Post Empirical Bayes Regression*

Sheng-Kai Chang
National Taiwan University

Yu-Chang Chen
National Taiwan University

Shuo-Chieh Huang
University of Chicago

Shen-Hsun Liao
National Taiwan University

February 29, 2024

Abstract

Empirical Bayes (EB) methods are widely utilized in economics for estimating individual and group-level fixed effects across diverse contexts, including teacher value-added, hospital qualities, and neighborhood effects. While estimates generated by EB are often incorporated into other statistical analyses like regression models, the econometric properties of post-EB regression have not been thoroughly investigated. This paper addresses this knowledge gap through two key contributions. First, we introduce a unified framework for two-step EB methods that applies to both linear and non-linear models, offering insights into their frequentist properties and assessing their robustness against model mis-specification. Second, we undertake a critical evaluation of commonly-used two-step EB methods in existing empirical research. Our analysis demonstrates that naive implementations of post EB regression can introduce a systematic bias, particularly in non-linear models.

Keywords: empirical Bayes, two-step methods, nonparametric empirical Bayes

*Yu-Chang Chen acknowledges the support from the National Science and Technology Council grant NSTC 112-2410-H-002-036-MY2. This draft is preliminary. Please do not share.

1 Introduction

Empirical Bayes (EB) methods are frequently employed for estimating fixed effects, particularly when the number of repeated observations per unit is limited. These methods offer a remedy to the incidental parameter problem (Neyman and Scott, 1948) often encountered in high-dimensional problems. Specifically, let $\hat{\theta}_i$ be the fixed effects estimates in a dummy variable regression. The EB estimators $\hat{\theta}_i^{EB}$, in its simplest form, shrink the raw fixed effects $\hat{\theta}_i$ to the "grand mean" $\bar{\bar{\theta}}$:

$$\hat{\theta}^{EB} = \lambda\hat{\theta}_i + (1-\lambda)\bar{\bar{\theta}},$$

where the shrinkage factor λ depends on the signal-to-noise ratio and the grand mean $\bar{\bar{\theta}}$ is the average of $\hat{\theta}_i$s. This approach has gained traction in empirical applications like teacher value-added models (see, e.g., Kane and Staiger, 2008; Chetty et al., 2014a; Jackson et al., 2014; Koedel et al., 2015), where the student-to-teacher ratio is often small. Other prominent applications in economics include neighborhood effects (Chetty and Hendren, 2018) and hospital qualities (Hull, 2018).

Beyond estimating the fixed effects per, EB estimates are frequently used as inputs for further statistical analyses. For instance, in studies of teacher qualities, researchers may regress students' labor market outcomes on the EB estimates of teacher qualities to quantify the long-term impact of teacher quality. The regression can be linear models that include transformed EB estimates or nonlinear models like logistic regression for binary dependent variable. We refer to the utilization of EB estimates in regression analyses as "post empirical Bayes regression".

In this paper, we introduce a general class of post empirical Bayes regression methods that enables researchers to consistently estimate both fixed effects and regression coefficients. Our methodology starts with the estimation of fixed effects and their underlying distributions, employing either nonparametric or parametric EB techniques. We then use the estimated

distribution of fixed effects—not the EB estimates themselves—to construct an estimator for the regression model. A key advantage of our method is its coherence and convenience, as it allows for the simultaneous estimation of fixed effects, their distributions, and the relevant regression coefficients. Additionally, our approach accommodates nonparametric EB and nonlinear regression while also allowing for dependent measurement errors, thereby offering a flexible and robust framework for empirical analysis.

Another key objective of this paper is to scrutinize empirical practices of post-EB regression. To this end, we formulate a general framework for two-step EB methods and employ it to assess common practices observed in existing literature. We pay special attention to cases where EB estimates serve as explanatory variables in linear regressions. Our findings reveal that while the two-step EB method applied to ordinary least squares (OLS) yields consistent estimators, the standard errors can be significantly downward-biased if corrections for generated regressors are not made. Consequently, t-tests conducted without these corrections may over-reject hypotheses, even with large sample sizes. In the context of non-linear models, we demonstrate that directly using EB estimates as explanatory variables can result in inconsistent estimators.

This paper will proceed as follows. In Section 2, we introduce use a simple example to illustrate the main idea of the method. Section 3 describes the general setup and the proposed method, and we present the theoretical results in 4. Simulation studies and empirical applications of the method are in Section 5.

1.1 Literature Review

From a theoretical standpoint, this project closely aligns with the errors-in-variables literature in econometrics. Interestingly, EB procedures can consistently estimate linear regression models with measurement errors, even though they were originally designed for different applications. This property was not first observed by us; the concept of using shrinkage to address measurement errors dates back to Whittemore (1989). Subsequent work by Guo

and Ghosh (2012) formalized this insight and established its consistency, while Efron (2016) also proposes an empirical Bayes approach for deconvolution problems. Our work also intersects with the literature on Bayesian methods for handling measurement errors (see, e.g., Carroll et al., 2006), as we aim to develop a unified theory for two-step EB methods that encompasses both linear and non-linear regression models.

Additionally, our project contributes to the growing body of literature on EB methods in both econometrics and statistics (see, e.g., Hansen, 2017; Meager, 2019; Ignatiadis and Wager, 2019; Azevedo et al., 2020; Armstrong et al., 2022; Bonhomme and Weidner, 2022). Although the original idea can be traced back as early as to Robbins (1956) and James and Stein (1961), there has been a resurgence in the application of EB methods, particularly in high-dimensional problems (Efron, 2012) as well as its increasing popularity in empirical studies (see, e.g., Chetty et al., 2014a,b; Chetty and Hendren, 2018; Hull, 2018; Angrist et al., 2021). This paper seeks to rigorously examine the statistical properties of two-step EB methods employed in these recent empirical works, thereby offering a comprehensive econometric analysis.

2 A Simple Case: OLS Regression with EB Estimates

One of the most common form of post-EB regression is a linear regression with EB estimates as regressors. In this section, we will use it as an example to help illustrate the research question and provide some theoretical results that will be further extended to the general case later in this paper.

EB models, which are derived from Bayesian methods, have two components: (1) the prior distribution for the latent variables and (2) the likelihood function of the observed variables given the latent ones. Formally, we can write

$$\theta_i \overset{i.i.d.}{\sim} \pi(\cdot), \tag{1}$$

$$x_{ij}|\theta_i \overset{i.i.d.}{\sim} f_{x|\theta}(\cdot), \tag{2}$$

in which $\pi(\cdot)$ is the prior distribution of θ_i and $f_{x|\theta}(\cdot)$ is the likelihood function. The fixed effects are unobserved and are only noisily measured by the measurements $x_{ij}, j = 1, 2, ..., m$.[1] In applications of teacher value-added models, θ_i are the teachers' (indexed by i) effects and x_{ij} are the students' (indexed by j) test scores.

Depending on the context, various modeling assumptions are made on the conditional distribution $f_{x|\theta}(\cdot)$. In empirical studies, the most frequently encountered set of assumptions—either explicitly stated or implicitly applied—pertains to the imposition that

$$\theta_i \overset{i.i.d.}{\sim} N(\mu_\theta, \tau^2), \tag{3}$$

$$x_{ij}|\theta_i \overset{i.i.d.}{\sim} N(\theta_i, \gamma^2), \tag{4}$$

i.e., a normal conjugate model. Since it is a conjugate model, the posterior distribution of θ given $x_1, x_2, ..., x_m$ is also normal.

EB estimator are essentially Bayes rules except that the prior is "estimated" from the

[1] For the ease of exposition, we assume the panel is balanced.

data. In the normal conjugate model, the EB estimator is

$$\hat{\theta}_i^{EB} = \frac{\hat{\tau}^2}{\hat{\gamma}^2 + \hat{\tau}^2} \cdot \bar{x}_i + \frac{\hat{\gamma}^2}{\hat{\gamma}^2 + \hat{\tau}^2} \cdot \hat{\mu}_\theta, \tag{5}$$

where

$$\bar{x}_i = \frac{1}{m} \sum_{j=1}^{m} x_{ij} \tag{6}$$

$$\hat{\mu}_\theta = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}, \tag{7}$$

$$\hat{\tau}^2 = \frac{1}{n} \sum_{i=1}^{n} (\bar{x}_i - \hat{\mu}_\theta)^2 - \frac{1}{m} \hat{\gamma}^2, \tag{8}$$

and

$$\hat{\gamma}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m-1} \sum_{j=1}^{m} (x_{ij} - \bar{x}_i)^2. \tag{9}$$

Notice that $\hat{\theta}_i^{EB}$ is the EB estimator as it is the posterior mean of θ given $x_{i1}, x_{i2}, ..., x_{im}$ with the parameters $(\mu_\theta, \tau^2, \gamma^2)$ estimated by their empirical analogues. That is, $\hat{\theta}_i^{EB}$ is the Bayes rule with an estimated prior. [2]

In a typical application of EB in economics, researchers are often interested in the prior distribution $\pi(\cdot)$, the fixed effects θ_i as well as the impact of θ_i on some other variables, say y_{ij}. More formally, the researcher is interested the coefficient β in the regression

$$y_{ij} = \alpha + \theta_i \beta + u_{ij}. \tag{10}$$

In the context of teacher value-added, y_{ij} may be some measures of the students' long-term outcome such as college attendance or earnings in the labor market, and the researcher's

[2]Strictly speaking, while EB methods are inspired by the Bayesian methodology, EB estimators do not perfectly fit into Bayesian paradigm. In this paper, we view the EB as frequentist methods and examine its frequentist properties such as consistency and asymptotic normality.

goal is to estimate the impact of θ_i on y_{ij} given that θ_i is unobserved and is only noisily measured by $x_{i1}, x_{i2}, ..., x_{im}$.[3]

One approach that is commonly seen in empirical practice is to regress the outcome variable y_{ij} on the EB estimates $\hat{\theta}_i^{EB}$

$$\hat{\beta}^{OLS,EB} = \frac{\sum_{i=1}^{n}(\hat{\theta}_i^{EB} - \bar{\hat{\theta}}^{EB})(y_{ij} - \bar{\bar{y}})}{\sum_{i=1}^{n}(\hat{\theta}_i^{EB} - \bar{\hat{\theta}}^{EB})^2}, \tag{11}$$

where $\bar{\bar{y}}$ and $\bar{\hat{\theta}}^{EB}$ are averages of $\bar{y}_i$ and $\hat{\theta}_i^{EB}$ respectively. The idea of regression on $\hat{\theta}_i^{EB}$ as if they were the true unobserved θ_i is quite popular, and it goes beyond linear regression to nonlinear models such as the logistic regression. While this approach does have intuitive appeal and is convenient once we have the EB estimates, the statistical properties of it remain unclear. In this paper, we aim to fill this gap.

We start our analysis by inspecting the asymptotic properties of $\bar{\hat{\theta}}^{EB}$. Lemma 1 provides a useful observation to understand $\hat{\beta}^{OLS,EB}$ as well as the more general form of post-EB regression. Note that since $x_{ij}, j = 1, ..., m$ are independent, unbiased measurements of θ_i, we can write

$$x_{ij} = \theta_i + v_{ij}, \tag{12}$$

where $\mathrm{E}[v_{ij}|\theta_i] = 0$ and $v_{ij} \perp v_{ij'}$ for $j \neq j'$. If we invert the role between x_{ij} and θ_i and

[3]While the example shares similarities with configurations often seen in the measurement error literature, this paper diverges in two key respects. First, our scope goes beyond the simple estimation of model parameters to include a critical examination of current empirical practices. As noted earlier, EB methods are commonly used to address measurement errors across various applications, even though there is a relative lack of theoretical substantiation for their effectiveness. Second, our concerns are not limited to estimating the regression coefficient β; we are also interested in estimating the latent fixed effects θ_i as well its distribution $\pi(\cdot)$.

consider the population regression of θ_i on $\bar{x}_i$, we can write

$$\theta_i = a + b\bar{x}_i + \tilde{v}_i, \tag{13}$$

$$\mathbb{E}[\tilde{v}_i] = \mathbb{E}[\tilde{v}_i \bar{x}_i] = 0. \tag{14}$$

Then Lemma 1 implies that the fitted value from the inverted regression is in fact, $\hat{\theta}_i^{EB}$.

Lemma 1. *Let $\mu_\theta = \mathbb{E}[\theta_i]$ be the grand mean and $\lambda = \frac{Var(\theta_i)}{Var(v_{ij})/m + Var(\theta_i)}$ be the signal-to-noise ratio. Then the regression coefficients $a = (1-\lambda) \cdot \mathbb{E}[\theta_i]$ and $b = \lambda$.*

Lemma 1 implies that the EB estimators $\hat{\theta}_i^{EB}$ are the best linear predictors of θ_i given $\bar{x}_i$, and we can use $\hat{\theta}_i^{EB}$ as consistent estimators of the fitted values of θ_i even though we do not observe θ_i. Viewing $\hat{\theta}_i^{EB}$ as fitted values of θ_i suggests that the post-EB regression estimator $\hat{\beta}^{OLS,EB}$ resembles a two-stage least square (2SLS) estimator, which is consistent provided that $\bar{x}_i$ is a valid instrument for θ_i.[4] We formalize this observation in Proposition 1 by showing that $\hat{\beta}^{OLS,EB}$ can be framed as a generalized method of moments (GMM) estimator (Hansen, 1982).

Proposition 1. *$\hat{\beta}^{OLS,EB}$ is consistent and asymptotically normal.*

Proof. Consider the population linear projection of θ_i on $\bar{x}_i$

$$\theta_i = a + b\bar{x}_i + \tilde{v}_i,$$

in which the coefficients are given by

$$a = (1-\lambda)\mu_\theta = (1-\lambda)E[\theta_i],$$

[4] Chetty, Friedman Rockoff (2014) made a similar observation, in which they refer to the instrument validity as the "forecast unbiasedness". Our formulation emphasizes that EB estimates are the "fitted value" in the first stage and that correction for generated regressors is needed when calculating the standard errors.

and

$$b = \frac{Var(\theta_i)}{Var(\theta_i) + \frac{1}{m}Var(v_{ij})} = \lambda.$$

Replace θ_i with its fitted value $a + b\bar{x}_i$ in the regression model $y_{ij} = \alpha + \theta_i\beta + u_{ij}$, we get

$$\begin{aligned}\bar{y}_i &= \alpha + \beta[(1-\lambda)\mu_\theta + \lambda\bar{x}_i + \tilde{v}_i] + \bar{u}_i \\ &= \alpha + \beta\underbrace{[(1-\lambda)\mu_\theta + \lambda\bar{x}_i]}_{x_i^{EB}} + (\beta\tilde{v}_i + \bar{u}_i).\end{aligned}$$

Therefore, together with the orthogonality condition, we can reformulate $\hat{\beta}^{\text{OLS,EB}}$ as an exact-identified GMM estimator with the following moment conditions for parameters $(\alpha, \beta, \mu_\theta, \lambda)$:

$$\begin{aligned}&\mathbb{E}[\bar{x}_i - \mu_\theta] = 0, \\ &\mathbb{E}[\frac{1}{m(m-1)}(x_{ij} - \bar{x}_i)'(x_i - \bar{x}_i) - (\bar{x}_i - \mu_\theta)^2\lambda] = 0, \\ &\mathbb{E}[y_{ij} - \alpha - \theta_i\beta] = 0, \\ &\mathbb{E}[(y_{ij} - \alpha - \theta_i\beta)\theta_i] = 0.\end{aligned}$$

Under regularity conditions, the standard (non-linear) GMM Hansen (1982) theory implies that $\hat{\beta}^{\text{OLS,EB}}$ is consistent and asymptotically normal. □

We are not the first to notice the consistency of $\hat{\beta}^{OLS,EB}$. The idea of using James-Stein's estimator to correct measurement error can be traced back as early to Whittemore (1989). Guo and Ghosh (2012) calculates the mean-squared error of the two-step estimator proposed by Whittemore (1989), implying that the estimator is consistent as a side result.[5] Our result further establishes that $\hat{\beta}^{OLS,EB}$ is also asymptotically normal.

[5]Another way to see the consistency of $\hat{\beta}^{OLS,EB}$ is to notice that the shrinkage factor λ is exactly the attenuation bias if one regresses y on the noisy measurement x. Direct calculation of the probability limit of $\hat{\beta}^{OLS,EB}$ shows that the shrinkage factor turns out to cancel out the attenuation bias, making $\hat{\beta}^{OLS,EB}$ consistent.

2.1 Beyond Linear Models

The example in Proposition 1 is simple and restrictive in several ways. First, the EB estimator $\hat{\theta}_i^{EB}$ is derived from the normal conjugate model. While the normal conjugate model is popular among empirical applications, other models, such as a beta-binomial model to accommodate binary $\mathbf{x}$, may also be of interest. Second, the regression model considered in Proposition 1 is a simple linear regression, which contains only one independent variable and excludes nonlinear regression models.

In this paper, our goal is to develop a general method of post-EB regression that allows for flexible specifications of EB estimators and encompasses both linear and nonlinear regression. However, as we can see from the proof of Proposition 1, the post-EB regression for simple linear regression is consistent since

$$\mathbb{E}[y_{ij}|\bar{x}_i] = \alpha + \mathbb{E}[\theta_i|\bar{x}_i] \cdot \beta \tag{15}$$

$$= \alpha + \theta_i^{EB} \cdot \beta. \tag{16}$$

So in the regression, in which we regress y_{ij} on θ_i^{EB}, is correctly specified. In nonlinear regression models, the above equality would amount to require that

$$\mathbb{E}[g(\theta_i, \beta)|\bar{x}_i] = g(\mathbb{E}[\theta_i|\bar{x}_i], \beta),$$

where $g(\cdot)$ is the conditional mean function of y conditional on θ. That is, the posterior mean of the transformation has to be the transformation of the posterior mean. The equality does not hold in general. Consequently, post-EB regression, if done naively (i.e., EB estimates $\hat{\theta}_i^{EB}$ are directly used as regressors as if they were the unobserved θ_i), can lead to biased estimates. For example, one can show that, if $g(\cdot)$ is the logit link function and (x_{ij}, θ_i) follows normal-normal, then the equality is violated, leading to an inconsistent estimator of β.[6] Hence, the empirical practice currently observed in the literature may lead to erroneous

[6]However, for probit models, in which $g(\cdot)$ is the cumulative distribution function of the normal distribu-

results.

Nevertheless, the shortcomings of the "naive" post-EB regression should not be interpreted as a general inadequacy of EB methods for correcting measurement errors in nonlinear models. In the following section, we outline the "proper" post-EB regression that are applicable to nonlinear models. Additionally, our generalized approach extends beyond the normal conjugate model for EB estimators, permitting the distribution $\pi(\cdot)$ to be identified in nonparametrically.

3 General Setup

Below, we describe the general setup we consider in this paper. As in most EB applications, we assume the data $\{\{y_{ij}, \mathbf{x}_{ij}, \boldsymbol{z}_{ij}\}_{j=1}^{m}, \theta_i\}_{i=1}^{n}$ has a panel structure. The latent variable $\boldsymbol{\theta}_i \in \mathbb{R}^{\dim(\boldsymbol{\theta})}$ is drawn from an unknown distribution π

$$\boldsymbol{\theta}_i \overset{i.i.d.}{\sim} \pi, \tag{17}$$

and is noisily measured by $\mathbf{x}_i \in \mathbb{R}^{m \times \dim(\mathbf{x})}$

$$\mathbf{x}_i | \boldsymbol{\theta}_i \overset{i.i.d.}{\sim} \mu_{\boldsymbol{\theta}}(\mathbf{x}; \boldsymbol{\gamma}), \tag{18}$$

where $\mu_{\boldsymbol{\theta}}(\mathbf{x}; \boldsymbol{\gamma})$ is the conditional distribution of $\mathbf{x}$ given $\boldsymbol{\theta}$. The parameter $\boldsymbol{\gamma} \in \mathbb{R}^{\dim(\boldsymbol{\gamma})}$ can be seen as a nuissance parameter that adds flexibility to model. For example, in the teacher quality context, the parameter $\boldsymbol{\gamma}$ may include the variance of within-class test score or the effect of control variables on the test score.[7] Note that, while we allow the prior π to be nonparametrically specified, we assume that the likelihood μ_θ belongs to a parametric family of distributions indexed by $\boldsymbol{\gamma}$.

tion, it is not hard to see that the equality would hold if (x_{ij}, θ_i) follows normal-normal.

[7] Consider the example $x_{ij}|\theta_i, z_{ij} \sim N(\theta_i + z_{ij}\gamma_1, \gamma_2)$, where x_{ij} is the test score and z_{ij} is a control variable such as family income. Define $\tilde{x}_{ij} = x_{ij} - z_{ij}\gamma_1$. We can write $\tilde{x}_{ij}|\theta_i \sim N(\theta_i, \gamma_2)$, and $\gamma = (\gamma_1, \gamma_2)$ is the nuisance parameter.

We consider the regression of the following form:

$$y_{ij} = g(\boldsymbol{\theta}_i, \boldsymbol{z}_{ij}) + \epsilon_{ij} \tag{19}$$

$$= \sum_{k=1}^{K} g_k(\boldsymbol{\theta}_i; \boldsymbol{\beta}) + \sum_{l=1}^{L} h_l(\boldsymbol{z}_{ij}; \boldsymbol{\delta}) + \epsilon_{ij}, \tag{20}$$

where $g(\cdot)$ is the conditional expectation function, $g_k(\cdot)$ and $h_l(\cdot)$ are functions of $\boldsymbol{\theta}_i$ and $\boldsymbol{z}_{ij}$ that comprise $g(\cdot)$, and $(\boldsymbol{\beta}_o, \boldsymbol{\delta}) \in \mathbb{R}^{\dim(\boldsymbol{\beta})} \times \mathbb{R}^{\dim(\boldsymbol{\delta})}$ is the vector of unknown parameters. We assume that the functional form of the functions $\{g_k(\cdot)\}_{k=1}^{K}$, and $\{h_l(\cdot)\}_{l=1}^{L}$ are known.

Central to our estimation strategy is to project the regression model onto the variable $\mathbf{x}_i$ so that we can identify the regression coefficients from a model that only consists of the observed variables. Specifically, in absence of functions $\{h_l(\cdot)\}_{l=1}^{L}$, our method amounts to estimating the parameter $\boldsymbol{\beta}$ by

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(y_{ij} - \sum_{k=1}^{K} \mathbb{E}(g_k(\boldsymbol{\theta}_i; \boldsymbol{\beta}) | \mathbf{x}_i) \right)^2, \tag{21}$$

provided that the posterior distribution of $\boldsymbol{\theta}_i$ given $\mathbf{x}_i$ is known.

Note that the estimator $\tilde{\boldsymbol{\beta}}$ is infeasible because it requires specifying a prior on the distribution of $\boldsymbol{\theta}$ for its implementation. In line with the empirical Bayes approach, we estimate the prior from data rather than specifying it. Denote this empirically derived prior as $\hat{\pi}$. Then, using the estimated prior $\hat{\pi}$, we can calculate the corresponding posterior and plug it into the regression model. Suppose the posterior admits a density $\hat{p}(\cdot|\mathbf{x})$, we can estimate (20) by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ y_{ij} - \sum_{k=1}^{K} \int g_k(\boldsymbol{\theta}; \boldsymbol{\beta}) \hat{p}(\boldsymbol{\theta}|\mathbf{x}_i) \mathrm{d}\boldsymbol{\theta} \right\}^2$$

which is the feasible counterpart of $\tilde{\boldsymbol{\beta}}$. In our proof, we will initially discuss the theoretical properties of $\tilde{\boldsymbol{\beta}}$ and subsequently extend the result to $\hat{\boldsymbol{\beta}}$.

When $\{h_l(\cdot)\}_{l=1}^L$ are present in the regression model, we have to first subtract $\sum_{l=1}^L h_l(\boldsymbol{z}_{ij};\boldsymbol{\delta})$ from y_{ij} before applying the procedure described above. In our setup, this can be done easily since we can view

$$\sum_{k=1}^{K} g_k(\boldsymbol{\theta}_i;\boldsymbol{\beta})$$

as fixed effects $\tilde{\theta}_i = \sum_{k=1}^K g_k(\boldsymbol{\theta}_i;\boldsymbol{\beta})$, and we can estimate $\boldsymbol{\delta}$ by a fixed effect regression

$$y_{ij} = \sum_{l=1}^{L} h_l(\boldsymbol{z}_{ij};\boldsymbol{\delta}) + \tilde{\theta}_i + \epsilon_{ij}.$$

Subtracting $\sum_{l=1}^L h_l(\boldsymbol{z}_{ij};\hat{\boldsymbol{\delta}})$ from y_{ij}, we can then proceed to the estimation of $\boldsymbol{\beta}$.

Below, we outline the steps to implement our method.

- Step 1: estimate the nuisance parameter $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. In most setup, this can be easily done by either method of moments or a fixed effect regression. Denote the estimators as $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$.

- Step 2: estimate the prior $\pi(\cdot)$ by the nonparametric maximum likelihood estimator (NPMLE):

  $$\hat{\pi} = \arg\max_{\pi} \frac{1}{n}\sum_i ln\left[\int \mu_\theta(\mathbf{x}_i;\hat{\boldsymbol{\gamma}})\mathrm{d}\pi(\theta)\right]$$

  that maximizes the likelihood over all possible distribution.

- Step 3: given the estimated prior $\hat{\pi}(\theta)$ and the likelihood $\mu_\theta(\mathbf{x};\hat{\boldsymbol{\gamma}})$, use the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{x}$ to calculate the posterior mean of the conditional mean functions

  $$\hat{\mathbb{E}}[g_k(\boldsymbol{\theta};\boldsymbol{\beta})|\mathbf{x}],\ k = 1, 2, ..., K.$$

  Finally, estimate β with the regression

  $$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{y_{ij} - \sum_{l=1}^{L} h_l(\boldsymbol{z}_{ij};\hat{\boldsymbol{\delta}}) - \sum_{k=1}^{K}\hat{\mathbb{E}}[g_k(\boldsymbol{\theta};\boldsymbol{\beta})|\mathbf{x}_i]\mathrm{d}\boldsymbol{\theta}\right\}^2.$$

Remark 1. *Equation (18) is a different framework from classical measurement error, which assumes the error is an additive independent random variable. Suppose* $\theta_i \sim \text{Beta}(a, b)$ *and* $x_i|\theta_i \sim \text{Bin}(2, \theta_i)$*. Then, the error*

$$e_i = 0.5x_i - \theta_i$$

is ***mean-independent*** *of* θ_i*, but* ***not independet*** *from* θ_i *(*$\mathbb{E}[e_i|\theta_i] = 0$ *but* $\mathbb{P}(e_i > 0|\theta_i = 0.5) \neq \mathbb{P}(e_i > 0|\theta_i = 1) = 0$*.) Thus, deconvolution, which relies on the decomposition of characteristic functions, is not useful.*

4 Theoretical Results

In this section, we derive the statistical properties of the proposed estimator. The structure of our argument will closely follow the three steps that we lay down earlier in the last section. We will show that, if consistent estimators $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$ for the nuisance parameters are available, then $\hat{\pi}$, the prior estimated by NPMLE, also converges to the true distribution when the sample size goes to infinity. The convergence of the posterior regression function $\hat{\mathbb{E}}[g_k(\boldsymbol{\theta}; \boldsymbol{\beta})|\mathbf{x}]$ can then be established, which further implies the regression estimator $\hat{\beta}$ is consistent and asymptotically normal.

4.1 Consistency of NPMLE $\hat{\pi}$

The idea of NPMLE can be traced as early back to the seminal paper of Kiefer and Wolfowitz (1956). Since then, a variety of theoretical extension of the baseline method as well as progresses in the computation of NPMLE has been done. In this paper, we apply and generalize NPMLE to mixture models with nuisance parameters. Noteworthily, we derive an equivalent formulation of NPMLE that could be of independent interest. Our formulation allows the utilization of the recent advancement in the optimal transport problem, which makes solving NPMLE in mixture models with multi-dimensional latent variables possible.

Given the distribution $\pi(\boldsymbol{\theta})$, we can write the marginal likelihood of $\mathbf{x}$ as

$$f(\mathbf{x};\pi,\boldsymbol{\gamma}) = \int_{\Theta} \mu_{\theta}(\mathbf{x};\boldsymbol{\gamma})\mathrm{d}\pi(\boldsymbol{\theta}),$$

that is, we can view the distribution of $\mathbf{x}$ as a mixture with mixing distribution $\pi(\boldsymbol{\theta})$. The idea of NPMLE is to search for the mixing distribution $\pi(\cdot)$ over Π, the space of all possible distribution on the support of $\boldsymbol{\theta}$. Formally, the NPMLE estimator $\hat{\pi}$ is given by:

$$\hat{\pi} = \max_{\pi\in\Pi} \sum_{i=1}^{n} \ln f(\mathbf{x}_i;\pi,\hat{\boldsymbol{\gamma}}).$$

To discuss the consistency of $\hat{\pi}$, we equip Π with the metric D_{KW}:

$$D_{KW}(\pi_1,\pi_2) = \int_{\Theta} |\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})| \ \exp(-||\boldsymbol{\theta}||_1)\mathrm{d}\boldsymbol{\theta}.$$

Notice that this metric is bounded on the space of c.d.f, and convergence in D_{KW} is equivalent to convergence in distribution, i.e., $D_{KW}(\pi_n,\pi) \to 0$ if and only if

$$\pi_n(\boldsymbol{\theta}) \to \pi(\boldsymbol{\theta})$$

for all continuity point $\boldsymbol{\theta} \in \Theta$.

The following assumptions guarantees that $\hat{\pi}$ is strongly consistent. Assumption 1 and 2 are standard in the literature except that we include nuisance parameter $\boldsymbol{\gamma}$ in the statement. Assumptions 3 and 4 are needed to assure that estimation error in $\hat{\boldsymbol{\gamma}}$ does not affect the consistency of $\hat{\pi}$ when the sample size n is large enough.

Assumption 1 (Identification)**.** *Let $F(\mathbf{x};\pi,\boldsymbol{\gamma})$ be the cumulative distribution function of $f(\mathbf{x};\pi,\boldsymbol{\gamma})$. For all $\boldsymbol{\gamma}$, if $F(\mathbf{x};\pi,\boldsymbol{\gamma}) = F(\mathbf{x};\pi',\boldsymbol{\gamma})$ for all $\mathbf{x}$, then $D_{KW}(\pi,\pi') = 0$.*

Assumption 2 (Contintuity)**.** *For all $\boldsymbol{\gamma}$, the function $f(\mathbf{x};\pi)$ is continuous in $\pi \in \Pi$.*

Assumption 3. *$\hat{\gamma}$ is a consistent estimator of $\boldsymbol{\gamma}$.*

Assumption 4. *Let* $Q(\pi, \boldsymbol{\gamma}) = \mathbb{E}_{\mathbf{x}}[\int_{\Theta} \mu_\theta(\mathbf{x}; \boldsymbol{\gamma}) \mathrm{d}\pi(\boldsymbol{\theta})]$. *There exists a neighborhood* $N_{\boldsymbol{\gamma}_o}$ *of* $\boldsymbol{\gamma}_o$ *s.t.* $\sup_{\boldsymbol{\gamma} \in N_{\boldsymbol{\gamma}_o}} \sup_{\pi \in \Pi} |\frac{\partial Q(\pi, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}| \leqslant M$ *for some* $M > 0$.

Proposition 2. *Under Assumption 1 - 4, when* $n \to \infty$,

$$D_{KW}(\hat{\pi}, \pi_o) \to 0 \quad a.s.,$$

i.e., $\hat{\pi}(\boldsymbol{\theta}) \xrightarrow{a.s.} \pi_o(\boldsymbol{\theta}) \ \forall \ \theta \in \Theta$.

Proof. The main part of the proof is to show that the sample criterion function

$$Q_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_i ln \left[\int \mu_\theta(\mathbf{x}_i; \hat{\boldsymbol{\gamma}}) \mathrm{d}\pi(\theta) \right]$$

converges uniformly to the population criterion function

$$Q(\boldsymbol{\beta}) = \mathbb{E} \left[\int \mu_\theta(\mathbf{x}_i; \boldsymbol{\gamma}_o) \mathrm{d}\pi(\theta) \right].$$

Consider the following functions:

$$\tilde{Q}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{n} \sum_i ln \left[\int \mu_\theta(\mathbf{x}_i; \boldsymbol{\gamma}) \mathrm{d}\pi(\theta) \right]$$

and

$$\tilde{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbb{E} \left[\int \mu_\theta(\mathbf{x}_i; \boldsymbol{\gamma}) \mathrm{d}\pi(\theta) \right].$$

Then

$$\begin{aligned} |Q_n(\boldsymbol{\beta}) - Q(\boldsymbol{\beta})| &= |\tilde{Q}_n(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}) - \tilde{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma}_o)| \\ &\leqslant |\tilde{Q}_n(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}) - \tilde{Q}(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}})| + |\tilde{Q}(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}) - \tilde{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma}_o)|. \end{aligned}$$

Note that the first term converges uniformly over $\boldsymbol{\beta}$ by as the standard NPMLE Chen (2017).

Using Taylor expansion, we can see that the second term is uniformly bounded by

$$M \cdot \|(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_o)\|,$$

which converges to zero as $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}_o$ □

We propose an algorithm to approximate the NPMLE solution when the latent variable $\boldsymbol{\theta} \in \mathbb{R}^d$ is a multi-dimensional vector. Suppose the likelihood of the observation vector $\mathbf{x}_i$ given the latent $\boldsymbol{\theta}_i$ has a density (or probability mass function) $\varphi_n(\mathbf{x}_i; \boldsymbol{\theta}_i)$.[8] The NPMLE estimator for the distribution of $\boldsymbol{\theta}_i$ is defined as

$$\hat{\pi} = \arg\max_{\pi} n^{-1} \sum_{i=1}^{n} \log \int \varphi_n(\mathbf{x}_i; \boldsymbol{\theta}) d\pi(\boldsymbol{\theta}),$$

where the maximum is taken over the space of all probability distributions on $\mathbb{R}^d$. It can be shown that, after some assumptions, $\hat{\pi}$ can be identified as

$$\hat{\pi} = \arg\min_{\pi} \inf_{\gamma \in \Pi(\pi, U)} \left[\int -\log \varphi_n(\mathbf{x}; \boldsymbol{\theta}) d\gamma(\boldsymbol{\theta}, \mathbf{x}) + I(\gamma) \right], \tag{22}$$

where $\Pi(\mu, \nu)$ denotes the set of couplings between probabilities μ and ν, U is the uniform on $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, and $I(\cdot)$ is the mutual information. In other words, $\hat{\pi}$ is the minimizer of the regularized optimal transport cost between π and U, with the cost measured by minus log likelihood.

With the identification (22), we can employ a gradient descent method to approximate $\hat{\pi}$. Let us pre-fix a grid points $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{k_n}\}$ in $\Theta \subset \mathbb{R}^d$, and initialize the probabilities $\{p_1, \ldots, p_{k_n}\}$ associated with the grid points. Consider the following regularized optimal

[8] We ignore $\boldsymbol{\gamma}$ as it does not affect the formulation.

transport problem, which is a smooth constrained optimization problem.

$$\begin{aligned} \min_{\gamma} \sum_{i=1}^{n} \sum_{j=1}^{k_n} -\gamma_{ij} \log \varphi_n(\mathbf{x}_i; \boldsymbol{\theta}_j) + I(\gamma) \\ \text{s.t.} \sum_i \gamma_{ij} = p_j, \quad j = 1, 2, \ldots, k_n \\ \sum_j \gamma_{ij} = \frac{1}{n}, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{23}$$

Then the gradient of the (minimized) transport cost with respect to $\mathbf{p} = (p_1, \ldots, p_{k_n})^\top$ is exactly the vector of Lagrange multipliers, $\boldsymbol{\lambda}$, associated with the set of constraints (23). Note that we can, without loss of generality, normalize $\boldsymbol{\lambda}$ so that $\sum_j^{k_n} \lambda_j = 0$. Therefore, the gradient update $\mathbf{p} + \eta\boldsymbol{\lambda}$ remains a valid probability on the grid points for any learning rate $\eta \in (0, 1)$. The following algorithm summarizes the above discussions.

Algorithm 1: Multi-dimensional NPMLE

Initialization: Probabilities $\mathbf{p}^{(0)} = (p_1, \ldots, p_{k_n})$, maximum number of iteration M, learning rate η

1 **for** $m = 1, 2, \ldots, M$ **do**
2 Compute the Lagrange multiplier vector $\boldsymbol{\lambda}$ associated with (23) for $\mathbf{p} = \mathbf{p}^{(m-1)}$.
3 If necessary, normalize $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}$ sums up to zero.
4 Update $\mathbf{p}^{(m)} = \mathbf{p}^{(m-1)} + \eta\boldsymbol{\lambda}$.
5 Early stopping if convergence.
6 **end**

4.2 Asymptotics of the infeasible estimator $\tilde{\beta}$

We begin our analysis of the proposed estimator $\hat{\boldsymbol{\beta}}$ by inspecting the properties of its oracle counterpart $\tilde{\boldsymbol{\beta}}$

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ y_{ij} - \sum_{l=1}^{L} h_l(\boldsymbol{z}_{ij}; \boldsymbol{\delta}) - \sum_{k=1}^{K} \mathbb{E}[g_k(\boldsymbol{\theta}; \boldsymbol{\beta}) | \mathbf{x}_i] \mathrm{d}\boldsymbol{\theta} \right\}^2, \tag{24}$$

in which the nuisance parameters $\pi(\cdot)$ and $\boldsymbol{\gamma}$ in $\mathbb{E}[g_k(\boldsymbol{\theta};\boldsymbol{\beta})|\mathbf{x}_i]$ and $\boldsymbol{\delta}$ are assumed to be known. Then, we will show that the $\hat{\boldsymbol{\beta}}$ is asymptotically equivalent to $\tilde{\boldsymbol{\beta}}$. Throughout the paper, we assume the parameter $\boldsymbol{\beta}$ belongs to a compact convex parameter space $\mathcal{B} \subseteq \mathbb{R}^d$. Moreover, the regression function $h(\cdot)$ is Lipschitz in $\boldsymbol{\beta}$ uniformly over Θ. That is, there exists $L_h < \infty$ such that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |h(\boldsymbol{\theta};\boldsymbol{\beta}) - h(\boldsymbol{\theta};\boldsymbol{\beta}')| \leqslant L_h \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|$$

for all $\boldsymbol{\beta}$, $\boldsymbol{\beta}'$ in $\mathcal{B}$.

Proposition 3. *Assume*

$$\mathbb{E}\left\{\mathbb{E}[h(\boldsymbol{\theta}_i;\boldsymbol{\beta}_o) - h(\boldsymbol{\theta}_i;\boldsymbol{\beta})|\mathbf{x}_i]\right\}^2 = 0 \tag{25}$$

if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_o$. In addition, assume

$$\mathbb{E}\sup_{\boldsymbol{\beta}\in\mathcal{B}}(h(\boldsymbol{\theta}_i;\boldsymbol{\beta}))^2 < \infty$$

and the moments $\mathbb{E}\,\epsilon_{ij}^2$ exists. Then $\tilde{\boldsymbol{\beta}}$ is consistent.

Proof. W.L.O.G., we can assume $\sum_{l=1}^{L} h_l(\boldsymbol{z}_{ij};\boldsymbol{\delta}_o) = 0$ as we can redefine the outcome as $\tilde{y}_{ij} = y_{ij} - \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} - \sum_{l=1}^{L} h_l(\boldsymbol{z}_{ij};\boldsymbol{\delta}_o)$ in the following proof. Write $Q_n(\boldsymbol{\beta}) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \mathbb{E}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta}_i;\boldsymbol{\beta})|\mathbf{x}_i))^2$. We first show that $Q_n(\boldsymbol{\beta})$ converges to a non-stochastic $Q(\cdot)$ in probability uniformly, which is uniquely minimized at $\boldsymbol{\beta}_o$. Write

$$\begin{aligned} Q_n(\boldsymbol{\beta}) =& \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\epsilon_{ij}^2 + n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}(h(\boldsymbol{\theta}_i;\boldsymbol{\beta}_o) - \mathbb{E}(h(\boldsymbol{\theta}_i;\boldsymbol{\beta})|\mathbf{x}_i))^2 \\ &+ \frac{2}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_i(h(\boldsymbol{\theta}_i;\boldsymbol{\beta}_o) - \mathbb{E}(h(\boldsymbol{\theta}_i;\boldsymbol{\beta})|\mathbf{x}_i)) \\ :=& A_{1,n} + A_{2,n} + A_{3,n}. \end{aligned}$$

Clearly $A_{1,n} \to \mathbb{E}(\epsilon_{ij}^2)$ in probability. Since $\mathcal{B}$ is compact, for each $\epsilon > 0$ we can choose a finite collection $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k\}$ in $\mathcal{B}$ such that for each $\boldsymbol{\beta} \in \mathcal{B}$ there exists some $\boldsymbol{\beta}_i$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_i\| < \epsilon$. Note that

$$\begin{aligned} A_{2,n} =& n^{-1} \sum_{i=1}^{n} (h(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o))^2 - \frac{2}{n} \sum_{i=1}^{n} h(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o) \mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}) | \mathbf{x}_i) + n^{-1} \sum_{i=1}^{n} (\mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}) | \mathbf{x}_i))^2 \\ :=& B_{1,n} + B_{2,n} + B_{3,n}. \end{aligned}$$

Clearly $B_{1,n} \to \mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o))^2$ in probability. To each $\boldsymbol{\beta}$ choose $\boldsymbol{\beta}^\dagger \in \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k\}$ such that $\|\boldsymbol{\beta}^\dagger - \boldsymbol{\beta}_i\| < \epsilon$. We have

$$B_{2,n} = -\frac{2}{n} \sum_{i=1}^{n} h(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o) \mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}^\dagger) | \mathbf{x}_i) - \frac{2}{n} \sum_{i=1}^{n} h(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o) \mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}) - h(\boldsymbol{\theta}_i; \boldsymbol{\beta}^\dagger) | \mathbf{x}_i),$$

so

$$\left| B_{2,n} + \frac{2}{n} \sum_{i=1}^{n} h(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o) \mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}^\dagger) | \mathbf{x}_i) \right| \leqslant \frac{2 L_h \epsilon}{n} \sum_{i=1}^{n} h(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o),$$

implying

$$\limsup_{n \to \infty} |B_{2,n} + 2\mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o) \mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}^\dagger | \mathbf{x}_i)| \leqslant 2 L_h \epsilon \mathbb{E} \sup_{\boldsymbol{\beta} \in \mathcal{B}} |h(\boldsymbol{\theta}_i; \boldsymbol{\beta})|$$

with probability tending to one. Since ϵ is arbitrary, standard arguments yield

$$B_{2,n} \to -2\mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o) \mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}) | \mathbf{x}_i)$$

in probability uniformly in $\boldsymbol{\beta} \in \mathcal{B}$. By a similar argument,

$$B_{3,n} \to \mathbb{E}(\mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}) | \mathbf{x}_i))^2$$

in probability uniformly in $\boldsymbol{\beta}$. Thus $A_{2,n} \to \mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}_o) - \mathbb{E}(h(\boldsymbol{\theta}_i; \boldsymbol{\beta}) | \mathbf{x}_i))^2$ in probability

uniformly in $\boldsymbol{\beta}$. Similarly, $A_{3,n} \to 0$ in probability uniformly in $\boldsymbol{\beta}$. We have shown

$$\begin{aligned} Q_n(\boldsymbol{\beta}) \to & \mathbb{E}(\epsilon_i^2) + \mathbb{E}(h(\boldsymbol{\theta}_i;\boldsymbol{\beta}_o) - \mathbb{E}(h(\boldsymbol{\theta}_i;\boldsymbol{\beta})|\mathbf{x}_i))^2 \\ = & \mathbb{E}(\epsilon_i^2) + \mathbb{E}(h(\boldsymbol{\theta}_i;\boldsymbol{\beta}_o) - \mathbb{E}(h(\boldsymbol{\theta}_i;\boldsymbol{\beta}_o)|\mathbf{x}_i))^2 + \mathbb{E}(\mathbb{E}(h(\boldsymbol{\theta}_i;\boldsymbol{\beta}_o) - h(\boldsymbol{\theta}_i;\boldsymbol{\beta})|\mathbf{x}_i))^2, \end{aligned}$$

which is uniquely minimized at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$ by our assumption. The desired result follows from standard results, e.g. Theorem 4.2.1 of Amemiya (1985). □

Remark 2. *The identification condition (25) is commonly adopted in the measurement error literature. In the linear regression model where $h(\boldsymbol{\theta};\boldsymbol{\beta}) = \boldsymbol{\theta}^\top\boldsymbol{\beta}$, (25) is equivalent to*

$$\mathbb{E}\left\{\mathbb{E}(\boldsymbol{\theta}_i|\mathbf{x}_i)\mathbb{E}(\boldsymbol{\theta}_i|\mathbf{x}_i)^\top\right\}$$

being positive definite.

Proposition 4. *Let $H_i(\boldsymbol{\beta}) = \mathbb{E}_\theta(h(\boldsymbol{\theta};\boldsymbol{\beta})|\mathbf{x}_i)$. Assuming regularity conditions for WLLN and CLT to hold. Assume also $H_i(\boldsymbol{\beta}) = \mathbb{E}(h(\boldsymbol{\theta};\boldsymbol{\beta})|\mathbf{x}_i)$ is twice continuously differentiable in $\boldsymbol{\beta}$, and the following matrices exist:*

$$\begin{aligned} \mathbf{M} = & \mathbb{E}\left[\nabla_{\boldsymbol{\beta}} H_i(\boldsymbol{\beta}_o)\nabla_{\boldsymbol{\beta}}^\top H_i(\boldsymbol{\beta}_o)\right], \\ \mathbf{V} = & \mathbb{E}\left[(y_i - H_i(\boldsymbol{\beta}_o))^2\nabla_{\boldsymbol{\beta}} H_i(\boldsymbol{\beta}_o)\nabla_{\boldsymbol{\beta}}^\top H_i(\boldsymbol{\beta}_o)\right]. \end{aligned}$$

Then

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \Rightarrow N(\mathbf{0}, \mathbf{M}^{-1}\mathbf{V}\mathbf{M}^{-1}). \tag{26}$$

Proof. The result follows from the proof of Proposition 3 and Delta method. □

Remark 3. *Proposition 3 and 4 can be applied to situations when researchers uses Bayesian (rather than empirical Bayes) procedure to generate regressors. For example, researchers may use Bayesian topic models to generate the topic of article for a regression analysis.*

Example 1. Consider the linear model $h(\boldsymbol{\theta};\boldsymbol{\beta}) = \boldsymbol{\theta}^\top\boldsymbol{\beta}$. Let $\mathbf{z}_i = \mathbb{E}(\boldsymbol{\theta}|\mathbf{x}_i)$. Then

$$\mathbf{M} = \mathbb{E}(\mathbf{z}_i\mathbf{z}_i^\top), \mathbf{V} = \mathbb{E}[(y_i - \mathbf{z}_i^\top\boldsymbol{\beta}_o)^2\mathbf{z}_i\mathbf{z}_i^\top],$$

and $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \Rightarrow N(0, \mathbf{M}^{-1}\mathbf{V}\mathbf{M}^{-1})$. When $\boldsymbol{\beta}_o = 0$, $\mathbf{M}^{-1}\mathbf{V}\mathbf{M}^{-1} = \sigma_\epsilon^2\mathbf{M}^{-1}$. In this case, naive EB coincides with this formula. Such equivalence when the true regression coefficient is zero is also seen in the instrumental variable regression.

Example 2. *For the nonlinear model $h(\theta;\beta) = \theta^2\beta$, we have $M = \mathbb{E}(w_i^2)$ and $V = \mathbb{E}[(y_i - w_i\beta_o)^2 w_i^2]$, where $w_i = \mathbb{E}(\theta^2|\mathbf{x}_i)$. In this case, ignoring the generated regressor can lead to biased standard error even when $\beta_o = 0$.*

4.3 Asymptotic results for the feasible estimator

In this subsection, we assume the likelihood measure $\mu_{\boldsymbol{\theta}}(\cdot)$ has a density $f_{\boldsymbol{\theta}}(\cdot)$ with respect to some common dominating measure λ (i.e. $\mu_{\boldsymbol{\theta}} \ll \lambda$ for each $\boldsymbol{\theta} \in \Theta$, where $\Theta \subseteq \mathbb{R}^p$). Let

$$\mathcal{G} = \left\{ \boldsymbol{\theta} \mapsto \frac{h(\boldsymbol{\theta};\boldsymbol{\beta}) f_{\boldsymbol{\theta}}(x)}{\int f_{\boldsymbol{\theta}}(x) d\pi(\boldsymbol{\theta})} : x \in \mathbb{R}^q, \boldsymbol{\beta} \in \mathcal{B}, \pi \in \mathcal{P} \right\}$$

be the collection of conditional expectation kernels for the conditional mean regression function, where $\mathcal{B} \subseteq \mathbb{R}^d$ is our parameter space. Observe that

$$\hat{\mathbb{E}}_{\boldsymbol{\theta}}\left[h(\boldsymbol{\theta};\boldsymbol{\beta})|\mathbf{x}_i = x\right] = \int \frac{h(\boldsymbol{\theta};\boldsymbol{\beta}) f_{\boldsymbol{\theta}}(x)}{\int f_{\boldsymbol{\theta}}(x) d\pi_n(\boldsymbol{\theta})} d\pi_n(\boldsymbol{\theta}),$$

and

$$\mathbb{E}_{\boldsymbol{\theta}}\left[h(\boldsymbol{\theta};\boldsymbol{\beta})|\mathbf{x}_i = x\right] = \int \frac{h(\boldsymbol{\theta};\boldsymbol{\beta}) f_{\boldsymbol{\theta}}(x)}{\int f_{\boldsymbol{\theta}}(x) d\pi_*(\boldsymbol{\theta})} d\pi_*(\boldsymbol{\theta}).$$

The following lemma gives a uniform convergence result of the empirical-Bayes posterior mean regression function to the infeasible Bayes posterior mean regression function.

Lemma 2. *Consider a sequence of prior distributions $\pi_n \in \mathcal{P}$ such that $\pi_n \Rightarrow \pi_*$. Assume that*

(1) there exists some $M < \infty$ such that

$$\sup_{x \in \mathbb{R}^q, \boldsymbol{\beta} \in \mathcal{B}, \pi \in \mathcal{P}, \boldsymbol{\theta} \in \Theta} \left| \frac{h(\boldsymbol{\theta}; \boldsymbol{\beta}) f_{\boldsymbol{\theta}}(x)}{\int f_{\boldsymbol{\theta}}(x) d\pi(\boldsymbol{\theta})} \right| \leqslant M;$$

(2) each element in $\mathcal{G}$ is continuous in $\boldsymbol{\theta}$;

(3) for each $\epsilon > 0$ there exist a compact $K_\epsilon \subseteq \Theta$ and a finite *subset $\mathcal{C}_\epsilon \subseteq \mathcal{G}$ such that to each $x \in \mathbb{R}^q$ and $\boldsymbol{\beta} \in \mathcal{B}$ corresponds an $g \in \mathcal{C}_\epsilon$ with*

$$\sup_{\boldsymbol{\theta} \in K_\epsilon, \pi \in \mathcal{P}} \left| \frac{h(\boldsymbol{\theta}; \boldsymbol{\beta}) f_{\boldsymbol{\theta}}(x)}{\int f_{\boldsymbol{\theta}}(x) d\pi(\boldsymbol{\theta})} - g(\boldsymbol{\theta}) \right| < \epsilon,$$

and $\pi_*(K_\epsilon) \geqslant 1 - \epsilon$ and $\pi_n(K_\epsilon) \geqslant 1 - \epsilon$ for all n.

Then,

$$\sup_{x \in \mathbb{R}^q, \boldsymbol{\beta} \in \mathcal{B}} \left| \hat{\mathbb{E}}_{\boldsymbol{\theta}} \left[h(\boldsymbol{\theta}; \boldsymbol{\beta}) | \mathbf{x}_i = x \right] - \mathbb{E}_{\boldsymbol{\theta}} \left[h(\boldsymbol{\theta}; \boldsymbol{\beta}) | \mathbf{x}_i = x \right] \right| \to 0.$$

Proof. Fixing x and $\boldsymbol{\beta}$, we can write

$$\left| \hat{\mathbb{E}}_{\boldsymbol{\theta}} \left[h(\boldsymbol{\theta}; \boldsymbol{\beta}) | \mathbf{x}_i = x \right] - \mathbb{E}_{\boldsymbol{\theta}} \left[h(\boldsymbol{\theta}; \boldsymbol{\beta}) | \mathbf{x}_i = x \right] \right| \leqslant R_1 + R_2 + R_3 + R_4,$$

where

$$
\begin{aligned}
R_1 &= \left| \int_{K_\epsilon^c} \frac{h(\boldsymbol{\theta};\boldsymbol{\beta}) f_{\boldsymbol{\theta}}(x)}{\int f_{\boldsymbol{\theta}}(x) d\pi_n(\boldsymbol{\theta})} d\pi_n(\boldsymbol{\theta}) - \int_{K_\epsilon^c} \frac{h(\boldsymbol{\theta};\boldsymbol{\beta}) f_{\boldsymbol{\theta}}(x)}{\int f_{\boldsymbol{\theta}}(x) d\pi_*(\boldsymbol{\theta})} d\pi_*(\boldsymbol{\theta}) \right| \\
R_2 &= \left| \int_{K_\epsilon} \frac{h(\boldsymbol{\theta};\boldsymbol{\beta}) f_{\boldsymbol{\theta}}(x)}{\int f_{\boldsymbol{\theta}}(x) d\pi_n(\boldsymbol{\theta})} d\pi_n(\boldsymbol{\theta}) - \int_{K_\epsilon} g(\boldsymbol{\theta}) d\pi_n(\boldsymbol{\theta}) \right| \\
R_3 &= \left| \int_{K_\epsilon} \frac{h(\boldsymbol{\theta};\boldsymbol{\beta}) f_{\boldsymbol{\theta}}(x)}{\int f_{\boldsymbol{\theta}}(x) d\pi_*(\boldsymbol{\theta})} d\pi_*(\boldsymbol{\theta}) - \int_{K_\epsilon} g(\boldsymbol{\theta}) d\pi_*(\boldsymbol{\theta}) \right| \\
R_4 &= \sup_{g \in \mathcal{C}_\epsilon} \left| \int_{K_\epsilon} g(\boldsymbol{\theta}) d\pi_n(\boldsymbol{\theta}) - \int_{K_\epsilon} g(\boldsymbol{\theta}) d\pi_*(\boldsymbol{\theta}) \right|.
\end{aligned}
$$

By condition (1) and (3), $R_1 \leqslant 2M\epsilon$. By condition (3), $R_2 + R_3 \leqslant 2\epsilon$. Since $g \in \mathcal{C}_\epsilon \subseteq \mathcal{G}$, it is bounded and continuous by conditions (1) and (2). Thus it follows from $\pi_n \Rightarrow \pi_*$, $R_4 \leqslant 2\epsilon + o(1)$. This concludes the proof. □

4.4 Equivalence to Bayes with oracle prior

Let

$$
L_n(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} (y_i - \mathbb{E}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta};\boldsymbol{\beta})|\mathbf{x}_i))^2
$$

and

$$
\hat{L}_n(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{\mathbb{E}}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta};\boldsymbol{\beta})|\mathbf{x}_i))^2.
$$

Recall that $\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} L_n(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} \hat{L}_n(\boldsymbol{\beta})$.

Lemma 3. *Assume (1) for each $\delta > 0$ there exists an $\epsilon > 0$ such that*

$$
P\left(\min_{\boldsymbol{\beta} \notin B_\delta(\tilde{\boldsymbol{\beta}})} L_n(\boldsymbol{\beta}) - L_n(\tilde{\boldsymbol{\beta}}) > \epsilon \right) \to 1;
$$

(2) $\mathcal{B}$ is compact, and (3)

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} |\hat{L}_n(\boldsymbol{\beta}) - L_n(\boldsymbol{\beta})| \to 0$$

in probability. Then,

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\| \to 0$$

in probability.

Proof. Fix $\delta > 0$ and let $\epsilon > 0$ satisfy condition (1). Then

$$\begin{aligned}
1 =& P(\hat{L}_n(\hat{\boldsymbol{\beta}}) \leqslant \hat{L}_n(\tilde{\boldsymbol{\beta}})) \\
\leqslant& P\left(L_n(\hat{\boldsymbol{\beta}}) \leqslant L_n(\tilde{\boldsymbol{\beta}}) + 2\sup_{\boldsymbol{\beta}\in\mathcal{B}} |\hat{L}_n(\boldsymbol{\beta}) - L_n(\boldsymbol{\beta})|\right) \\
\leqslant& P(L_n(\hat{\boldsymbol{\beta}}) \leqslant L_n(\tilde{\boldsymbol{\beta}}) + \epsilon) + P\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}} |\hat{L}_n(\boldsymbol{\beta}) - L_n(\boldsymbol{\beta})| > \epsilon/2\right) \\
\leqslant& P(\hat{\boldsymbol{\beta}} \in B_\delta(\tilde{\boldsymbol{\beta}})) + o(1).
\end{aligned}$$

The desired result follows. □

In Proposition 1, we have shown that $\tilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_o$. Therefore, a direct consequence of Lemma 3 is that $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_o$ in probability.

Proposition 5. *Assume WLLN and CLT. If $\hat{\boldsymbol{\beta}}$ is the feasible EB estimator computed using prior π_n with $\pi_n \Rightarrow \pi_*$, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \Rightarrow N(0, \mathbf{M}^{-1}\mathbf{V}\mathbf{M}^{-1}),$$

where $\mathbf{M}$ and $\mathbf{V}$ are the same as those in (26).

Proof. The proof is standard delta method. By F.O.C. and mean value theorem,

$$
\begin{aligned}
0 =& -\frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{\mathbb{E}}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta};\hat{\boldsymbol{\beta}}|\mathbf{x}_i))\nabla_{\boldsymbol{\beta}}\hat{\mathbb{E}}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta};\hat{\boldsymbol{\beta}})|\mathbf{x}_i) \\
=& \frac{2}{n}\left[\sum_{i=1}^{n}\nabla_{\boldsymbol{\beta}}\hat{\mathbb{E}}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta};\check{\boldsymbol{\beta}}|\mathbf{x}_i))\nabla_{\boldsymbol{\beta}}^{\top}\hat{\mathbb{E}}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta};\check{\boldsymbol{\beta}}|\mathbf{x}_i)) - \sum_{i=1}^{n}(y_i - \hat{\mathbb{E}}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta};\check{\boldsymbol{\beta}}|\mathbf{x}_i))\nabla_{\boldsymbol{\beta}}^{2}\hat{\mathbb{E}}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta};\check{\boldsymbol{\beta}})|\mathbf{x}_i)\right](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \\
& - \frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{\mathbb{E}}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta};\boldsymbol{\beta}_o|\mathbf{x}_i)))\nabla_{\boldsymbol{\beta}}\hat{\mathbb{E}}_{\boldsymbol{\theta}}(h(\boldsymbol{\theta};\boldsymbol{\beta}_o)|\mathbf{x}_i).
\end{aligned}
$$

Then apply CLT and WLLN. □

Proposition 3 implies at least under quite general assumptions, using the empirical Bayes as denoiser is **equivalent** to using Bayes denoiser with oracle prior.

5 Simulation

5.1 Biased Nonlinear Second Stage

5.1.1 Log Model

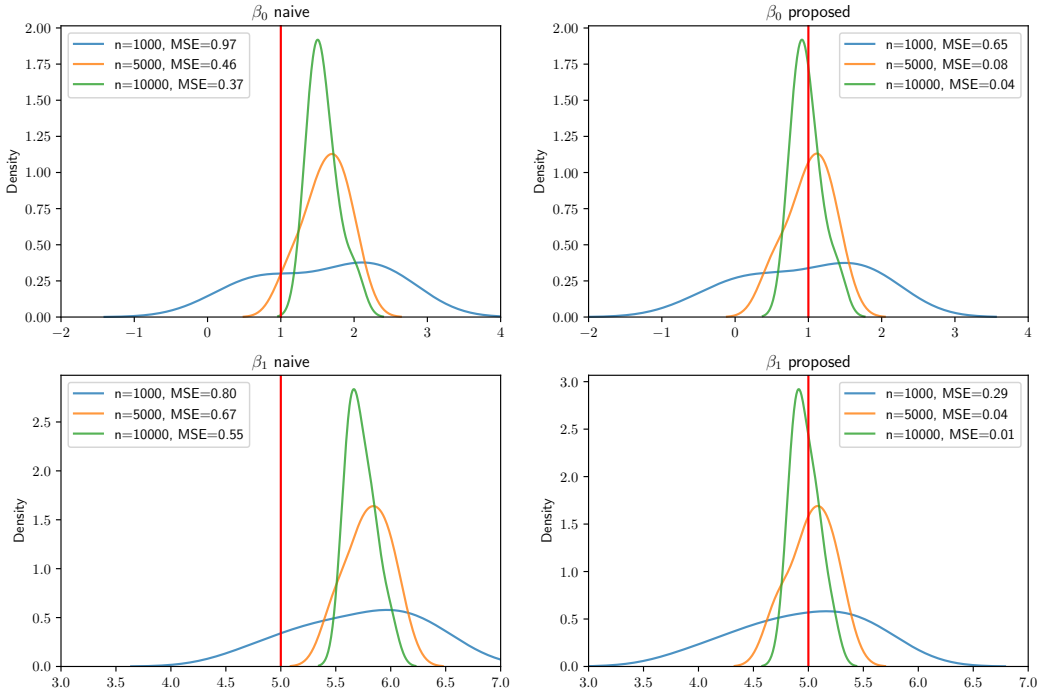
$$\theta_i \sim Beta(a, b)$$

$$10X_i \mid \theta_i \sim Binom(10, \theta_i)$$

$$Y_i = \beta_0 + \beta_1 \ln\theta_i + e_i,\ e_i \sim \mathcal{N}(0, \eta^2)$$

$i = 1 \dots n$,

For this simulation, $m = 10$, $a = 2$, $b = 7$, $\eta = 1.5$ $\beta_0 = 1$, $\beta_1 = 5$. Data are repeatedly generated $B = 500$ times, and below we present the distribution of the estimators and their mean square error performances for different values of n. The result show that the our proposed method is consistent and asymptotically normal, whereas the naive estimator (replacing the unobserved θ_i directly by $\hat{\theta}^{EB}$ in the regression) can lead to bias.



$\widehat{\beta}_0$	naive	proposed
n=1000	0.97	0.65
n=5000	0.46	0.08
n=10000	0.37	0.04

$\widehat{\beta}_1$	naive	proposed
n=1000	0.80	0.29
n=5000	0.67	0.04
n=10000	0.55	0.01

5.1.2 Logit Model

$$\theta_i \sim \mathcal{N}(\mu, \tau^2)$$

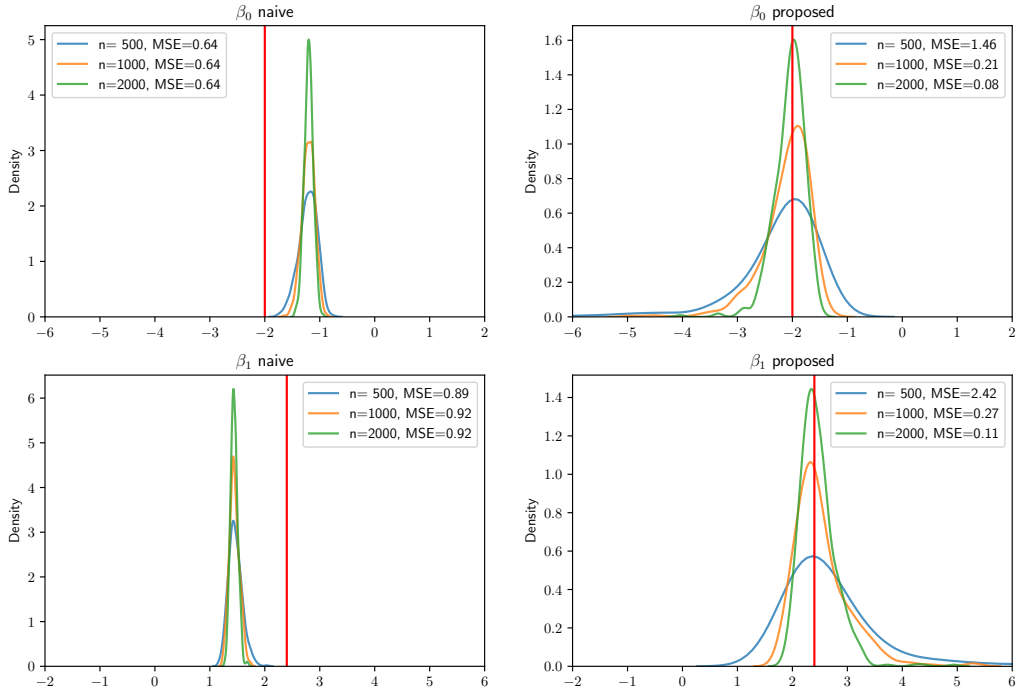
$$X_{ij} \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma^2)$$

$$P(Y_{ij} = 1) = \Lambda(\beta_0 + \beta_1\theta_i);\ P(Y_{ij} = 0) = 1 - \Lambda(\beta_0 + \beta_1\theta_i),$$

where $\Lambda(x) = \frac{1}{1+e^{-x}}$ a logistic function. $i = 1 \ldots n$, $j = 1 \ldots m$.

For this simulation, $m = 10$, $\mu = 0$, $\tau = 4$, $\beta_0 = -2$, $\beta_1 = 2.4$. Data are repeatedly generated $B = 500$ times, and below we present the distribution of the estimators and their mean square error performances for different values of n. Similar to the previous result, the

simulation shows that the our proposed method is consistent and asymptotically normal, whereas the naive estimator can lead to bias.



	$\widehat{\beta}_0$	naive	proposed
	n=500	0.64	1.46
	n=1000	0.64	0.21
	n=2000	0.64	0.08

	$\widehat{\beta}_1$	naive	proposed
	n=500	0.89	2.42
	n=1000	0.92	0.27
	n=2000	0.92	0.11

5.2 Nonparametric Priors

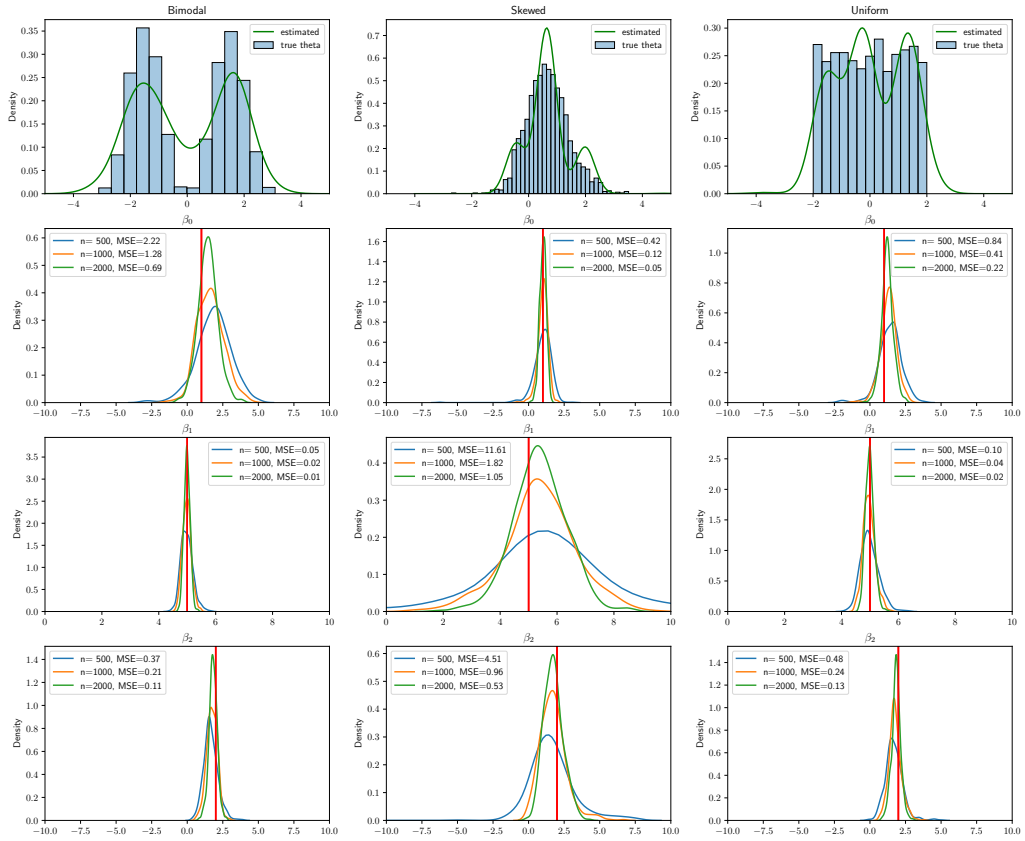
$$\theta_i \sim G$$

$$X_{ij} \mid \theta_i \sim \mathcal{N}(0, \sigma^2)$$

$$Y_{ij} = \beta_0 + \beta_1\theta_i + \beta_2\theta_i^2 + \alpha Z_{ij} + e_{ij},\ e_{ij} \sim \mathcal{N}(0, \eta^2)$$

$\sigma = 3,\ \eta = 2,\ \beta_0 = 1,\ \beta_1 = 5,\ \beta_2 = 2,\ \alpha = 1.2\ m = 15,\ B = 500$

In this simulation exercise, we show that our method can nonparametrically identify the prior and generate consistent estimates of the regression parameter. By contrast, parametric methods with misspecified prior can lead to bias.

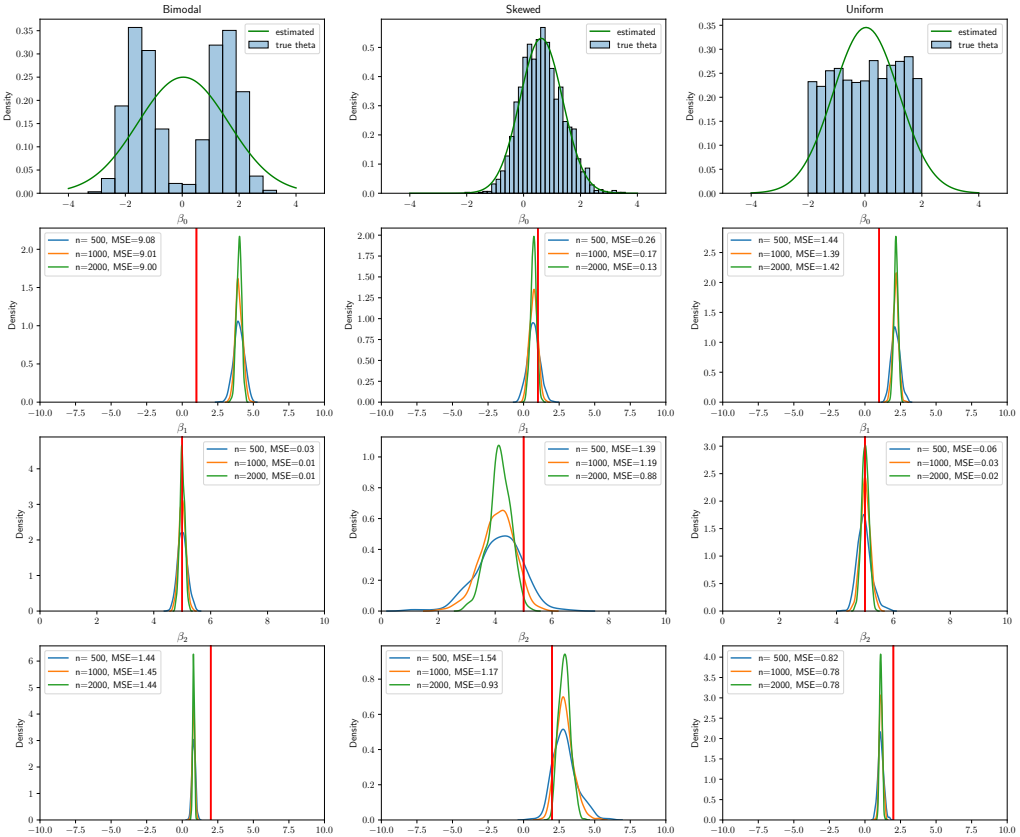


$\widehat{\beta}_0$	Bimodal	Skewed	Uniform
n=500	2.33	0.42	0.84
n=1000	1.28	0.12	0.41
n=2000	0.69	0.05	0.22

$\widehat{\beta}_1$	Bimodal	Skewed	Uniform
n=500	0.05	11.61	0.10
n=1000	0.02	1.82	0.04
n=2000	0.01	1.05	0.02

$\widehat{\beta}_2$	Bimodal	Skewed	Uniform
n=500	0.37	4.51	0.48
n=1000	0.21	0.96	0.24
n=2000	0.11	0.53	0.13

Misspecified prior:



References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Angrist, J., Hull, P., Pathak, P. A., and Walters, C. (2021). Credible school value-added with undersubscribed school lotteries. *The Review of Economics and Statistics*, pages 1–46.

Armstrong, T. B., Kolesár, M., and Plagborg-Møller, M. (2022). Robust empirical bayes confidence intervals. *Econometrica*, 90(6):2567–2602.

Azevedo, E. M., Deng, A., Montiel Olea, J. L., Rao, J., and Weyl, E. G. (2020). A/b testing with fat tails. *Journal of Political Economy*, 128(12):4614–000.

Bonhomme, S. and Weidner, M. (2022). Posterior average effects. *Journal of Business & Economic Statistics*, 40(4):1849–1862.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Chen, J. (2017). Consistency of the mle under mixture models.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9):2633–79.

Chetty, R. and Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, 133(3):1163–1228.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Efron, B. (2016). Empirical bayes deconvolution estimates. *Biometrika*, 103(1):1–20.

Guo, M. and Ghosh, M. (2012). Mean squared error of james–stein estimators for measurement error models. *Statistics & Probability Letters*, 82(11):2033–2043.

Hansen, B. E. (2017). Stein-like 2sls estimator. *Econometric Reviews*, 36(6-9):840–852.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hull, P. (2018). Estimating hospital quality with quasi-experimental data. *Available at SSRN 3118358*.

Ignatiadis, N. and Wager, S. (2019). Covariate-powered empirical bayes estimation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jackson, C. K., Rockoff, J. E., and Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annu. Rev. Econ.*, 6(1):801–825.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, Calif. University of California Press.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906.

Koedel, C., Mihaly, K., and Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, 47:180–195.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32.

Robbins, H. (1956). An empirical bayes approach to statistics. Technical report, COLUMBIA UNIVERSITY New York City United States.

Whittemore, A. S. (1989). Errors-in-variables regression using stein estimates. *The American Statistician*, 43(4):226–228.