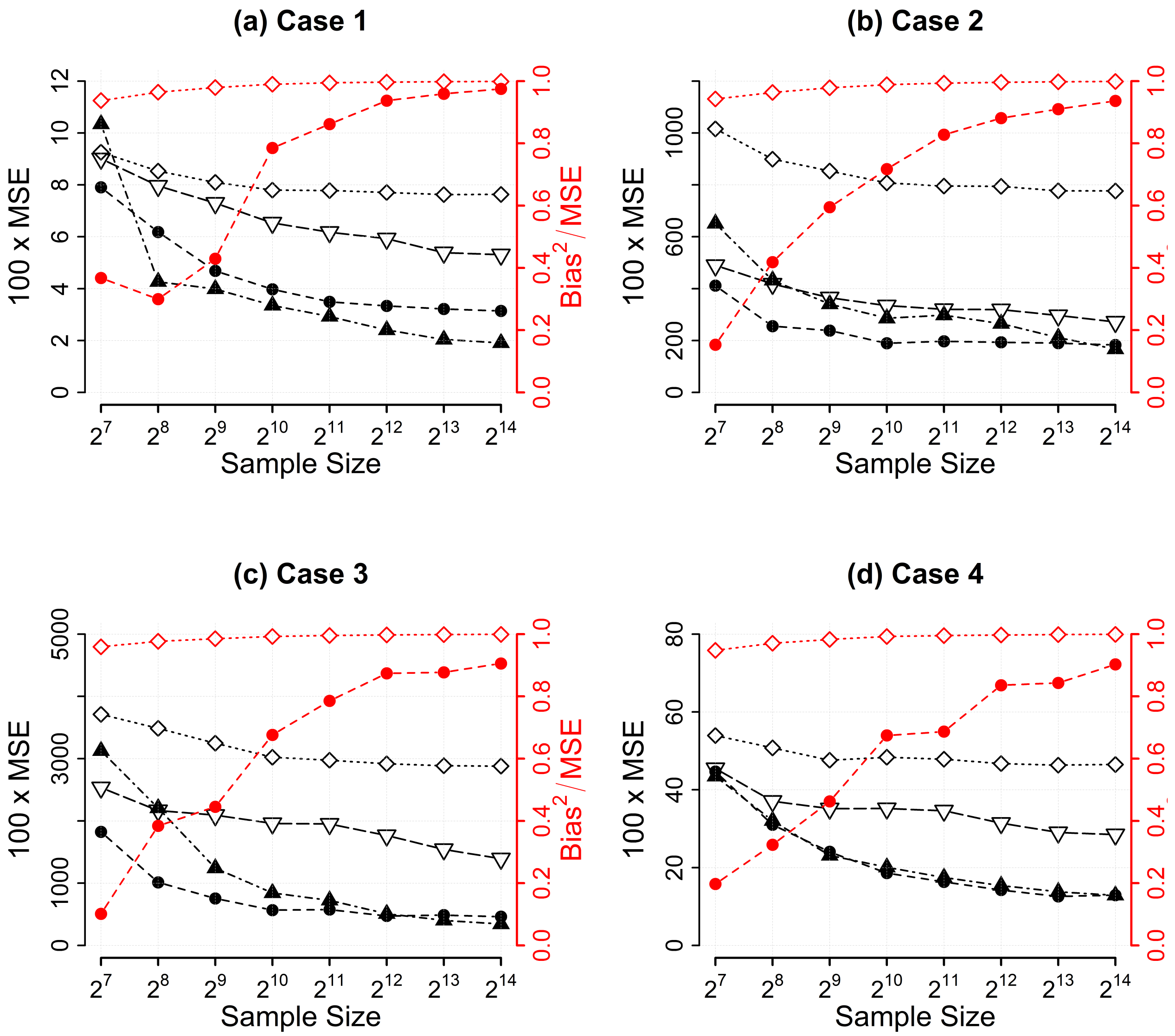

Figure 5. MSE results when x is sampled from uniform distribution. MSE (black) for cases 1–4 computed for the naive approach (open diamonds), the BLPE (open triangles), the Bayesian approach (solid triangles), and the conditional approach (solid circles). The mean squared bias/MSE is shown in red for the naive and the conditional estimators.

the Huang and Zhou (2017) bandwidth selector, which follows Delaigle and Hall (2008), produced unstable results with both Delaigle et al. (2009) and Huang and Zhou (2017) estimators. Out of the 96 combinations, this bandwidth selector appeared six times only all with Delaigle et al. (2009) estimator. Huang and Zhou (2017) estimator was selected once when using the plug-in bandwidth selector. Staudenmayer and Ruppert (2004) was selected 33 times (mostly with $n \leq 2^{12}$), followed by Fan and Truong (1993) estimator (31 times) and Delaigle et al. (2009) estimator when using the plug-in method (25 times).

We demonstrated through this simulation study that the conditional approach is superior to the best local polynomial estimator even when the model is incorrectly specified and is competitive with the Bayesian approach.

4.1 Comparisons with Sarkar et al. (2014)

Ideally we would like to include Sarkar's et al (2014) method, henceforth SMC, in the simulation study conducted in the previous section. However, this method is very computationally intensive and timely consuming, prohibiting its use in our simulations. For example, on a single dataset with $n = 2^{10}$ only, it took about 40 minutes to execute. Instead, a light version of the simulation study is performed using 40 simulated datasets with $n = 2^9$.

As recommended by Sarkar et al. (2014), we run 10000 MCMC iterations including a burn-in period of 5000 before applying thinning of factor 5. The R-code to execute SMC can be found in the journal website where the research is published. Three replicates are generated since the implementation of SMC depends on availability of replicates to estimate $\sigma^2_{w|x}$ which previously was assumed to be known. The conditional approach estimates $\sigma^2_{w|x}$ using the pooled sample variance of $\boldsymbol{w}$. The parameter σ^2_x is estimated by the average of the sample variances of replicates of $\boldsymbol{w}$ minus the estimate of $\sigma^2_{w|x}$. Then the algorithm runs separately on each replicate and predictions are generated before they are combined by taking their average. We found that this procedure can yield more accurate predictions than applying the procedure once on the average of the replicates.

Table 1

$100 \times$ MSE results for the conditional approach and Sarkar's et al. (2014) method (SMC). The comparison is based on $n = 2^9$ and involves cases 1 & 2 and three distributions of x. Three replicates are available.

	Conditional	SMC
Case 1		
Normal	1.18	1.20
Skew Normal	5.80	9.51
Uniform	4.34	3.10
Case 2		
Normal	27.64	2797.58
Skew Normal	128.22	3146.27
Uniform	171.59	2597.85

The results are shown in Table 1.

Under Case–1, the conditional approach has a comparable performance with SMC when x has a normal distribution and a better performance when x is sampled from skew normal distribution although it incorrectly specifies the distribution of x. Note that SMC does not make any distributional assumption on the distribution of x. SMC yielded more accurate predictions when x is sampled from uniform distribution. Under Case–2, the conditional approach has evidently produced superior performance. The improvement in MSE over the SMC method approximately ranged between 15 times to 101 times.

In almost all cases and distributions, the performance of the conditional approach has improved in presence of replications despite $\sigma^2_{w|x}$ being unknown. The improvement is minimal had we decided to take a single run on the data and perform the algorithm on the average of the replicates.

Table 2 reports the average time in seconds to execute the conditional approach when $\sigma^2_{w|x}$ is known (no replicates) and $\sigma^2_{w|x}$ is unknown (three replicates), and SMC (three replicates) on a

Table 2

Average calculation times in seconds for the SMC method and the conditional approach when $\sigma^2_{w|x}$ is known and when $\sigma^2_{w|x}$ is unknown.

Sample Size	2^7	2^9	2^{10}	2^{14}
SMC	239.62	755.45	2360.04	
Conditional ($\sigma^2_{w|x}$ known)	12.89	13.02	20.150	177.70
Conditional ($\sigma^2_{w|x}$ unknown)	18.91	22.68	24.112	197.41

single dataset. SMC's computation time when $n = 2^7$ exceeds the conditional approach computing time when $n = 2^{14}$ regardless $\sigma^2_{w|x}$ is known or not. The conditional approach is about 98 times faster to run than SMC at $n = 2^{10}$.

5 Application

Berry et al. (2002) analyzed a clinical trial experiment comparing a treatment group to a control group. The response variable is a score representing severity of a disease with lower scores indicating more severe disease. The main interest is estimating the score difference between the two groups adjusting for the baselines scores that are believed to covary non-linearly with the response variable. Basically, the model contains a dummy variable representing the group and 30 spline terms in the baseline score variable along their interactions with the dummy variable. The measurement error variance in the baselines scores is estimated at 0.35. The sample consists of 490 patients distributed evenly between the two groups.

Figure 6 shows the fitted curves representing the treatment effect (difference in the average scores between the treatment and the control groups) across various values of baselines scores. Notice that the naive approach that ignores measurement error concludes that treatment has a negligible effect on the severity of the disease in contrast to the Bayesian approach and the

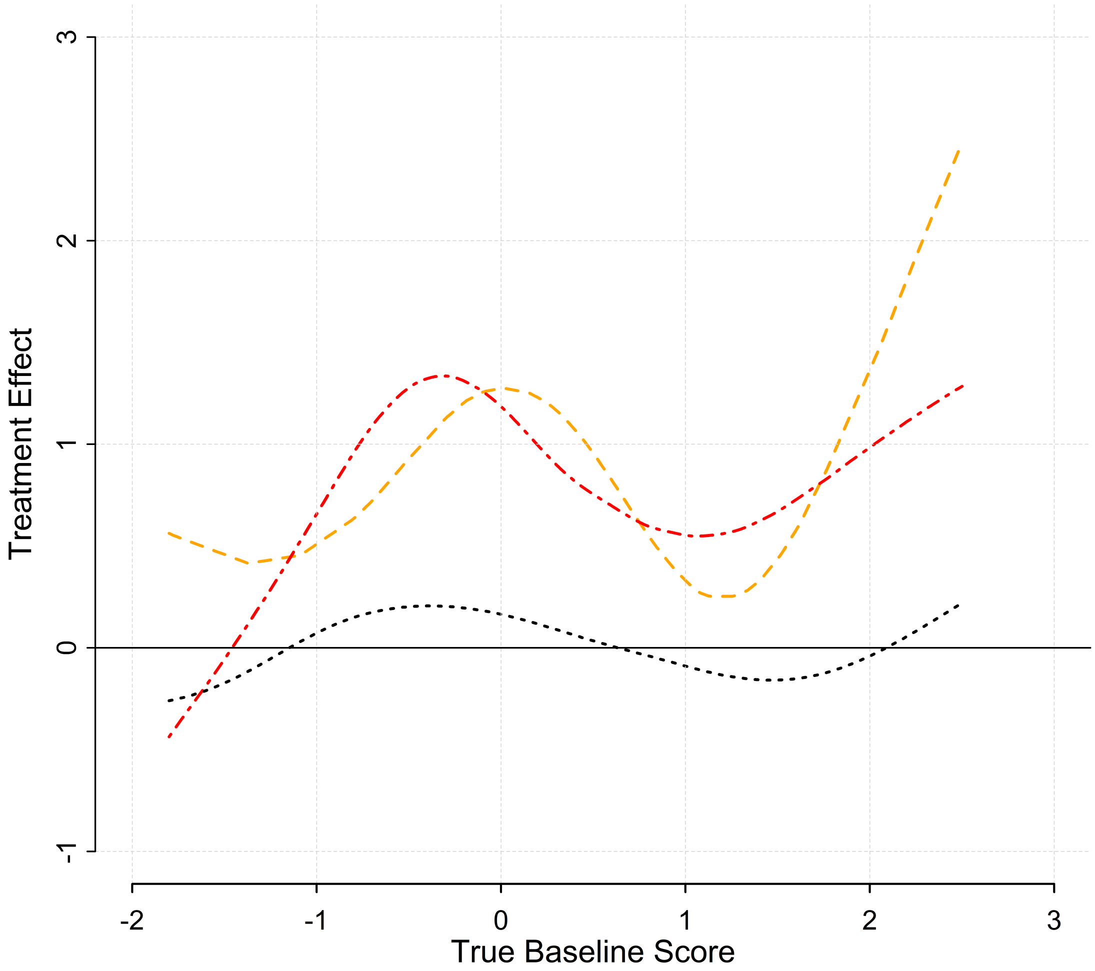

Figure 6. Estimating treatment effect utilizing naive approach (dotted lines), the Bayesian approach (dot-dashed lines), and the conditional approach (dashed lines).

conditional approach both of which agree to a large extent but markedly differ on the boundaries.

6 Conclusions

Ignoring measurement error in predictors may lead to highly biased inferences. The estimated regression function will misrepresent the true functional relationship between the response variable and the predictors. We demonstrated that the bias that results from failing to take into account measurement error remains huge regardless of the sample size. For this purpose, we devised an iterative frequentist approach that depends on modelling the conditional mean and the variance of the response variable given the observed predictor and the random effects. We showed through simulations that the conditional method largely outperformed the naive approach and the best local polynomial estimator. This remained true even when the conditional approach mistakenly assumed the normality of x. The conditional approach also displayed a competitive performance with the Bayesian approach and in some cases it was superior.

As stated earlier, this work can be naturally extended to additive models when more than one predictor exists. It may be some or all of the predictors are measured with errors. Inclusion of interactions may complicate the mathematics but remains feasible with few additional assumptions regarding the joint relationships between the predictors. Another future direction includes utilizing other basis functions such as O'Sullivan splines or B-splines rather than truncated splines could be tried. A far more interesting direction is exploring how to generalize methods presented in this article to account for measurement errors in logistic and Poisson regression models. Finally, we assumed that x has a normal distribution. The family of skew normal distributions is a very rich family that includes normal distributions and long tailed distributions that are frequently seen in practice. It is desirable for future work to extend the methodology presented here to when x follows a skew normal distribution. Also, a mixture of normal is another family of interest.

Appendix A: Proof of Theorem 2

We will demonstrate (10) only since (7), (8), and (9) will follow similarly. First since $b > a$,

$$\begin{aligned}\int_{-\infty}^{+\infty}(x-a)_+(x-b)_+f(x)\,\mathrm{d}x &= \int_{b}^{+\infty}(x-a)(x-b)f(x)\,\mathrm{d}x\\ &= \int_{b}^{+\infty}(x-\mu+\mu-a)(x-\mu+\mu-b)f(x)\,\mathrm{d}x\\ &= \int_{b}^{+\infty}(x-\mu)^2f(x)\,\mathrm{d}x+(2\mu-a-b)\int_{b}^{+\infty}(x-\mu)f(x)\,\mathrm{d}x\\ &+(\mu-a)(\mu-b)(1-F(b)) \qquad (20)\end{aligned}$$

The first term in (26)

$$\begin{aligned}\int_{b}^{+\infty}(x-\mu)^2f(x)\,\mathrm{d}x &= \int_{b}^{+\infty}(2\pi s^2)^{-1/2}(x-\mu)^2\exp(-(x-\mu)^2/(2s^2))\,\mathrm{d}x\\ &= \int_{b-\mu}^{+\infty}(2\pi s^2)^{-1/2}t^2\exp(-t^2/(2s^2))\,\mathrm{d}t\\ &= -(2\pi s^2)^{-1/2}s^2t\exp(-t^2/(2s^2))\Big|_{b-\mu}^{+\infty}+s^2\int_{b-\mu}^{+\infty}(2\pi s^2)^{-1/2}\exp(-t^2/(2s^2))\,\mathrm{d}t\\ &= (2\pi s^2)^{-1/2}s^2(b-\mu)\exp(-(\mu-b)^2/(2s^2))+s^2(1-F(b))\\ &= s^2(b-\mu)f(b)+s^2(1-F(b)) \qquad (21)\end{aligned}$$

Now, the second term in (26)

$$\begin{aligned}(2\mu-a-b)\int\limits_{b}^{+\infty}(x-\mu)f(x)\,\mathrm{d}x &= (2\mu-a-b)\int\limits_{b}^{+\infty}(2\pi s^2)^{-1/2}(x-\mu)\exp(-(x-\mu)^2/(2s^2))\,\mathrm{d}x \\ &= (2\mu-a-b)\int\limits_{(b-\mu)^2/2}^{+\infty}(2\pi s^2)^{-1/2}\exp(-t/s^2)\,\mathrm{d}t \\ &= (2\mu-a-b)(-s^2(2\pi s^2)^{-1/2}\exp(-t/s^2))\Big|_{(b-\mu)^2/2}^{+\infty} \\ &= (2\mu-a-b)s^2 f(b) \qquad (22)\end{aligned}$$

Substituting (27) and (28) in (26) gives (10).

Appendix B: Cov($\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u}$)

Using result 1, the (i,j) entry of Cov$(\boldsymbol{X\beta}|\boldsymbol{w})$ is

$$\mathrm{Cov}(\beta_0+\beta_1 x_i,\beta_0+\beta_1 x_j|\boldsymbol{w}) = \beta_1^2\mathrm{Cov}(x_i,x_j|\boldsymbol{w}) = \begin{cases} 0 & i\neq j \\ \beta_1^2\frac{\sigma_x^2\sigma_{w|x}^2}{\sigma_x^2+\sigma_{w|x}^2} & i=j\end{cases} \qquad (23)$$

Cov$(\boldsymbol{X\beta}|\boldsymbol{w})$ is a diagonal matrix since the x_i's are independent given $\boldsymbol{w}$. Similarly, the ith element of the vector $\boldsymbol{Zu}$ is a function of x_i only. Therefore, Cov$(\boldsymbol{Zu}|\boldsymbol{w},\boldsymbol{u})$ is a diagonal matrix as well. The (i,i) entry of this matrix is

$$\mathrm{Cov}(\boldsymbol{Zu}|\boldsymbol{w},\boldsymbol{u})_{ii} = \mathrm{Var}\left(\sum_{j=1}^{k}u_j(x_i-k_j)_+|w_i,\boldsymbol{u}\right) \qquad (24)$$

More compactly using vectors, (24) can be re-expressed as

$$\mathrm{Cov}(\boldsymbol{Z}\boldsymbol{u}|\boldsymbol{w},\boldsymbol{u})_{ii} = \boldsymbol{u}^\top \mathrm{Cov}(\boldsymbol{z}_i|w_i)\boldsymbol{u} \tag{25}$$

where z_i is the ith row vector of $\boldsymbol{Z}$. Specifically, $\boldsymbol{z}_i = \big[(x_i-k_1)_+,\ \ldots,(x_i-k_k)_+\big]^\top$. The (l,r) entry of $\mathrm{Cov}(\boldsymbol{z}_i|w_i)$

$$\mathrm{Cov}\big((x_i-k_l)_+,(x_i-k_r)_+|w_i\big) = \mathrm{E}\big((x_i-k_l)_+(x_i-k_r)_+|w_i\big) - \mathrm{E}\big((x_i-k_l)_+|w_i\big)\mathrm{E}\big(x_i-k_r)_+|w_i\big) \tag{26}$$

The last term can be found similar to (12) whereas the first term, by (10),

$$\begin{aligned}\mathrm{E}\big((x_i-k_l)_+(x_i-k_r)_+|w_i\big) &= \frac{\sigma_x^2\sigma_{w|x}^2}{\sigma_x^2+\sigma_{w|x}^2}f_i(k_r)\left(\frac{\sigma_x^2 w_i+\mu_x\sigma_{w|x}^2}{\sigma_x^2+\sigma_{w|x}^2}-k_l\right)\\ &+ \left(\frac{\sigma_x^2\sigma_{w|x}^2}{\sigma_x^2+\sigma_{w|x}^2}+\left(\frac{\sigma_x^2 w_i+\mu_x\sigma_{w|x}^2}{\sigma_x^2+\sigma_{w|x}^2}-k_l\right)\times\left(\frac{\sigma_x^2 w_i+\mu_x\sigma_{w|x}^2}{\sigma_x^2+\sigma_{w|x}^2}-k_r\right)\right)\\ &\times \big(1-F_i(k_r)\big)\end{aligned} \tag{27}$$

assuming $l \le r$.

Finally, similar to above, $\mathrm{Cov}(\boldsymbol{X\beta},\boldsymbol{Z}\boldsymbol{u}|\boldsymbol{w},\boldsymbol{u})$ is a diagonal matrix since the ith element for both vectors $\boldsymbol{X\beta}$ and $\boldsymbol{Z}\boldsymbol{u}$ depends on x_i alone. The (i,i) entry of this matrix is

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{X\beta},\boldsymbol{Z}\boldsymbol{u}|\boldsymbol{w},\boldsymbol{u})_{ii} &= \mathrm{Cov}(\beta_0+\beta_1 x_i,\boldsymbol{z}_i^\top\boldsymbol{u}|w_i,\boldsymbol{u})\\ &= \beta_1\mathrm{Cov}(x_i,\boldsymbol{z}_i|w_i)\boldsymbol{u}\\ &= \beta_1\sum_{j=1}^k u_j\mathrm{Cov}\big(x_i,(x_i-k_j)_+|w_i\big)\end{aligned} \tag{28}$$

which can be directly found using (9).

Let v_i be (i,i) entry of the $\mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u})$. Putting (5), (23), (27) and (28) together concludes that $\mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{u})$ is a diagonal matrix with v_i being

$$v_i = \sigma_e^2 + \beta_1^2 \frac{\sigma_x^2 \sigma_{w|x}^2}{\sigma_x^2 + \sigma_{w|x}^2} + 2\beta_1 \mathrm{Cov}(x_i, \boldsymbol{z}_i|w_i)\boldsymbol{u} + \boldsymbol{u}^\top \mathrm{Cov}(\boldsymbol{z}_i|w_i)\boldsymbol{u} \tag{29}$$

Appendix C: Varying the Number of Knots

Figure 6 shows the MSE performance for the conditional approach for Case 1 while varying k the number of knots. It basically repeats the analysis in panel (a) of Figure 2 for the conditional approach with $k = 10, 20, 30, 40$ and 50. Except when n is small, the performance of the estimator is slightly affected by choice of k.

Appendix D: MSE Pointwise Assessment

Figure 7 provides pointwise MSE for six points for the conditional and the Bayesian approaches for Case 1. Two points are the on the boundary of the grid $\{-2, 2\}$, three critical points $\{-1, 0, 0.43\}$, and one inflection point at 0.81.

References

[1] Azzalini, A. (1985), "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics,* 12, 171–178.

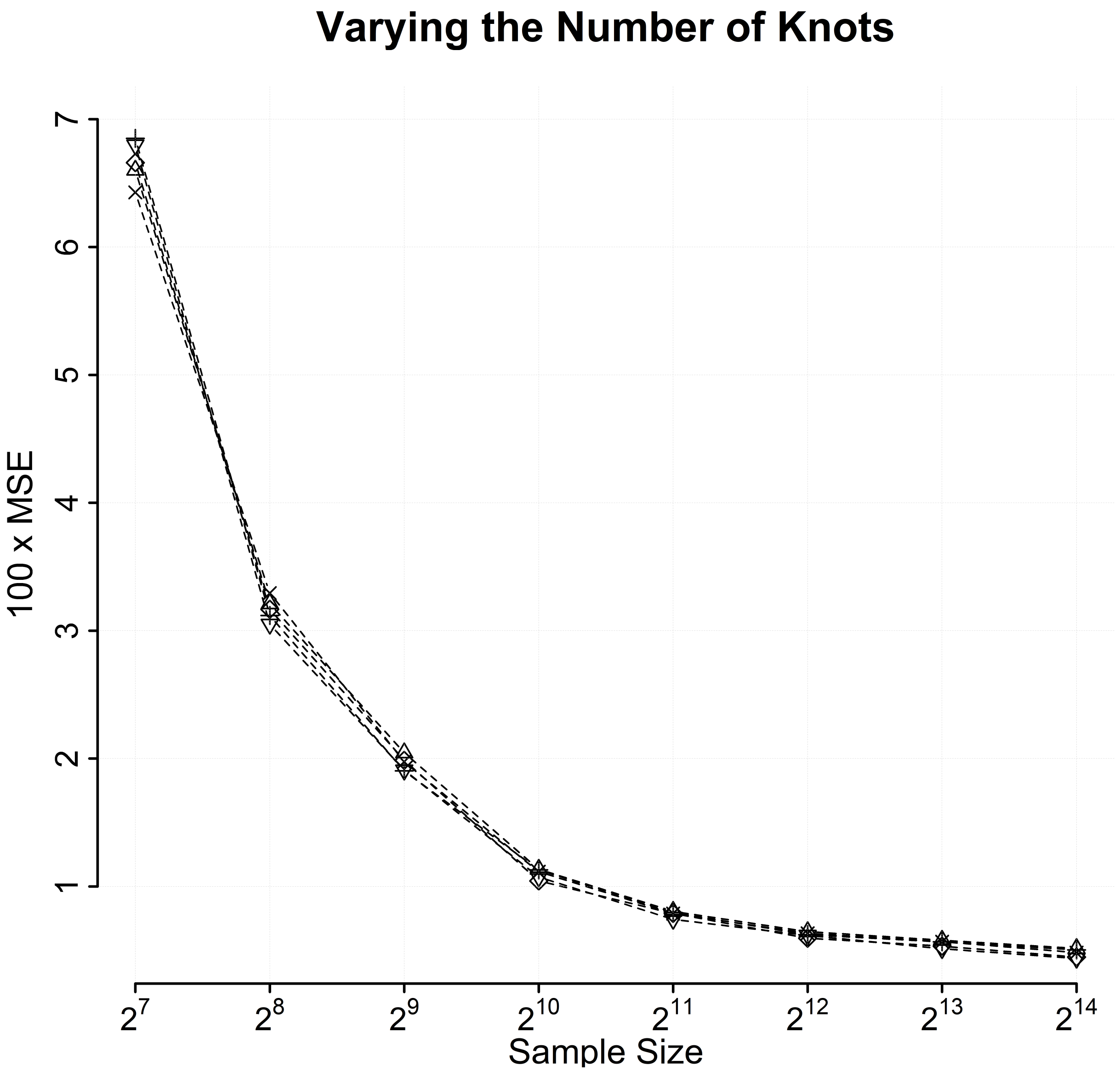

Figure 7. MSE for Case 1 computed for the conditional approach for different number of knots; $k = 10$ (triangle), $k = 20$ (plus), $k = 30$ (cross), $k = 40$ (diamond) and $k = 50$ (inverted triangle)

[2] Azzalini, A. (2013), *The skew-normal and related families,* Cambridge: Cambridge University Press.

[3] 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.

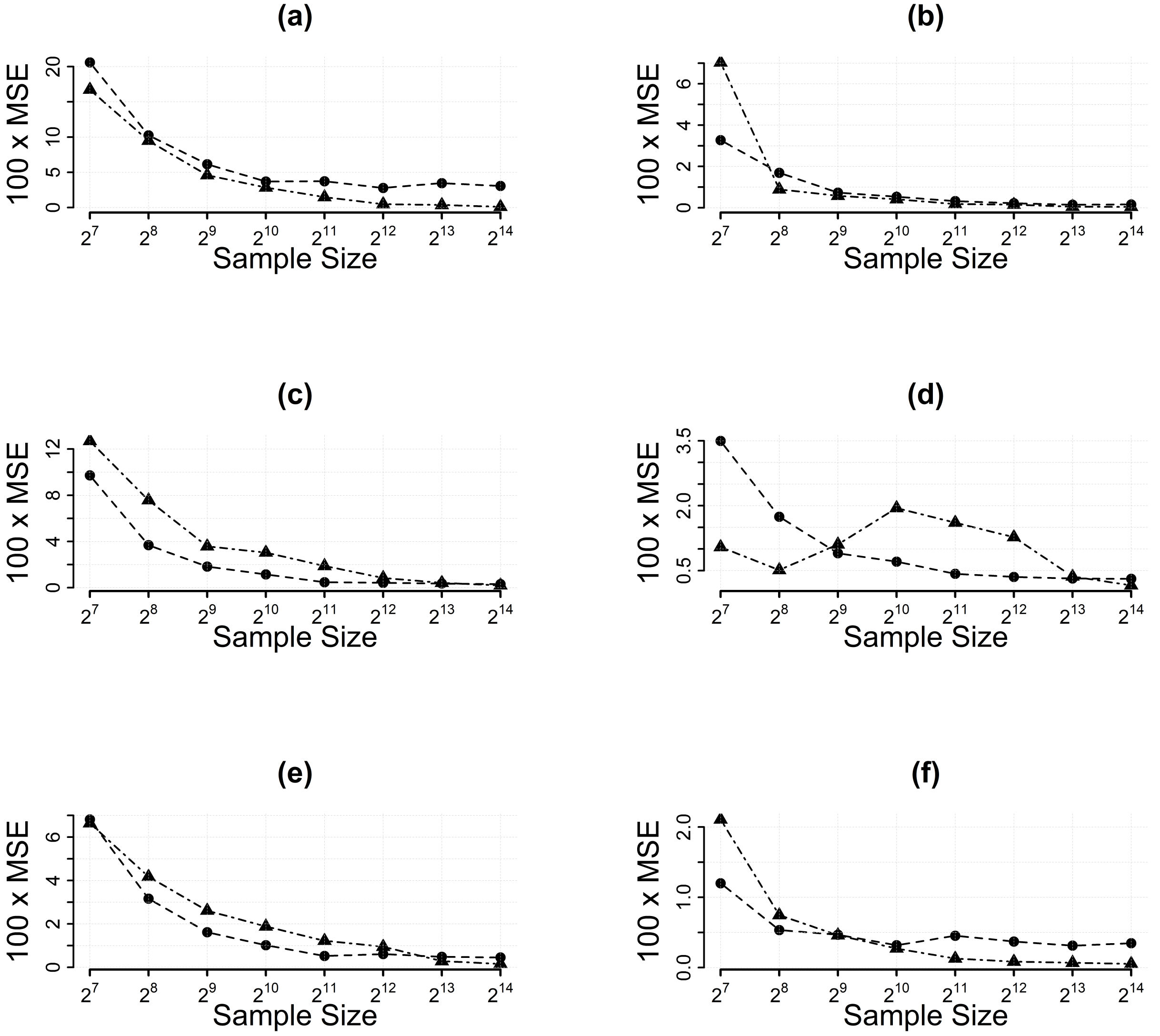

Figure 8. Pointwise MSE for Case 1 computed at six points for the Bayesian approach (solid triangles), and the conditional approach (solid circles). The boundary points $\{-2, 2\}$ are shown in panels (a) and (b), the critical points $\{-1, 0, 0.43\}$ are shown in panels (c),(d),(e), and the inflection point 0.81 is shown in (f).

[4] Carroll, R. J. (1989), "Covariance Analysis in Generalized Linear Measurement Error Models," *Statistics in Medicine,* 8 , 1075–1093.

[5] Carroll, R. J., Maca, J. D., and Ruppert, D. (1999), "Nonparametric Regression With Errors in Covariates," *Biometrika,* 86, 541–554.

[6] Carroll, R. J., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006), *Measurement Error in Nonlinear Models: a Modern Perspective (2nd ed.),* Boca Raton: Chapman and Hall.

[7] 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.

[8] Davidian, M., and Giltinan, D.M. (1995), *Nonlinear Models for Repeated Measurement Data,* New York: Chapman and Hall.

[9] Delaigle, A., and Gijbels, I. (2002), "Estimation of integrated squared density derivatives from a contaminated sample", *Journal of the Royal Statistical Society: Series B,* 64, 869–886.

[10] Delaigle, A., and Gijbels, I. (2004), "Practical bandwidth selection in deconvolution kernel density estimation", *Computational Statistics and Data Analysis,* 45, 249–267.

[11] Delaigle, A., Fan, J., and Carroll R.J. (2009), "A Design-Adaptive Local Polynomial Estimator for the Errors-in-Variables Problem", *Journal of the American Statistical Association,* 104, 348–359.

[12] Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing with B-Splines and Penalties" (with discussion), *Statistical Science,* 11, 89—102.

[13] Fan, J., and Truong, Y. K. (1993), "Nonparametric Regression with Errors in Variables", *The Annals of Statistics,* 21, 1900-–1925.

[14] Fuller, W.A. (1987), *Measurement Error Models,* New York: John Wiley & Sons.

[15] Ganguli, B., Staudenmayer, J., and Wand, M.P. (2005), "Additive Models with Predictors Subject to Measurement Error," *Australian & New Zealand Journal of Statistics,* 47, 193–202.

[16] Huang, X., and Zhou, H. (2017), “An alternative local polynomial estimator for the error-in-variables problem,” *Journal of Nonparametric Statistics,* 29, 301–325.

[17] Ruppert, D., and Carroll, R. J. (2000), “Spatially Adaptive Penalties for Spline Fitting,” *Australia and New Zealand Journal of Statistics,* 42, 205-–223.

[18] Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression,* Cambridge, UK: Cambridge University Press.

[19] Sarkar, A., Mallick, B. K., and Carroll, R. J. (2014), “Bayesian Semiparametric Regression in the Presence of Conditionally Heteroscedastic Measurement and Regression Errors,” *Biometrics,* 70, 823–834.

[20] Spiegelman, D., Rosner, B., and Logan, R. (2000), “Estimation and Inference for Logistic Regression with Covariate Misclassification and Measurement error, in Main Study/Validation Study Designs,” *Journal of the American Statistical Association,* 95, 51–61.

[21] Staudenmayer, J., and Ruppert, D. (2004). “Local polynomial regression and simulation–extrapolation,” *Journal of the Royal Statistical Society. Series B,* 66, 17–30.

[22] Wang, X., Shen, J., and Ruppert, D. (2011). “On the asymptotics of penalized spline smoothing,” *Electronic Journal of Statistics,* 5, 1–17.