# TESTING STATISTICAL HYPOTHESES FOR LATENT VARIABLE MODELS AND SOME COMPUTATIONAL ISSUES

A Dissertation

by

DONG HYUK LEE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

| | |
|---|---|
| Chair of Committee, | Samrian Sinha |
| Committee Members, | Raymond J. Carroll |
| | Lan Zhou |
| | Roger S. Zoh |
| Head of Department, | Valen E. Johnson |

August 2018

Major Subject: Statistics

Copyright 2018 Dong Hyuk Lee

# ABSTRACT

In this dissertation, I address unorthodox statistical problems concerning goodness-of-fit tests in the latent variable context and efficient statistical computations.

In epidemiological and biomedical studies observations with measurement errors are quite common, especially when it is difficult to calibrate true signals accurately. In this first problem, I develop a statistical test for testing equality of two distributions when the observed contaminated data follow the classical additive measurement error model. The fact is that the two-sample homogeneity tests, such as Kolmogorov-Smirnov, Anderson-Darling, or von Mises test, are not consistent when observations are subject to measurement error. To develop a consistent test, first the characteristic functions of unobservable true random variables are estimated from the contaminated data, and then the test statistic is defined as the integrated difference between the two estimated characteristic functions. It is shown that when the sample size is large and the null hypothesis holds, the test statistic converges to an integral of a squared Gaussian process. However, enumeration of this distribution to obtain the rejection region is not simple. Therefore, I propose a bootstrap approach to compute the $p$-value of the test statistic. The operating characteristics of the proposed test is assessed and compared with the other approaches via extensive simulation studies. The proposed method is then applied to analyze the National Health and Nutrition Examination Survey (NHANES) dataset. Although researchers considered estimation of the regression parameters in the presence of exposure measurement error, this testing problem is completely new and no one has considered it before.

In the next problem, I consider the stochastic frontier model (SFM) which is a widely used model for measuring firms' efficiency. In productivity or cost studies in the field of econometrics, there is a discrepancy between the theoretically optimal product and the actual output for a certain amount of inputs and this gap is called technical inefficiency. To assess this inefficiency, the stochastic frontier model is in use to include this gap as a latent variable in addition to the usual statistical noise. Since it is unable to observe this gap, estimation and inference depend on the distributional assumption of the technical inefficiency term. Usually, an exponential or half-normal distribution is widely assumed for the inefficiency term. In that sense, I develop a Bayesian test for testing whether this parametric assumption is correct. I construct a broad semiparametric family which approximate or contain the true distribution as an alternative and then define a Bayes factor. I show the Bayes factor consistency under certain conditions and present the finite sample performance via Monte-Carlo simulations.

The second part of my dissertation is about statistical computational problems. Frequentist standard errors are of interest to evaluate uncertainty of an estimator and utilized for many statistical inference problems. In this dissertation, I consider standard error calculation for Bayes estimators. Except some hypothetical scenarios, estimating frequentist variability of any estimator possibly involves bootstrapping to approximate the sampling distribution of the estimator. In addition, for a Bayesian modeling combined with Markov chain Monte Carlo (MCMC) and bootstrap the computation of the standard error of Bayes estimator is computationally expensive and impractical. Specifically, repeated application of the MCMC on each of the bootstrapped data make everything computationally inefficient. To overcome this difficulty, I propose a clever use of the importance sampling technique to reduce the computational burden. I apply this proposed technique to several examples including logistic regression, linear measurement error model, Weibull regression model and vector autoregressive model.

In the second computational problem, I explore the binary regression with flexible skew-probit link function which contains traditional probit link function as a special case. The skew-probit model is useful for modelling success probability of binary response or count data where the success probability is not a symmetric function of continuous regressors. In this topic, I investigate the parameter identifiability of skew-probit model. I then demonstrate that the maximum likelihood estimator (MLE) of the skewness parameter is highly biased. I develop a penalized likelihood approach based on three penalty functions to reduce the finite sample bias of the MLE of the skew-probit model. The performances of each penalized MLE are compared through extensive simulations and I analyze the heart-disease data using the proposed approaches.

# DEDICATION

*To my family for their constant support and encouragement*

# ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my advisor, Dr. Samiran Sinha, for his enormous guidance and encouragement throughout my doctoral study. He provided me with very interesting topics and showed me how to conduct scientific researches. I also remain indebted for his in-depth advice and supervision during the times when I was applying for the next position. Without his invaluable help, these works wouldn't have been possible. It has been a long journey since I was taking his STAT 605 class which made me study with him.

I also would like to thank Dr. Raymond J. Carroll not only for his thoughtful and constructive advice which improved the work in Chapter 4 substantially but also for a very strong letter of recommendation for me that genuinely helped me to the next position. Moreover, it was a great opportunity to attend his measurement error course where I was able to learn important details of measurement error models for this dissertation.

I would also like to express my gratitude to my committee members, Dr. Lan Zhou and Dr. Roger S. Zoh, for their willingness to serve on my committee and their helpful comments and suggestions. Additionally, I thank Dr. Alan Dabney for teaching two very helpful courses, one on biostatistics and the other on bioinformatics and his nice letter of recommendation. These courses greatly help me in consulting and collaborative works.

I am grateful to my friends, Dr. Jaehong Jeong, Dr. Yei Eun Shin, Raanju Sundararajan, Soutrik Mandal, and Antik Chakraborty for helping me get through the entire program. Thanks also to Dr. Michael Longnecker for his generous support and Ms. Andrea Dawson for her assistance on paperworks.

Last but not the least, I would like to thank my family for their unconditional support and encouragement on this endeavor.

# CONTRIBUTORS AND FUNDING SOURCES

## Contributors

This work was supported by a dissertation committee consisting of Dr. Samrian Sinha (advisor), Dr. Raymond J. Carroll and Dr. Lan Zhou of the Department of Statistics and Dr. Roger S. Zoh of the Department of Epidemiology and Biostatistics.

All work conducted for the dissertation was completed by the student independently.

## Funding Sources

Graduate study was supported by a teaching assistantship from Texas A&M University.

TABLE OF CONTENTS

# LIST OF FIGURES

FIGURE Page

# LIST OF TABLES

TABLE Page

# 1. INTRODUCTION

## 1.1 Goodness-of-fit test in the latent variable context

### 1.1.1 A Two Sample Problem

Suppose that we have observations $D_x = \{X_1, \ldots, X_{n_x}\}$ and $D_y = \{Y_1, \ldots, Y_{n_y}\}$ where $X_1, \ldots, X_{n_x}$ are $n_x$ independently and identically distributed (iid) observations from a distribution $F_x$ and $Y_1, \ldots, Y_{n_y}$ are $n_y$ iid observations from another distribution $F_y$. Moreover assume that $D_x$ and $D_y$ are independent each other. Testing $H_0 : F_x = F_y$ based on observations $D_x$ and $D_y$ has been extensively studied in the literature.

The Kolmogorov-Smirnov test is one of the most widely used test based on the empirical distribution functions (EDF) (Kolmogorov, 1933; Smirnov, 1939a,b). The test statistic is $\sup_t |\widehat{F}_x(t) - \widehat{F}_y(t)|$, where $\widehat{F}_x(t) = (1/n_x) \sum_{i=1}^{n_x} I(X_i \leq t)$ and $\widehat{F}_y(t) = (1/n_y) \sum_{i=1}^{n_y} I(Y_i \leq t)$. Kuiper (1960) proposed a similar test statistic $\sup_t(\widehat{F}_x(t) - \widehat{F}_y(t)) - \inf_t(\widehat{F}_x(t) - \widehat{F}_y(t))$ and Maag and Stephens (1968) provided tables of the above test statistic.

An alternative to the Kolmogorov-Smirnov test is Cramér-von Mises type test (Cramér, 1928; von Mises, 1931; Smirnov, 1936, 1937). The test statistic has form of $\int_{-\infty}^{\infty} \{\widehat{F}_x(t) - \widehat{F}_y(t)\}^2 \omega(t) d\widehat{F}(t)$, where $(n_x + n_y)\widehat{F} = n_x\widehat{F}_x + n_y\widehat{F}_y$ is the EDF of pooled sample $D_x$ and $D_y$, $\omega(t) = 1$ corresponds to von Mises statistic and $\omega(t) = \{\widehat{F}(1 - \widehat{F})\}^{-1}$ corresponds to Anderson-Darling statistic (Rosenblatt, 1952; Darling, 1957; Kiefer, 1959; Fisz, 1960; Anderson, 1962; Pettitt, 1976; Scholz and Stephens, 1987). See Stephens (1992) and references therein for more information on tests based on the EDF.

Other than aforementioned tests, one can construct tests based on the empirical characteristic function (Fan, 1997; Alba et al., 2001; Jiménez-Gamero et al., 2009). Zhang (2002, 2006) developed goodness-of-fit test using the likelihood ratio statistic following the Cressie-Read family of divergence statistics (Cressie and Read, 1984). He demonstrated that the tests derived from the likelihood ratio statistic are as powerful as traditional EDF based test for location difference problems, while they are more powerful in scale or shape change.

However, in observational studies, $D_x$ and $D_y$ may not be available, rather one can have replicated contaminated observations for $D_x$ and $D_y$. In that case, one can form averages out of replicated observations, and apply the traditional two-sample tests on the averages. If the number of replications is not large, this naive approach can produce misleading results. To circumvent this issue, we propose a consistent two-sample test when direct observations on $D_x$ and $D_y$ are not available. The detailed methodologies are discussed in Chapter 2.

### 1.1.2 Stochastic Frontier Model

Consider the following Cobb-Douglas production frontier model

$$\log(Q_i) = \beta_0 + \beta_1 \log(K_i) + \beta_2 \log(L_i) - u_i + v_i, i = 1, \dots, n, \tag{1.1}$$

where $Q_i$ is total production (the real value of all goods produced in a year), $K_i$ is capital input (the real value of all machinery, equipment etc) and $L_i$ is labor input (the total number of person-hours worked in a year) for the $i^{th}$ company. Here $\log(Q_i^*) = \beta_0 + \beta_1 \log(K_i) + \beta_2 \log(L_i) + v_i$ is considered as the optimal frontier goal such as maximum production or minimum costs, where $v_i$ is a random error outside of capital and labor input. However, there are discrepancies between the actual production and the theoretical maximum production. This gap is called technical inefficiency, $u_i$, and making inference on this inefficiency term is the key purpose of the considering this production model (Aigner et al., 1977; Meeusen and van den Broeck, 1977). For interpretation, $\log(Q_i) - \log(Q_i^*) = -u_i \leq 0$ and this implies $Q_i/Q_i^* = e^{-u}$, where $Q_i$ is the actual production and $Q_i^*$ is the theoretical optimal production. Therefore, $u_i$ is assumed to be a positive random variable, while $u_i = 0$ means that the company attains the full efficiency as a special case.

In general, (1.1) can be written in the usual linear regression form given by $y_i = \beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta}_1 - u_i + v_i$. In terms of the statistical inference, the difficulty arises because $u_i$ is latent and unobservable quantity. Thus inferences are based on the specific assumption of the distribution of $u_i$ while $v_i$ is generally assumed to be $\mathrm{Normal}(0, \sigma_{\mathrm{v}}^2)$. Aigner et al. (1977) considered a half-normal distribution, Meeusen and van den Broeck (1977) assumed exponential distribution, Stevenson (1980) used truncated normal distribution and Greene (1990) adopted gamma distribution for $u_i$. For example, suppose $u_i \sim \mathrm{Normal}^{+}(0, \sigma_{\mathrm{u}}^2)$. The density function of $u$ is $f_u(u) = 2\exp\left(-u^2/2\sigma_u^2\right)/\sqrt{2\pi\sigma_u^2}$, and from this assumption one can derive the density function of $\epsilon = -u + v$:

$$\begin{aligned} f_\epsilon(\epsilon) &= \int_0^\infty f_{\epsilon|u}(\epsilon|u) f_u(u) du = \int_0^\infty \frac{2}{2\pi\sigma_u\sigma_v} \exp\left\{-\frac{(\epsilon+u)^2}{2\sigma_v^2} - \frac{u^2}{2\sigma_u^2}\right\} du \\ &= \frac{2}{\sigma}\phi\left(\frac{\epsilon}{\sigma}\right)\Phi\left(-\lambda\frac{\epsilon}{\sigma}\right), \end{aligned}$$

where $\phi$ and $\Phi$ are the density function and the distribution function of the $\mathrm{Normal}(0,1)$, $\sigma^2 = \sigma_v^2 + \sigma_u^2$ and $\lambda = \sigma_u/\sigma_v$. This reparameterization $(\sigma^2, \lambda)$ is conventional in the literature. The maximum likelihood estimator (MLE) of $\beta_0, \boldsymbol{\beta}_1, \sigma^2, \lambda$ can be obtained by maximizing the log-likelihood function $\ell(\beta_0, \boldsymbol{\beta}_1, \sigma, \lambda) = -n\log\sigma - (1/2\sigma^2)\sum_{i=1}^n (y_i - \beta_0 - \boldsymbol{x}_i'\boldsymbol{\beta}_1) + \sum_{i=1}^n \log[\Phi\{-\lambda(y_i - \beta_0 - \boldsymbol{x}_i'\boldsymbol{\beta}_1)/\sigma\}]$.

Once the MLE of $(\beta_0, \boldsymbol{\beta}_1, \sigma, \lambda)$ is computed, it is possible to calculate the conditional density of $u_i$ given $\epsilon_i$ which makes the prediction of individual technical inefficiency possible (Jondrow et al., 1982). Under the half normal distribution assumption, $u_i|\epsilon_i$ is $\mathrm{Normal}(\mu_*, \sigma_*^2)$ truncated at 0, where $\mu_* = -\sigma_u^2\epsilon_i/\sigma^2$ and $\sigma_*^2 = \sigma_u^2\sigma_v^2/\sigma^2$, and $\mu_* = E(u_i|\epsilon_i)$ is used as the predictor of the technical inefficiency of the $i^{th}$ producer (Jondrow et al., 1982). Estimation and prediction details for other parametric models are well summarized in Kumbhakar and Lovell (2003).

In addition to the above estimation approaches, van den Broeck et al. (1994) developed a Bayesian framework under Gamma distribution assumption with shape parameters 1, 2 and 3 of the technical inefficiency. Griffin and Steel (2004) proposed a semiparametric Bayesian approach by considering a non-parametric prior distribution on the distribution of $u_i$. Instead of linearity assumption between the input and output variables, Simar et al. (2017) adopted a nonparametric method to estimate the relationship between these two variables.

As discussed so far, the estimation and prediction heavily depend on the distributional assumption of the technical inefficiency. Specifically, different assumption of $u_i$ will change the conditional distribution of $u_i$ given $\epsilon_i$ and so does the predictor $E(u_i|\epsilon_i)$. Therefore, diagnostics and model checking is a crucial step after fitting the model. In this dissertation, I propose a Bayesian goodness-of-fit test for checking the distributional assumption of the technical inefficiency without specifying the alternative model. This is described in Chapter 3.

## 1.2 Efficient Statistical Computation

### 1.2.1 Standard error of Bayes Estimators

I begin with a simple example illustrating the basic concept. Suppose that $X_1, \ldots, X_n$ are a sample from the $\mathrm{Normal}(\mu, \sigma^2)$ distribution with both unknown mean $\mu$ and variance $\sigma^2$. Assume that given $\sigma^2$, a prior $\mu|\sigma^2 \sim \mathrm{Normal}(m, \tau\sigma^2)$, and $\sigma^2 \sim \mathrm{IG}(a, b)$, where $\mathrm{IG}(a, b)$ refers to the inverse gamma density, i.e., $\pi(\sigma^2) = \exp(-1/b\sigma^2)/\Gamma(a)b^a(\sigma^2)^{a+1}$. I also assume that the hyperparameters $a, b, \tau$ are specified. Then the joint posterior density of $(\mu, \sigma^2)$ is given by

$$\pi(\mu, \sigma^2|X_1, \ldots, X_n) \propto (\sigma^2)^{-n/2-1/2-a-1} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{1}{2\tau\sigma^2}(\mu-m)^2 - \frac{1}{b\sigma^2}\right\}.$$

Therefore,

$$\mu|\sigma^2, X_1, \ldots, X_n \sim \mathrm{Normal}\left(\frac{n\bar{X}+m/\tau}{n+1/\tau}, \frac{\sigma^2}{n+1/\tau}\right),$$

and

$$\sigma^2|X_1, \ldots, X_n \sim IG\left(\frac{n}{2}+a, \left\{\frac{(n-1)s^2}{2} + \frac{n\bar{X}^2}{2} + \frac{m^2}{2\tau} + \frac{1}{b} - \frac{(n\bar{X}+m/\tau)^2}{2(n+1/\tau)}\right\}^{-1}\right),$$

where $s^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. The posterior mean of $\mu$ is $\widehat{\mu} = (n\bar{X} + m/\tau)/(n + 1/\tau)$. Note that its variance is

$$\mathrm{var}(\widehat{\mu}) = \mathrm{var}\left(\frac{n\bar{X}+m/\tau}{n+1/\tau}\right) = \frac{n}{(n+1/\tau)^2}\sigma^2.$$

Similarly, the posterior mean of $\sigma^2$ is

$$\widehat{\sigma}^2 = \Big(\frac{n}{2} + a - 1\Big)^{-1} \left\{ \frac{(n-1)s^2}{2} + \frac{n\bar{X}^2}{2} + \frac{m^2}{2\tau} + \frac{1}{b} - \frac{(n\bar{X} + m/\tau)^2}{2(n+1/\tau)} \right\}$$

and its variance is

$$\text{var}(\widehat{\sigma}^2) = \Big(\frac{n}{2} + a - 1\Big)^{-2} \left\{ \frac{n\sigma^2(2\mu^2 + \sigma^2)}{2} - \frac{(2\mu^2 + \sigma^2/n)\sigma^2/n}{2(n+1/\tau)^2} \right\}.$$

These variabilities var$(\widehat{\mu})$ and var$(\widehat{\sigma}^2)$ measure the accuracy of the posterior expectation and serve as a proxy for evaluating uncertainty of a procedure. Efron (2015) pointed out that frequentist accuracy can be a criterion to choose a non-informative prior when trustful past information is unavailable. In computational aspects, however, it is not always possible to have a closed form expression of the variance or the standard error of a posterior mean as in this example. Moreover in general, posterior expectation is not of only interest, rather other posterior summaries are needed such as posterior mode, $\alpha^{th}$ quantile. The standard error of those posterior summaries usually does not have a closed form even in the simple examples. I investigate a general approach to estimate the standard error of Bayes estimators, that is computationally more efficient than the naive approach. This topic is investigated in Chapter 4.

### 1.2.2 A Binary Regression with Skew-Probit Link

Suppose that we observe $\{(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$ where $Y_i$ is the binary response, i.e., $Y_i = 1$ if $i^{th}$ subject experiences the primary outcome and $Y_i = 0$ otherwise and $\boldsymbol{X}_i$ is a vector of covariates. Then for modelling conditional probability pr$(Y = 1|\boldsymbol{X})$ usually logistic or probit model is considered. One characteristic of them is that they are symmetric in the sense that they approach to 0 and 1 at the same rate. This is because their densities are symmetric around 0. In a practical setting, there is no reason to believe that success probability must be modeled via symmetric link. Rather we should fit a flexible model to the data, and allow the data to choose an appropriate model (symmetric and asymmetric link).

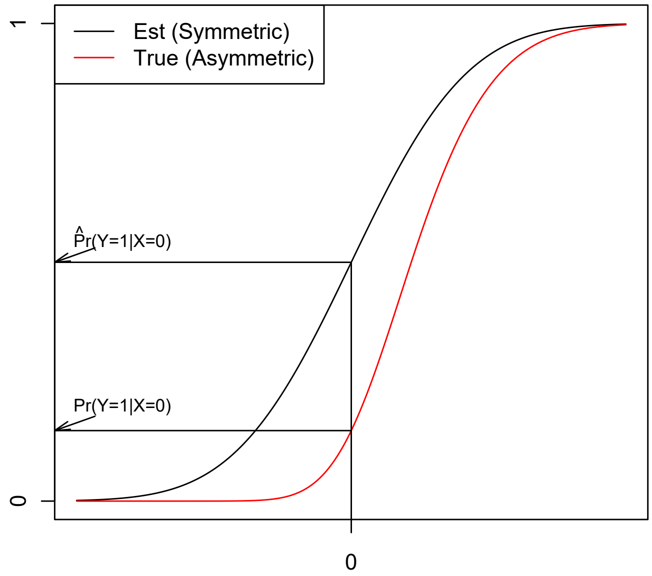

Figure 1.1: A hypothetical example of the true conditional probability pr$(Y|X)$ (red line) and the estimated pr$(Y|X)$ (black line).

Figure 1 shows the difference between a symmetric and asymmetric link function. The black line shows how the success probability of a binary response variable Y is changing with a continuous X when the probit (symmetric) link function is used. In contrast, the red line shows the change when the success probability is regulated by an asymmetric link. Noticeably, under the asymmetric link, pr(Y=1|X) is very close to zero when X<0.

Chen et al. (1999) proposed a class of asymmetric link functions for binary regression in a Bayesian context. One special link function they considered is the skew-probit link where pr$(Y|X)$ is modeled using the CDF of the standard skew normal distribution (Azzalini, 1985). As the name implies the skew-probit link function includes probit link function as a special case. In Chapter 5, I explore two crucial problems, identifiability and the bias of the MLE of the model parameters.

# 2. A TEST OF HOMOGENEITY OF DISTRIBUTIONS WHEN OBSERVATIONS ARE SUBJECT TO MEASUREMENT ERRORS

## 2.1 Background and literature review

As discussed in Section 1.1.1, mismeasured variables are common in epidemiological and biomedical studies. Failure to account these errors in the measured variables might lead to incorrect statistical inference. One such motivating data comes from the National Health and Nutrition Examination Survey (NHANES) which is designed to assess the health and nutritional status of adults and children in the United States. Besides dietary intakes, a number of biomarkers, such as blood pressure, albumin level, creatinine level are difficult to measure accurately. To handle uncertainty in the measurements, in the NHANES study multiple measurements are taken on these variables.

A common public health question is how the behavioral factors are associated with a biomarker, a health outcome, or a surrogate of a health outcome (Hogan et al., 2007; Puddey and Beilin, 2006; Primatesta et al., 2001). In particular, like others, I am interested to verify if alcohol consumption and the systolic blood pressure are associated. In an attempt to answer this question one may use the NHANES data, and apply a two-sample nonparametric test to the average of multiple measurements from the two behavioral groups, alcoholic and non-alcoholic. However, the standard testing tools are inappropriate as the observed data are contaminated with measurement errors. As shown in the simulation study, in this contaminated data scenario, the standard tests that ignore measurement errors likely to result in a wrong conclusion. This motivates to develop a new two-sample testing method when the available data are measured with errors.

In this chapter, I consider testing $H_0 : F_x = F_y$ when neither $D_x$ nor $D_y$ is observed, rather, we observe replicated erroneous observations for $D_x$ and $D_y$. In particular, our observed data are $D_w = \{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{n_x}\}$ and $D_v = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{n_y}\}$, where $\boldsymbol{W}_j^T = (W_{j1}, \ldots, W_{jm_x})$ and $\boldsymbol{V}_k^T = (V_{k1}, \ldots, V_{km_y})$ for $j = 1, \ldots, n_x$ and $k = 1, \ldots, n_y$. Assume that $m_x \geq 2$ and $m_y \geq 2$, and the observed $W$'s are related with the unobserved $X$'s through the classical additive measurement error model, that means,

$$W_{jl} = X_j + U_{x,jl},\ l = 1, \ldots, m_x,\ j = 1, \ldots, n_x, \tag{2.1}$$

and the measurement error $U_{x,jl}$'s are assumed to be iid, independent of $X_j$, and follows the distribution $F_{u_x}$ that is symmetric around $0$. Similarly, further assume that

$$V_{kl} = Y_k + U_{y,kl},\ l = 1, \ldots, m_y,\ k = 1, \ldots, n_y, \tag{2.2}$$

and the measurement error $U_{y,kl}$ are assumed to be iid, independent of $Y_k$, and follows the distribution $F_{u_y}$ that is symmetric around $0$. Specifically, here $D_x$ and $D_y$ are latent observations and are never observed. The above mentioned CDF $F_x$, $F_y$, $F_{u_x}$ and $F_{u_y}$ are assumed to be absolutely continuous but otherwise left unknown.

Although errors in measured variables have received considerable attention from density estimation perspective (Carroll and Hall, 1988; Delaigle and Hall, 2016) and in the regression context (Gustafson, 2003; Carroll et al., 2006), no one has ever considered testing of homogeneity of two distributions when the observed data are subject to measurement errors that are common in observational studies. The statistical test of homogeneity of distributions is widely used in social and medical sciences and in the field of Engineering. Given its importance in various fields and error contamination in the observed data are commonplace, a consistent test is urgently needed.

Like errors in covariates in regression models, this problem can be tackled in several ways. First, one may model all the distributions, $F_x, F_y, F_{u_x}, F_{u_y}$ parametrically, and then test $H_0$ by checking equality of a set of parameters. However, any parametric approach may face misspecification bias. Therefore, we do not wish to use any parametric model assumption. In the nonparametric context, one may estimate the densities of $X$ and $Y$ from the contaminated data using any density deconvolution approach available in the literature (Delaigle et al., 2008; Delaigle and Hall, 2016). Then carry out a test based on the deconvoluted densities. Numerical instability is a well known phenomenon of deconvoluted density estimation, and that is due to the inverse transformation of the characteristic function. To circumvent this problem, I design a test that is directly based on the characteristic functions, and the test statistic itself does not depend on the deconvoluted density.

Briefly, proposed approach can be described into two simple steps. First, the characteristic functions of $X$ and $Y$ from the contaminated data is estimated. Next, a test statistic is formulated by using the two estimated characteristic functions. Then the asymptotic distribution of the proposed test statistic under $H_0$ will be derived. The limit distribution has a complex form and it involves different unknown population parameters, making it less appealing to use for calibrating the test statistic. Motivated by this problem, I propose a novel Bootstrap approach under the measurement error framework that gives a theoretically valid data generation procedure under the null hypothesis, and that also constitutes an important contribution of this work. In addition to theoretical investigation of the large sample properties of the Bootstrap based testing procedure, finite sample properties of the test are judged via simulation studies. The results of the simulation study show that the proposed testing method has competitive performance in terms of maintaining the size of the test, and superior power properties compared to its competitors, even when the two population distributions are not drastically different. Finally, I analyze the real datasets that motivated me to consider this research problem.

Chapter 2 is organized as follows. Section 2.2 gives the formulation of the test statistic based on the estimated characteristic functions under the measurement error model (2.1) and (2.2) and investigates its asymptotic properties. Section 2.3 describes the details of the Bootstrap method and proves its theoretical validity. Results from a moderately large simulation study are given in Section 2.4, showing the performance of the proposed testing method under the null and under different alternatives. An application of the methodology to an NHANES 2009-2010 survey data is given in Section 2.5, followed by some concluding remarks in Section 2.6. Proofs of the main results are given in the Appendix A.

## 2.2 Testing methodology

### 2.2.1 Notation

First, let us introduce notations that will be used throughout this chapter. Let $\phi_x$, $\phi_y$, $\phi_{u_x}$ and $\phi_{u_y}$ be the characteristic functions of $X$, $Y$, $U_x$ and $U_y$, respectively. Let $a_x(t)$ and $b_x(t)$ be the real and imaginary parts of $\phi_x(t)$, respectively. Similarly, define $a_y(t)$ and $b_y(t)$ from $\phi_y(t)$. For future reference, denote estimators of $F_x, F_y, F_{u_x}$, and $F_{u_y}$, by $\widehat{F}_x$, $\widehat{F}_y$, $\widehat{F}_{u_x}$, and $\widehat{F}_{u_y}$ respectively. Suppose that $F = F_x = F_y$ denotes the common distribution under $H_0$, and write $\widehat{F}$ to denote its estimator. Further, define $\overline{W}_j = \sum_{l=1}^{m_x} W_{jl}/m_x$, $\overline{V}_k = \sum_{l=1}^{m_y} V_{kl}/m_y$, $M_x = m_x(m_x - 1)/2$, $M_y = m_y(m_y - 1)/2$, $N_x = n_x M_x$, $N_y = n_y M_y$. Note that the characteristic function of $\overline{W}_j$ is given by $\phi_1(t) = \phi_x(t)\{\phi_{u_x}(t/m_x)\}^{m_x}$, and that of $\overline{V}_k$ is $\phi_2(t) = \phi_y(t)\{\phi_{u_y}(t/m_y)\}^{m_y}$.

In the naive approach that ignores measurement errors in the observed data, one may first compute $\{\overline{W}_1, \ldots, \overline{W}_{n_x}\}$ and $\{\overline{V}_1, \ldots, \overline{V}_{n_y}\}$ and then apply any nonparametric testing procedure directly on these transformed data. Indeed, this naive method is usually inconsistent, that means, it fails to maintain the nominal type-I error level. If $m_x = m_y = m$, and $F_{u_x} = F_{u_y} = F_u$, then $\phi_1(t) = \phi_x(t)\{\phi_u(t/m)\}^m$ and $\phi_2(t) = \phi_y(t)\{\phi_u(t/m)\}^m$. Consequently the null hypothesis $H_0 : \phi_x(t) = \phi_y(t)$ implies $\phi_1(t) = \phi_2(t)$. That means, testing $H_0$ becomes equivalent to testing $H_0 : F_1 = F_2$, where $F_1$ and $F_2$ are the distribution functions of $\overline{W}_j$ and $\overline{V}_k$. Thus, when $m_x = m_y$ and $F_{u_x} = F_{u_y}$, the naive testing procedure is consistent for testing $H_0 : F_x = F_y$. However, if either $m_x \neq m_y$ or $F_{u_x} \neq F_{u_y}$, the naive test may not be consistent.

### 2.2.2 Development of the test statistic

I shall work under the standard condition (Delaigle et al., 2008) that $\phi_{u_x}(t)$ and $\phi_{u_y}(t)$ are real-valued function and do not vanish on $\mathbb{R}$, but do not impose any such conditions on the characteristic functions $\phi_x$ and $\phi_y$ of the (true) latent variables. The real valued characteristic function condition results from the assumption that the error distribution is symmetric around zero. Further, as is well known (Stefanski and Carroll, 1990), the non vanishing assumption is also due to overcome the identifiability problem. Under these conditions, the characteristic function for the measurement error can be recovered using the difference between two observations $W_1 - W_2$, where $W_1 = X + U_{x,1}$ and $W_2 = X + U_{x,2}$. Then, $\phi_{W_1 - W_2}(t) = E[\exp\{it(W_1 - W_2)\}] = E[\exp\{it(U_{x,1} - U_{x,2})\}] = E[\exp\{itU_{x,1}\}]E[\exp\{-itU_{x,2}\}] = \{\phi_{u_x}(t)\}^2$, where $i^2 = -1$. Hence $\phi_{u_x}$ is estimable from the data by using all possible pairwise differences of the $W_{jk}$ variables. On the other hand, $\phi_1(t)$ is directly estimable from the data, using the means of the replicated measurements. Consequently, $\phi_x(t)$ is estimable exploiting the relationship $\phi_1(t) = \phi_x(t)\{\phi_{u_x}(t/m_x)\}^{m_x}$. Specifically, estimators for $\phi_1(t)$ and $\phi_{u_x}(t)$ are given by

$$\widehat{\phi}_1(t) = \frac{1}{n_x}\sum_{j=1}^{n_x} \exp(it\overline{W}_j),$$

$$\widehat{\phi}_{u_x}(t) = \sqrt{|\widehat{\phi}_{W_1-W_2}(t)|} = \sqrt{\left|\frac{1}{n_x}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{t(W_{jl_1} - W_{jl_2})\}\right|}, \tag{2.3}$$

respectively, where $\mathcal{S}_x = \{(l_1, l_2) : 1 \le l_1 < l_2 \le m_x\}$. Note that the non-vanishing and continuity assumption on $\phi_{u_x}(t)$, and $\phi_{u_x}(0) = 1$ imply that $\phi_{u_x}(t)$ is a positive real valued function. Thus, the above estimator of $\phi_{u_x}(t)$ is positive on compact subsets with high probability, for $n_x$ large. Now, we propose to estimate $\phi_x(t)$ by

$$\begin{aligned}\widehat{\phi}_x(t) &= \frac{\widehat{\phi}_1(t)}{\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}} = \frac{n_x^{-1}\sum_{j=1}^{n_x}\cos(t\overline{W}_j) + in_x^{-1}\sum_{j=1}^{n_x}\sin(t\overline{W}_j)}{|n_x^{-1}\sum_{j=1}^{n_x} M_x^{-1}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{(t/m_x)(W_{jl_1} - W_{jl_2})\}|^{m_x/2}} \\ &= \widehat{a}_x(t) + i\widehat{b}_x(t),\end{aligned}$$

where $\widehat{a}_x(t)$ and $\widehat{b}_x(t)$ are the real and imaginary part of $\widehat{\phi}_x(t)$, respectively, and we write

$$\widehat{a}_x(t) = \frac{n_x^{-1}\sum_{j=1}^{n_x} c_{jw}(t)}{\widehat{a}_{2x}(t)},\ \widehat{b}_x(t) = \frac{n_x^{-1}\sum_{j=1}^{n_x} d_{jw}(t)}{\widehat{a}_{2x}(t)}, \tag{2.4}$$

with $c_{jw}(t) = \cos(t\overline{W}_j)$, $d_{jw}(t) = \sin(t\overline{W}_j)$, and

$$\widehat{a}_{2x}(t) = |n_x^{-1}\sum_{j=1}^{n_x} M_x^{-1}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{(t/m_x)(W_{jl_1} - W_{jl_2})\}|^{m_x/2}. \tag{2.5}$$

Similarly, $\phi_y(t)$ can be estimated by $\widehat{\phi}_y(t) = \widehat{a}_y(t) + i\widehat{b}_y(t)$, where $\widehat{a}_y(t) = n_y^{-1} \sum_{j=1}^{n_y} c_{jv}(t)/\widehat{a}_{2y}(t)$ and $\widehat{b}_y(t) = n_y^{-1} \sum_{j=1}^{n_y} d_{jv}(t)/\widehat{a}_{2y}(t)$, with $c_{jv}(t) = \cos(t\overline{V}_j)$, $d_{jv}(t) = \sin(t\overline{V}_j)$, and

$$\widehat{a}_{2y}(t) = |n_y^{-1} \sum_{j=1}^{n_y} M_y^{-1} \sum_{(l_1,l_2)\in\mathcal{S}_y} \cos\{(t/m_y)(V_{jl_1} - V_{jl_2})\}|^{m_y/2}.$$

Under the null hypothesis $F_x = F_y$, $(\widehat{a}_x(t), \widehat{b}_x(t))$ is expected to be close to $(\widehat{a}_y(t), \widehat{b}_y(t))$. When the null hypothesis does not hold, the difference between them is expected to be large, and this fact motivates to form the following test statistic to test the hypothesis $H_0 : F_x = F_y$:

$$T_{n_x} = \int_{-\infty}^{\infty} n_x[\{\widehat{a}_x(t) - \widehat{a}_y(t)\}^2 + \{\widehat{b}_x(t) - \widehat{b}_y(t)\}^2]\omega(t)dt, \tag{2.6}$$

for a properly chosen non-negative weight function $\omega(t)$. The test function is

$$\Phi = \begin{cases} 1 & \text{if } T_{n_x} > t_{n_x,\alpha} \\ 0 & \text{otherwise,} \end{cases}$$

where the critical value $t_{n_x,\alpha}$ satisfies $\text{pr}(T_{n_x} > t_{n_x,\alpha}) = \alpha$ under $H_0$, for a given $\alpha \in (0, 1)$.

In (2.6), the weight function $\omega(t)$ is used for ensuring the finiteness of the integral on the right side, and it is typically taken as a compactly supported function. As expected, the power of the test depends on the weight function $\omega(t)$. In a related work, Epps and Pulley (1983) proposed a test for normality based on the empirical characteristic function of the observed data without measurement errors and described some desirable properties of $\omega(t)$. Here we follow Epps and Pulley (1983)'s guidance and take $\omega(t)$ to be a piece-wise continuous positive valued function with a compact support $[t_1, t_2]$ that includes 0, and $\omega(t) = 0$ for $t > t_2$ or $t < t_1$. For more details on some practical choices for $t_1$ and $t_2$, see the simulation and data analysis section.

### 2.2.3 Large Sample properties of the test statistic

The first result gives the null distribution of the test statistic.

**Theorem 1.** *Under the null hypothesis, as $n_x, n_y \to \infty$ and $\sqrt{n_x/n_y} \to \rho \in (0, \infty)$, the test statistic $T_{n_x}$ converges to a random variable, given by*

$$\int [\xi_1(t)^2 + \xi_2(t)^2]\omega(t)dt$$

*where $\xi_1(\cdot)$ and $\xi_2(\cdot)$ are independent zero mean Gaussian processes with continuous sample paths, with probability one. The covariance functions of $\xi_j(\cdot)$, $j = 1, 2$ are rational functions of the (real and imaginary parts of the ) characteristic functions of $\overline{W}_1$, $\overline{V}_1$, $U_{x,1}$ and $U_{y,1}$, and are given in the Appendix A.*

It follows from the statement of Theorem 1 that the limit distribution of the test statistic can also be expressed as an infinite sum of weighted, independent Chi-squared random variables with degrees of freedom 1. However, the weights in the infinite series representation or the covariance function of the Gaussian processes $\xi_j(\cdot)$, $j = 1, 2$ in the integral representation above are complicated functions of unknown population parameters that are difficult to estimate under the measurement error model. As a result, a Bootstrap method is developed in order to devise alternative approximations to the null distribution of the test statistic that can be used for calibrating the test.

The next result shows that under mild conditions, the power of the test statistic under alternative hypothesis tends to one. To state it, define $D_a(t) = a_x(t) - a_y(t)$ and $D_b(t) = b_x(t) - b_y(t)$.

**Theorem 2.** *Suppose that $\sqrt{n_x/n_y} \to \rho \in (0, \infty)$ and that the alternative hypothesis $\int\{D_a^2(t) + D_b^2(t)\}\omega(t)dt \neq 0$ holds. Then, for any $\alpha \in (0, 1)$, the power of the size $\alpha$ test, pr$(T_{n_x} > t_{n_x,\alpha})$ tends to $1$ as $n_x, n_y \to \infty$.*

## 2.3 The proposed Bootstrap method

### 2.3.1 Outline of the Bootstrap procedure

In this section, I describe a novel Bootstrap method for approximating the null distribution of the test statistic given in Theorem 1. Note that due to the presence of the measurement error, simple resampling from the original data will not capture the distributions of the latent variables and the error variables precisely. In addition, resampling the observations directly will also fail to ensure that the data are generated under the null hypothesis. Therefore, I propose to generate observations from a suitable estimated common distribution $\widehat{F}$ of the two populations for the latent variables, enforcing the null distribution. I also independently generate observations from estimated distribution functions $\widehat{F}_{u_x}$ and $\widehat{F}_{u_y}$ of the two sets of error variables and combine them to define the Bootstrap analogues of $W$ and $V$. Exact constructions of $\widehat{F}$ and $\widehat{F}_{u_x}$ (and $\widehat{F}_{u_y}$) are described in Sections 2.3.2 and 2.3.3 below.

A Bootstrap sample will consist of $D_w^* = \{\boldsymbol{W}_1^*, \ldots, \boldsymbol{W}_{n_x}^*\}$ and $D_v^* = \{\boldsymbol{V}_1^*, \ldots, \boldsymbol{V}_{n_y}^*\}$, where $\boldsymbol{W}_j^* = (W_{j1}^*, \ldots, W_{jm_x}^*)^T$, $j = 1, \ldots, n_x$ and $\boldsymbol{V}_k^* = (V_{j1}^*, \ldots, V_{km_y}^*)^T$, $k = 1, \ldots, n_y$, with $W_{jl}^* = X_j^* + U_{x,jl}^*$ and $V_{kl}^* = Y_k^* + U_{y,kl}^*$. Here, $X_1^*, \ldots, X_{n_x}^*, Y_1^*, \ldots, Y_{n_y}^*$ are iid draws from the estimated common distribution $\widehat{F}$, and $U_{x,jl}^*$ are iid draws from $\widehat{F}_{u_x}$ and $U_{y,kl}^*$ are iid draws from $\widehat{F}_{u_y}$. For each Bootstrap sample, we would compute the test statistic. Suppose that $T_{b,n_x}^*$ denotes the test statistic corresponding to the $b^{th}$ Bootstrap sample. Then the estimated $p$-value is $\sum_{b=1}^{B} I(T_{b,n_x}^* > T_{n_x})/B$ based on $B$ Bootstrap samples. We reject $H_0$ at the $100\alpha\%$ level of significance if the $p$-value is less than a given $\alpha$. Now I describe how to estimate $F$, $F_{u_x}$, and $F_{u_y}$ nonparametrically. Validity of the Bootstrap approximation is proved in Section 2.3.4.

### 2.3.2 Estimation of the common distribution $F$

Let $g$ be a density function of $\overline{W}$, the mean of $m_x$ repeated observations. Then for a symmetric kernel $K$ and given bandwidth $h_w$, $\widehat{g}(w) = (n_x h_w)^{-1} \sum_{j=1}^{n_x} K\{(w - \overline{W}_j)/h_w\}$ is a kernel density estimator for $g$, and consequently the estimated characteristic function of $\overline{W}$ is

$$\widehat{\phi}_{\overline{W}}(t) = \int \exp(itw)\widehat{g}(w)dw = \frac{1}{n_x}\sum_{j=1}^{n_x} \exp(it\overline{W}_j) \int \exp(ith_w z)K(z)dz = \widehat{\phi}_1(t)\phi_K(h_w t),$$

where $\widehat{\phi}_1(t)$ is the empirical characteristic function of $\overline{W}$ and $\phi_K(t)$ is the characteristic function of the kernel $K$. Therefore, the estimated characteristic function $\widehat{\phi}_x(t) = \widehat{\phi}_{\overline{W}}(t)/\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x} = \widehat{\phi}_1(t)\phi_K(h_w t)/\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}$. I want to point out that due to integrability requirement for the estimated characteristic function, here I am using a different approach to estimating $\phi_x(t)$ than that used in Section 2.2.2. Similarly, I estimate $\phi_y(t)$ by $\widehat{\phi}_y(t) = \widehat{\phi}_2(t)\phi_K(h_v t)/\{\widehat{\phi}_{u_y}(t/m_y)\}^{m_y}$. Although an estimator of the characteristic function of the common distribution $F$ can be defined in many ways, for simplicity I have decided to consider the estimator to be $\widehat{\phi}(t) = \{\widehat{\phi}_x(t) + \widehat{\phi}_y(t)\}/2$. Next using the inversion formula along with the conditions $\sup_t |\phi_K(t)/\phi_{u_x}(t/h_w)| < \infty$, $\int |\phi_K(t)/\phi_{u_x}(t/h_w)|dt < \infty$, $\sup_t |\phi_K(t)/\phi_{u_y}(t/h_v)| < \infty$ and $\int |\phi_K(t)/\phi_{u_y}(t/h_v)|dt < \infty$ for fixed $h_w, h_v > 0$ (Stefanski and Carroll, 1990), a deconvoluted density estimator can be obtained, given by:

$$
\begin{aligned}
\widehat{f}(r) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itr)\widehat{\phi}(t)dt \\
&= \frac{1}{4\pi}\int_{-\infty}^{\infty} \exp(-itr)\left[\frac{\sum_{j=1}^{n_x}\exp(it\overline{W}_j)\phi_K(h_w t)/n_x}{\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}} + \frac{\sum_{j=1}^{n_y}\exp(it\overline{V}_j)\phi_K(h_v t)/n_y}{\{\widehat{\phi}_{u_y}(t/m_y)\}^{m_y}}\right]dt \\
&= \frac{1}{2n_x}\sum_{j=1}^{n_x}\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\{-it(r-\overline{W}_j)\}\frac{\phi_K(h_w t)}{\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}}dt \\
&\quad + \frac{1}{2n_y}\sum_{j=1}^{n_y}\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\{-it(r-\overline{V}_j)\}\frac{\phi_K(h_v t)}{\{\widehat{\phi}_{u_y}(t/m_y)\}^{m_y}}dt \\
&= \frac{1}{2n_x h_w}\sum_{j=1}^{n_x}\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\{-it(r-\overline{W}_j)/h_w\}\frac{\phi_K(t)}{\{\widehat{\phi}_{u_x}(t/h_w m_x)\}^{m_x}}dt \\
&\quad + \frac{1}{2n_y h_v}\sum_{j=1}^{n_y}\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\{-it(r-\overline{V}_j)/h_v\}\frac{\phi_K(t)}{\{\widehat{\phi}_{u_x}(t/h_v m_y)\}^{m_y}}dt \\
&= \frac{1}{2n_x h_w}\sum_{j=1}^{n_x} L_x\left(\frac{r-\overline{W}_j}{h_w}\right) + \frac{1}{2n_y h_v}\sum_{j=1}^{n_y} L_y\left(\frac{r-\overline{V}_j}{h_v}\right),
\end{aligned}
$$

where

$$L_x(u) = (1/2\pi)\int_{-\infty}^{\infty}\exp(-itu)\phi_K(t)/\{\widehat{\phi}_{u_x}(t/h_w m_x)\}^{m_x}dt$$

and

$$L_y(u) = (1/2\pi)\int_{-\infty}^{\infty}\exp(-itu)\phi_K(t)/\{\widehat{\phi}_{u_y}(t/h_v m_y)\}^{m_y}dt.$$

Although the common population CDF $F$ may not have a density, this density estimator is well defined. I am using this formula only to motivate the definition of the CDF estimator given next.

Indeed, replacing $\phi_{u_x}$ by its estimator given in (2.3) and $\phi_{u_y}$ by the corresponding estimator, and replacing $\phi_K(t)$ by $(1-t^2)^3 1_{[-1,1]}(t)$, and using the integration formula (A.1) of Hall and Lahiri (2008), the estimator of the common distribution is

$$
\begin{aligned}
&\widehat{F}(r) = \\
&\frac{1}{n_x}\sum_{j=1}^{n_x}\left[\frac{1}{2}+\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\sin\{t(r-\overline{W}_j)\}}{t}\frac{(1-h_w^2t^2)^3 1_{[-1,1]}(h_wt)}{|N_x^{-1}\sum_{j=1}^{n_x}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{(t/m_x)(W_{jl_1}-W_{jl_2})\}|^{m_x/2}}dt\right] \\
&+\frac{1}{n_y}\sum_{j=1}^{n_y}\left[\frac{1}{2}+\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\sin\{t(r-\overline{V}_j)\}}{t}\frac{(1-h_v^2t^2)^3 1_{[-1,1]}(h_vt)}{|N_y^{-1}\sum_{j=1}^{n_y}\sum_{(l_1,l_2)\in\mathcal{S}_y}\cos\{(t/m_y)(V_{jl_1}-V_{jl_2})\}|^{m_y/2}}dt\right] \\
&=\frac{1}{2}+\frac{1}{2n_x\pi}\int_0^{1/h_w}\frac{(1-h_w^2t^2)^3\sum_{j=1}^{n_x}\sin\{t(r-\overline{W}_j)\}}{t|N_x^{-1}\sum_{j=1}^{n_x}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{(t/m_x)(W_{jl_1}-W_{jl_2})\}|^{m_x/2}}dt \\
&+\frac{1}{2n_y\pi}\int_0^{1/h_v}\frac{(1-h_v^2t^2)^3\sum_{j=1}^{n_y}\sin\{t(r-\overline{V}_j)\}}{t|N_y^{-1}\sum_{j=1}^{n_y}\sum_{(l_1,l_2)\in\mathcal{S}_y}\cos\{(t/m_y)(V_{jl_1}-V_{jl_2})\}|^{m_y/2}}dt.
\end{aligned}
$$

For generating random numbers we shall use a monotonized version of $\widehat{F}$, and following the general technique of Hall and Lahiri (2008), we define $\tilde{F}(r)=\sup\{\widehat{F}(r^*):r^*\le r\}$, and then estimate the quantile for a given $p\in(0,1)$ as $q=\sup\{r:\tilde{F}(r)\le p\}$.

Next I would like to point out the optimal $h_w$. I shall use Hall and Lahiri (2008)'s method that is relatively straight forward to apply. According to Theorem 4.1 of that paper, I choose the optimal $h_w$ that minimizes $n_x^{-1}I(h)+B_xh^4$, where $2\pi I(h)=\int t^{-2}[1-\phi_K(ht)/\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}]^2dt$, $B_x=\kappa_2^2/(16\sqrt{\pi}\widehat{\sigma}_x^3)$ with $\kappa_2=\int x^2K(x)dx$. For our choice of kernel, $\kappa_2=6$. Also, var($\overline{W}$) $=$ var($X$) $+$ var($U_x$)$/m_x$, so we estimate $\sigma_x^2$ by $\widehat{\sigma}_x^2=\widehat{\sigma}_{\overline{W}}^2-\widehat{\sigma}_{u_x}^2/m_x$, where $\widehat{\sigma}_{\overline{W}}^2=(n_x-1)^{-1}\sum_{j=1}^{n_x}(\overline{W}_j-\overline{W}_{..})^2$, $\widehat{\sigma}_{u_x}^2=(n_x)^{-1}\sum_{j=1}^{n_x}(m_x-1)^{-1}\sum_{l=1}^{m_x}(W_{jl}-\overline{W}_j)^2$, and $\overline{W}_{..}=(n_xm_x)^{-1}\sum_{j=1}^{n_x}\sum_{l=1}^{m_x}W_{jl}$. A numerical integration technique is applied to evaluate $I(h_w)$. Similarly, I shall determine the optimal $h_v$.

### 2.3.3 Estimation of $F_{u_x}$ and $F_{u_y}$

In this section, I shall describe the estimation of $F_{u_x}$, and the estimation of $F_{u_y}$ follows similar steps, so is omitted. Observe that $W_{jl_1}-W_{jl_2}=U_{x,jl_1}-U_{x,jl_2}$, where $U_{x,jl_1}$ and $U_{x,jl_2}$ are iid copies of the random variable $U_x$ and $(l_1,l_2)\in\mathcal{S}_x$. Hence the density of the difference of the iid copies can be estimated by the kernel method

$$\widehat{f}_{U_{x,1}-U_{x,2}}(u^*) = \frac{1}{hn_x}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x} K\{\frac{u^*-(W_{jl_1}-W_{jl_2})}{h}\},$$

where we take $h = 1.06\widehat{\sigma}_{d,u_x}n_x^{-1/5}$ (Sheather, 2004), where $\widehat{\sigma}^2_{d,u_x} = (n_x-1)^{-1}\sum_{j=1}^{n_x} 2\{m_x(m_x-1)\}^{-1}\sum_{(l_1,l_2)\in\mathcal{S}_x}[(W_{jl_1}-W_{jl_2})-n_x^{-1}\sum_{j'=1}^{n_x}2\{m_x(m_x-1)\}^{-1}\sum_{(l_1,l_2)\in\mathcal{S}_x}(W_{j'l_1}-W_{j'l_2})]^2$. Then, the characteristic function estimator of $U_{x,1}-U_{x,2}$ is given by

$$\begin{aligned}
\widehat{\phi}_{U_{x,1}-U_{x,2}}(t) &= \int \exp(itu^*)\widehat{f}_{U_{x,1}-U_{x,2}}(u^*)du^* \\
&= \int \exp(itu^*)\frac{1}{n_xh}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x} K\{\frac{u^*-(W_{jl_1}-W_{jl_2})}{h}\}du^* \\
&= \frac{1}{n_xh}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x}\int \exp(itu^*)K\{\frac{u^*-(W_{jl_1}-W_{jl_2})}{h}\}du^* \\
&= \frac{1}{n_x}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x}\int \exp[it\{(W_{jl_1}-W_{jl_2})+hz\}]K(z)dz \\
&= \frac{1}{n_x}\sum_{j=1}^{n_x}\frac{2}{m_x(m_x-1)}\sum_{(l_1,l_2)\in\mathcal{S}_x}\exp\{it(W_{jl_1}-W_{jl_2})\}\phi_K(ht).
\end{aligned}$$

Since $E[\exp\{it(U_{x,1}-U_{x,2})\}] = \{\phi_{u_x}(t)\}^2$ due to the symmetry of $U_x$, and using $\phi_K(t) = (1-t^2)^3 1_{[-1,1]}(t)$, we estimate $\phi_{u_x}(t)$ by

$$\begin{aligned}
\widehat{\phi}_{u_x}(t) &= \sqrt{\widehat{E}[\exp\{it(U_{x,1}-U_{x,2})\}]} \\
&= \sqrt{\widehat{\phi}_{U_{x,1}-U_{x,2}}(t)} \\
&= \sqrt{\left|\sum_{j=1}^{n_x}\sum_{(l_1,l_2)\in\mathcal{S}_x}\frac{2\cos\{t(W_{jl_1}-W_{jl_2})\}}{n_xm_x(m_x-1)}(1-h^2t^2)^3 1_{[-1,1]}(ht)\right|}. \qquad (2.7)
\end{aligned}$$

Due to the presence of the indicator function, $\int |\widehat{\phi}_{u_x}(t)|dt < \infty$, and this integrability is a sufficient condition for the following inversion. Hence, we estimate $F_{u_x}(u)$ by

$$\begin{aligned}
\widehat{F}_{u_x}(u) &= \frac{1}{2} + \frac{1}{2\pi}\int_0^{\infty} \frac{\exp(itu)\widehat{\phi}_{u_x}(-t) - \exp(-itu)\widehat{\phi}_{u_x}(t)}{it} dt \\
&= \frac{1}{2} + \frac{1}{2\pi}\int_0^{\infty} \frac{\widehat{\phi}_{u_x}(t)\{\exp(itu) - \exp(-itu)\}}{it} dt \\
&= \frac{1}{2} + \frac{1}{\pi}\int_0^{\infty} \widehat{\phi}_{u_x}(t)\frac{\sin(tu)}{t} dt \\
&= \frac{1}{2} + \frac{1}{\pi}\int_0^{\infty} \frac{\sin(tu)}{t}\sqrt{\left|\sum_{j=1}^{n_x}\sum_{(l_1,l_2)\in\mathcal{S}_x} \frac{2\cos\{t(W_{jl_1} - W_{jl_2})\}}{n_x m_x(m_x - 1)}(1 - h^2t^2)^3 1_{[-1,1]}(ht)\right|} dt \\
&= \frac{1}{2} + \frac{1}{\pi}\int_0^{1/h} \frac{\sin(tu)}{t}\sqrt{\left|\sum_{j=1}^{n_x}\sum_{(l_1,l_2)\in\mathcal{S}_x} \frac{2\cos\{t(W_{jl_1} - W_{jl_2})\}}{n_x m_x(m_x - 1)}\right|}(1 - h^2t^2)^{3/2} dt.
\end{aligned}$$

I shall evaluate this integration by the Gauss-Legendre quadrature formula. As before, for simulating random numbers from this distribution we define the $p^{th}$ $(0 < p < 1)$ percentile as $q = \sup\{r : \tilde{F}_{u_x}(r) \leq p\}$, where $\tilde{F}_{u_x}(r) \equiv \sup\{\widehat{F}_{u_x}(r^*) : r^* \leq r\}$.

### 2.3.4 Validity of the Bootstrap

I now show that under some regularity conditions, the proposed Bootstrap method produces valid approximation to the distribution of the test statistic under the null. The Bootstrap probability is denoted by $P_*$.

**Theorem 3.** *Suppose that $H_0 : F_x = F_y$ holds and as $n_x, n_y \to \infty$, $\sqrt{n_x/n_y} \to \rho \in (0, \infty)$. Also suppose that the bandwidths $h_w > 0$ and $h_v > 0$ are such that $[\{h_w + (n_x h_w)^{-1}\} + \{h_v + (n_x h_v)^{-1}\}] \to 0$. Then,*

$$\lim_{n_x\to\infty}\sup_{t\geq 0}\left|P(T_{n_x} \leq t) - P_*(T^*_{n_x} \leq t)\right| = 0, \textit{ almost surely.}$$

Next, for $\alpha \in (0, 1)$, let $\widehat{t}_{n_x,\alpha}$ denote the $(1-\alpha)$- quantile the Bootstrapped statistic $T^*_{n_x}$. Then, an immediate consequence of this result is that for any $\alpha \in (0, 1)$, $\widehat{t}_{n_x,\alpha} - t_{n_x,\alpha} \to 0$ almost surely.

As a consequence, under the conditions of Theorem 3,

$$\text{pr}(T_{n_x} > \hat{t}_\alpha) \to \alpha.$$

Thus, the Bootstrap method provides a valid method for calibrating the test statistic without having to estimate the covariance structure of the limit distribution of $T_{n_x}$. Finite sample properties of the Bootstrap approximation are presented in the next section.

**Remark 1.** *It may be noted that the formula for $\widehat{F}$ in Section 2.3.2 implicitly assumes that the median of $F(\cdot)$ is zero, i.e., the median of $F_x$ and $F_y$ are zero. However, this does not pose any problem for Bootstrapping the null distribution of the test statistic $T_{n_x}$. To appreciate why, note that $H_0 : F_x = F_y$ is equivalent to $H_0' : F_{x,a} = F_{y,a}$ for any $a \in \mathbb{R}$, where $F_{x,a}(t) = F_x(t+a)$ and $F_{y,a}(t) = F_y(t+a)$, $t \in \mathbb{R}$. Thus, if necessary, by subtracting a common constant $a \in \mathbb{R}$, we can, without loss of generality, assume that under the null hypothesis, the medians of $F_x$ and $F_y$ are zero. Indeed, noting that the test statistic $T_{n_x}$ can be written as $T_{n_x} = \int |\widehat{\phi}_x(t) - \widehat{\phi}_y(t)|^2 \omega(t) dt$, it follows that $T_{n_x}$ is invariant under a common location change. As a result, one gets a valid approximation to the null distribution of $T_{n_x}$ by using the estimator $\widehat{F}$ in Section 2.3.2 even when the median of the common distribution $F$ is different from zero. This observation also highlights the challenges and complexities associated with formulation of a valid Bootstrap method in the two sample testing problem in presence of measurement error.*

### 2.4 Simulation studies

**Simulation designs:** In this section, I present the numerical performance of the proposed test via Monte-Carlo simulations. I simulated datasets that consisted of two samples, $\{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{n_x}\}$ and $\{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{n_y}\}$, where $\boldsymbol{W}_j = (W_{j1}, \ldots, W_{jm_x})^T$ and $\boldsymbol{V}_k = (V_{k1}, \ldots, V_{km_y})^T$. I considered $n_x = n_y = 50, 200$ and $500$ while two different scenarios corresponding to the number of repetitions are considered: 1) $m_x = m_y = 2$ and 2) $m_x = 2, m_y = 3$. Type I error rate was examined in the following four designs (D1, D2, D3, D4), while power of the test was examined in designs D5, D6, D7, and D8. In addition, D9 and D10 were designed to explore robustness of the proposed method towards the symmetric measurement error assumption.

D1 $X, Y \sim \text{Normal}(0, 1)$ and $U_x, U_y \sim DE(0, 0.35)$

D2 $X, Y \sim \text{Normal}(0, 1)$ and $U_x, U_y \sim N(0, 0.5^2)$

D3 $X, Y \sim \text{Normal}(0, 1)$ and $U_x \sim DE(0, 0.35), U_y \sim N(0, 0.5^2)$

D4 $X, Y \sim (\chi^2_1 - 1)/\sqrt{2}$ and $U_x \sim DE(0, 0.35), U_y \sim DE(0, 0.2)$

D5 $X \sim \text{Normal}(0, 1), Y \sim \text{Normal}(0.2, 1)$ and $U_x, U_y \sim DE(0, 0.35)$

D6 $X \sim \text{Normal}(0, 1), Y \sim DE(0, 0.7)$ and $U_x, U_y \sim DE(0, 0.35)$

D7 $X \sim \text{Normal}(0, 1), Y \sim DE(0, 0.7)$ and $U_x \sim DE(0, 0.35), U_y \sim N(0, 0.5^2)$

D8 $X \sim 0.5\text{Normal}(-0.9, 0.45^2) + 0.5\text{Normal}(0.9, 0.45^2), Y \sim \text{Normal}(0, 1)$ and $U_x, U_y \sim DE(0, 0.35)$

D9 $X, Y \sim \text{Normal}(0, 1)$ and $U_x, U_y \sim EXP(0.5) - 0.5$

D10 $X \sim \text{Normal}(0, 1), Y \sim DE(0, 0.7)$ and $U_x, U_y \sim EXP(0.5) - 0.5$

Here $DE(a, b)$ stands for the double exponential distribution with mean $a$ and variance $2b^2$ and $EXP(a)$ denotes the exponential distribution with mean $a$. In the first three designs, both measurement error variances associated with $X$ and $Y$ are 25% of the variability of $X$ (or $Y$). In D4, both $X$ and $Y$ follow the modified chi-square distribution with degrees of freedom $1$, mean $0$ and variance $1$. The choice of the true signals (the distribution of $X$ or $Y$) and the measurement error variance were somewhat similar to that of Delaigle et al. (2008). In D4, measurement error variances corresponding to $X$ and $Y$ are different, and consequently the variances of the convoluted observations are different, i.e., $\text{var}(W_{jl}) \neq \text{var}(V_{kl^*})$. The designs are also different in terms of the smoothness of their measurement error distributions, I considered the ordinary smooth class (D1, D4, D6, D8), the supersmooth class (D2), the mixed case (D3, D5, D7). For the alternative hypotheses, I included cases where there are differences in the location (D5) and in the shape (D6, D7, D8). In D9 and D10, we considered centered exponential distribution for the measurement error with variability 25% of that of the true signal.

**Method of analysis:** For each dataset, I carried out hypothesis test at the $5\%$ level of significance. For the proposed method I rejected the null hypothesis $H_0 : F_x(r) = F_y(r)$ against $H_a : F_x(r) \neq F_y(r)$ if the $p$-value calculated using $B = 1,000$ Bootstrap samples was less than $\alpha = 0.05$. I also analyzed each data set using the naive testing methods that included the Kolomogorov-Smirnov test (K-S) and the two sample Anderson-Darling test (A-D) based on the averages $\{\overline{W}_j, j = 1, \ldots, n_x\}$ and $\{\overline{V}_k, k = 1, \ldots, n_y\}$. In these naive tests, $\{\overline{W}_j, j = 1, \ldots, n_x\}$ and $\{\overline{V}_k, k = 1, \ldots, n_y\}$ are considered as random samples from $F_x$ and $F_y$, respectively.

Regarding the choice of $\omega(t)$, it is worth highlighting the desirable properties of $\omega(t)$ advocated by Epps and Pulley (1983). First, $\omega(t)$ should have more weight where the underlying difference between the two characteristic functions is large, and that difference is usually large in an interval near zero. Second, the weight $\omega(t)$ should be large where the estimators $\widehat{a}_x(t) - \widehat{a}_y(t)$ and $\widehat{b}_x(t) - \widehat{b}_y(t)$ are highly precise. In fact, the precision decreases as $t$ moves away from zero. Furthermore, for the ordinary smooth and supersmooth class of measurement error distributions (Fan, 1991), the characteristics functions are polynomially and exponentially decreasing, respectively. Consequently, for a small $\varepsilon > 0$, $|\widehat{\phi}_{u_x}(t)| \leq \varepsilon$ whenever $|t| \geq t^*$ for some $t^* > 0$, that in turn results in highly variable estimators $\widehat{a}_x(t)$, $\widehat{a}_y(t)$, $\widehat{b}_x(t)$, $\widehat{b}_y(t)$ when $|t| > t^*$. Based on these considerations, for the proposed approach, I used different weights, the normal weight $\omega(t) = \exp(-t^2/2)I(t_1 < t < t_2)$ and the uniform weight $\omega(t) = I(t_1 < t < t_2)$. For each weight, I considered two sets of $(t_1, t_2)$. In the first set I took $t_1 = \min(F_x^{-1}(0.005), F_y^{-1}(0.005))$ and $t_2 = \max(F_x^{-1}(0.995), F_y^{-1}(0.995))$, and the corresponding weights are referred to as $\mathrm{norm}_{0.99}$ and $\mathrm{unif}_{0.99}$ for the normal and uniform weight, respectively. In the second set I took $t_1 = \min(F_x^{-1}(0.1), F_y^{-1}(0.1))$ and $t_2 = \max(F_x^{-1}(0.9), F_y^{-1}(0.9))$, and the corresponding weights are referred to as $\mathrm{norm}_{0.8}$ and $\mathrm{unif}_{0.8}$. Results for these four different weights show how the performance of the test depends on the weight function.

**Results:** For each scenario I simulated 5,000 datasets, and for each scenario I computed the power of each test. The power represents the proportion of times rejecting $H_0$ at the 5% level out of $5,000$ replications. Tables 3.1 and 2.2 contain the simulation results for 1) $m_x = m_y$ and 2)

Table 2.1: The entries of the table show the proportion of the rejection of $H_0$ at the $5\%$ level for the simulation study with sample sizes $n_x = n_y = n$ and $m_x = m_y = 2$ based on $5,000$ replications. Here K-S, A-D, and C-F refer to the Kolmogorov-Smirnov, Anderson-Darling, and the proposed characteristic function based test, respectively. The entries corresponding to designs D1-D4 show the Type-I error rate, and the other entries are power.

| | | | | C-F | | | |
|---|---|---|---|---|---|---|---|
| | $n$ | K-S | A-D | $\text{unif}_{0.99}$ | $\text{unif}_{0.8}$ | $\text{norm}_{0.99}$ | $\text{norm}_{0.8}$ |
| D1 | 50 | 0.038 | 0.044 | 0.033 | 0.036 | 0.038 | 0.037 |
| | 200 | 0.039 | 0.050 | 0.038 | 0.043 | 0.044 | 0.044 |
| | 500 | 0.052 | 0.052 | 0.047 | 0.050 | 0.048 | 0.048 |
| D2 | 50 | 0.039 | 0.050 | 0.029 | 0.036 | 0.036 | 0.037 |
| | 200 | 0.038 | 0.052 | 0.039 | 0.043 | 0.041 | 0.044 |
| | 500 | 0.049 | 0.051 | 0.037 | 0.043 | 0.041 | 0.044 |
| D3 | 50 | 0.038 | 0.049 | 0.033 | 0.042 | 0.037 | 0.042 |
| | 200 | 0.039 | 0.053 | 0.039 | 0.047 | 0.043 | 0.047 |
| | 500 | 0.055 | 0.051 | 0.045 | 0.046 | 0.045 | 0.045 |
| D4 | 50 | 0.041 | 0.056 | 0.010 | 0.038 | 0.035 | 0.037 |
| | 200 | 0.052 | 0.083 | 0.033 | 0.036 | 0.037 | 0.036 |
| | 500 | 0.120 | 0.198 | 0.040 | 0.036 | 0.039 | 0.035 |
| D5 | 50 | 0.099 | 0.137 | 0.053 | 0.108 | 0.092 | 0.115 |
| | 200 | 0.327 | 0.443 | 0.156 | 0.369 | 0.322 | 0.393 |
| | 500 | 0.728 | 0.820 | 0.401 | 0.749 | 0.695 | 0.775 |
| D6 | 50 | 0.063 | 0.069 | 0.095 | 0.053 | 0.085 | 0.051 |
| | 200 | 0.147 | 0.154 | 0.439 | 0.110 | 0.315 | 0.095 |
| | 500 | 0.423 | 0.470 | 0.863 | 0.275 | 0.745 | 0.222 |
| D7 | 50 | 0.060 | 0.069 | 0.082 | 0.054 | 0.085 | 0.051 |
| | 200 | 0.133 | 0.149 | 0.403 | 0.107 | 0.300 | 0.089 |
| | 500 | 0.391 | 0.425 | 0.845 | 0.262 | 0.738 | 0.207 |
| D8 | 50 | 0.097 | 0.084 | 0.218 | 0.054 | 0.101 | 0.053 |
| | 200 | 0.326 | 0.251 | 0.812 | 0.066 | 0.392 | 0.061 |
| | 500 | 0.828 | 0.832 | 0.997 | 0.132 | 0.896 | 0.104 |
| D9 | 50 | 0.042 | 0.047 | 0.038 | 0.044 | 0.039 | 0.042 |
| | 200 | 0.037 | 0.050 | 0.042 | 0.045 | 0.047 | 0.045 |
| | 500 | 0.046 | 0.046 | 0.047 | 0.045 | 0.047 | 0.045 |
| D10 | 50 | 0.062 | 0.069 | 0.109 | 0.052 | 0.092 | 0.046 |
| | 200 | 0.150 | 0.159 | 0.450 | 0.121 | 0.319 | 0.099 |
| | 500 | 0.443 | 0.472 | 0.875 | 0.280 | 0.752 | 0.217 |

Table 2.2: The entries of the table show the proportion of the rejection of $H_0$ at the $5\%$ level for the simulation study with sample sizes $n_x = n_y = n$ and $m_x = 2, m_y = 3$ based on $5,000$ replications. Here K-S, A-D, and C-F refer to the Kolmogorov-Smirnov, Anderson-Darling, and the proposed characteristic function based test, respectively. The entries corresponding to designs D1-D4 show the Type-I error rate, and the other entries are power.

| | | | | C-F | | | |
|---|---|---|---|---|---|---|---|
| | $n$ | K-S | A-D | $\text{unif}_{0.99}$ | $\text{unif}_{0.8}$ | $\text{norm}_{0.99}$ | $\text{norm}_{0.8}$ |
| D1 | 50 | 0.038 | 0.043 | 0.037 | 0.038 | 0.038 | 0.038 |
| | 200 | 0.039 | 0.048 | 0.041 | 0.047 | 0.047 | 0.045 |
| | 500 | 0.055 | 0.056 | 0.046 | 0.051 | 0.051 | 0.052 |
| D2 | 50 | 0.039 | 0.050 | 0.032 | 0.038 | 0.037 | 0.038 |
| | 200 | 0.046 | 0.052 | 0.039 | 0.043 | 0.042 | 0.044 |
| | 500 | 0.050 | 0.053 | 0.038 | 0.046 | 0.041 | 0.047 |
| D3 | 50 | 0.036 | 0.049 | 0.036 | 0.038 | 0.039 | 0.039 |
| | 200 | 0.038 | 0.054 | 0.042 | 0.046 | 0.046 | 0.048 |
| | 500 | 0.049 | 0.053 | 0.044 | 0.048 | 0.047 | 0.048 |
| D4 | 50 | 0.049 | 0.066 | 0.009 | 0.039 | 0.038 | 0.040 |
| | 200 | 0.082 | 0.162 | 0.026 | 0.034 | 0.035 | 0.035 |
| | 500 | 0.287 | 0.551 | 0.036 | 0.034 | 0.034 | 0.034 |
| D5 | 50 | 0.105 | 0.141 | 0.059 | 0.112 | 0.100 | 0.120 |
| | 200 | 0.342 | 0.457 | 0.179 | 0.384 | 0.339 | 0.409 |
| | 500 | 0.746 | 0.828 | 0.434 | 0.764 | 0.718 | 0.786 |
| D6 | 50 | 0.071 | 0.075 | 0.116 | 0.055 | 0.095 | 0.051 |
| | 200 | 0.194 | 0.210 | 0.483 | 0.113 | 0.328 | 0.099 |
| | 500 | 0.583 | 0.647 | 0.900 | 0.285 | 0.778 | 0.230 |
| D7 | 50 | 0.070 | 0.075 | 0.112 | 0.059 | 0.089 | 0.056 |
| | 200 | 0.190 | 0.210 | 0.475 | 0.113 | 0.318 | 0.094 |
| | 500 | 0.558 | 0.621 | 0.890 | 0.282 | 0.769 | 0.219 |
| D8 | 50 | 0.104 | 0.088 | 0.243 | 0.057 | 0.104 | 0.056 |
| | 200 | 0.364 | 0.312 | 0.837 | 0.073 | 0.420 | 0.064 |
| | 500 | 0.869 | 0.873 | 0.998 | 0.131 | 0.904 | 0.099 |
| D9 | 50 | 0.042 | 0.046 | 0.041 | 0.041 | 0.039 | 0.042 |
| | 200 | 0.038 | 0.050 | 0.046 | 0.047 | 0.047 | 0.046 |
| | 500 | 0.046 | 0.050 | 0.051 | 0.046 | 0.045 | 0.045 |
| D10 | 50 | 0.067 | 0.076 | 0.130 | 0.052 | 0.094 | 0.048 |
| | 200 | 0.195 | 0.215 | 0.497 | 0.123 | 0.343 | 0.101 |
| | 500 | 0.575 | 0.636 | 0.906 | 0.288 | 0.774 | 0.226 |

$m_x \neq m_y$ cases, respectively. The results indicate that the proposed test maintains the nominal level for all designs (D1 - D4) and for different weights. For D4, the naive tests fail to maintain the nominal level, and their power seems to be increasing with the sample size for both cases, 1) $m_x = m_y$ and 2) $m_x \neq m_y$. The intuitive reason is that although the means are the same $E(\overline{W}) = E(\overline{V})$, the variances are different, var$(\overline{W}) = 1 + 0.25/m_x$ and var$(\overline{V}) = 1 + 0.02/m_y$. Therefore, K-S or A-D test based on the empirical distributions of $(\overline{W}_1, \ldots, \overline{W}_{n_x})$ and $(\overline{V}_1, \ldots, \overline{V}_{n_y})$ are likely to reject $H_0$. For the scenarios D1-D3 when $m_x = 2$ and $m_y = 3$, although the type-I error rate of the K-S and A-D seems to be under the nominal level, a further simulation with $n_x = n_y = 2000$ revealed that the type-I error rate is exceeds the nominal level as powers for K-S (A-D) test are 0.0544 (0.0572), 0.0542 (0.061), and 0.054 (0.061) for designs D1, D2, and D3, respectively.

For the cases, where the alternative hypothesis holds, the power of the proposed test is increasing with the sample size. For D5, where the distribution of $X$ and $Y$ differ only by a location parameter, the power of the proposed test is somewhat lower than that of the naive approaches. Here is an intuitive explanation. Since the difference in the location parameters for the $X$ and $Y$ distributions is well reflected in the difference between the CDFs of $\overline{W}$ and $\overline{V}$ when the distribution of $F_{u_x}$ and $F_{u_y}$ are the same, the naive methods are capable of differentiating the two underlying distributions. Although the proposed method detects the difference between $F_x$ and $F_y$ in terms of the location parameter, the actual difference is somewhat masked out by the variability of the estimator of the characteristic functions of the true signal and the measurement error. For scenarios D6, D7, and D8, the power of the proposed approach is significantly better than the other methods, even for sample size $n = 50$. In D6, D7, and D8, the mean and variance of the convoluted observations from the two samples are almost the same, $E(W_{jl}) = E(V_{kl^*}) = 0$ and var$(W_{jl}) \approx$ var$(V_{kl^*})$, and also the first two moments of $\overline{W}_j$ are the same as that of $\overline{V}_k$, i.e., $E(\overline{W}_j) = E(\overline{V}_k) = 0$ and var$(\overline{W}_j) \approx$ var$(\overline{V}_k)$ for $m_x = m_y$ case. Additionally, the shapes of the distribution of $\overline{W}_j$ and $\overline{V}_k$ are not dramatically different, especially for $m_x = m_y$ case. Therefore, the power of the K-S or A-D is lower than that of the proposed method. Naturally the power of the naive approaches improve from $m_x = m_y = 2$ to $m_x = 2, m_y = 3$ scenario as the variance of $\overline{W}$ and $\overline{V}$ become different due to different replications. For the asymmetric measurement error model (D9 and D10), the proposed test maintains the level and gives better power ($\mathrm{unif}_{0.99}$ and $\mathrm{norm}_{0.99}$) than the K-S or A-D tests. These results indicate that the proposed test is quite robust towards the violation of the symmetric error assumption.

In summary, the simulation results indicate that the proposed test is consistent, while the naive tests could be inconsistent. In the absence of any specific knowledge about the characteristic function of the underlying distributions, in our opinion, the $\mathrm{unif}_{0.99}$ weight is preferable as it covers a wide range of $t$-values and gives equal importance to the difference between the two characteristic functions at any $t$.

### 2.5 Numerical study using the NHANES data

I shall apply the proposed method to analyze the NHANES data that are publicly available at `https://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm`. In real data, there could be a number of covariates that may affect the variable of interest. Thus it is important to eliminate the confounding effect, and one general approach is to regress out these covariates. Then the residual can be considered to be independent of the confounding variables, and we may apply the testing procedure on the residuals. To be specific let me discuss our first data example.

**Blood pressure example:** For the illustration purpose I consider the NHANES 2009-2010 survey data, and focus only non-Hispanic white males whose ages are between 35 and 55 years (middle-aged adults) so that we have a more or less homogeneous group with a lesser extent of the confounding issue. The goal is to test equality of the distribution of systolic blood pressure between two groups, non-alcoholic and alcoholic. Alcohol consumption data are collected through two 24-hour recall interviews. Define a subject as non-alcoholic if both measurements are less than 14 grams, otherwise the subject is considered to be alcoholic. Since fourteen grams is considered to be the amount of alcohol in a standard drink, I use this value to define the two behavioral groups. This classification results in $n_x = 207$ (non-alcoholic) and $n_y = 126$ (alcoholic). Since an accurate measurement of blood pressure is difficult to obtain, at least three measurements were taken in the mobile examination center. For the analysis I consider the first three measurements for each subjects, i.e., $m_x = m_y = 3$. Suppose that $A_{jk} = \log(\text{systolic blood pressure}_{jk})$ denote the logarithm of the $k^{th}$ blood pressure measurement of the $j$th individual, $k = 1, 2, 3$, and $j = 1, \ldots, (n_x + n_y)$.

The dataset contains potential confounding variables such as body mass index (BMI), a continuous variable, and income, an ordinal categorical variable. Suppose that $\boldsymbol{Z}_j$ and $A_j$ denote the set of confounding variables, and the logarithm of the true systolic blood pressure of subject $j$.

Next, assume that $A_j = \beta_0 + \boldsymbol{Z}_j^T \boldsymbol{\beta}_1 + \varepsilon_j$ for some $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$, and observed data are $A_{jk} = A_j +$ measurement error, for $k = 1, \ldots, m$. Note that we assume one common $\boldsymbol{\beta}$ regardless of the group (non-alcoholic or alcoholic), otherwise $\varepsilon_j$ does not have the group information. To estimate $\boldsymbol{\beta}$ we simply regress $\overline{A}_j = \sum_k A_{jk}/m$ on $\boldsymbol{Z}_j$. Then define

$$W_{jl} = A_{jl} - \widehat{\beta}_0 - \boldsymbol{Z}_j^T \widehat{\boldsymbol{\beta}}_1 = X_j + U_{x,jl},\ l = 1, \ldots, m_x,\ j = 1, \ldots, n_x,$$

and

$$V_{kl} = A_{kl} - \widehat{\beta}_0 - \boldsymbol{Z}_k^T \widehat{\boldsymbol{\beta}}_1 = Y_k + U_{y,kl},\ l = 1, \ldots, m_y,\ k = n_x + 1, \ldots, n_y.$$

Particularly, for this example, $W_{jl} = A_{jl} - \widehat{\beta}_0 - \widehat{\beta}_1 \text{BMI}_j - \widehat{\beta}_2 \text{income}_j$, $j = 1, \ldots, n_x$, and $V_{kl} = A_{(n_x+k)l} - \widehat{\beta}_0 - \widehat{\beta}_1 \text{BMI}_{n_x+k} - \widehat{\beta}_2 \text{income}_{n_x+k}$, for $k = 1, \ldots, n_y$. Next, I shall apply the proposed test and the two naive tests on those residuals. As discussed in the simulation study, the $\text{unif}_{0.99}$ weight function is considered. However, to calculate $t_1$ and $t_2$ I use the deconvoluted distribution functions $\widehat{F}_x$ and $\widehat{F}_y$ instead of $F_x$ and $F_y$ as the later two are unknown in the real data.

The resulting $p$-values are given in the first row of Table 2.3. At the $5\%$ level, the proposed method strongly rejects $H_0$ while the naive approaches contradict each other so that it is difficult to make a decision. For this and the next application, I use $10,000$ Bootstrap samples and the $\text{unif}_{0.99}$ weight function to calculate the $p$-value for our proposed method. The conclusion based on the proposed test affirms the medical science that usually alcohol consumption and high blood pressure are

Table 2.3: The table shows the $p$-values for testing of hypothesis using the real data. Here K-S, A-D, and C-F refer to the Kolmogorov-Smirnov, Anderson-Darling, and the proposed characteristic function based test, respectively. Also, ACR≡ Albumin-to-creatinine ratio, and BP≡ Systolic blood pressure. For the proposed method, the number of Bootstrap samples was $10,000$.

| Variable | K-S | A-D | C-F |
|---|---|---|---|
| BP | 0.058 | 0.001 | 0.001 |
| ACR | 0.043 | 0.143 | 0.008 |

associated. Moreover, repeated binge drinking for a long time may cause elevated blood pressure (`http://www.mayoclinic.org/diseases-conditions/high-blood-pressure/expert-answers/blood-pressure/faq-20058254`).

**Albumin-to-creatinine ratio (ACR) example:** In this application we check if the distribution of albumin-to-creatinine ratio (ACR) differs by smoking status. Albumin is a protein and creatinine is a chemical waste, and their ratio ACR is used to assess renal functionality. Usually higher level of ACR is associated with a higher risk of renal events. Our interest is in testing equality of the distribution of ACR among non-smoking and smoking group after adjusting the effect of the confounding variables.

In the NHANES study (2009-2010 survey data), urinary albumin and creatinine were measured twice for each participants, the first sample was collected in the mobile examination center (MEC) and the second sample was collected during the interview at home. We consider these two measurements (samples) as the two noisy measurements of the same underlying truth, and hence $m_x = m_y = m = 2$.

For this test I consider only non-Hispanic white males who are older than 60 years as the renal issue is more prevalent in the older group. I define a person as a non-smoker if he smoked less than 100 cigarettes in his lifetime, otherwise the person is called a smoker, and based on this classification we obtain $n_x = 161$ (non-smoking) and $n_y = 290$ (smoking). For the $j$th individual define $A_{jk} = \log(\text{albumin}_{jk}/\text{creatinine}_{jk})$ for $k = 1, 2$, and $j = 1, \ldots, (n_x + n_y)$. As in the previous application, to remove the effect of BMI and income, I regress $\overline{A}_j = \sum_k A_{jk}/m$ on

BMI and income and obtain $\widehat{\boldsymbol{\beta}}$. Next define the residuals $W_{jl} = A_{jl} - \widehat{\beta}_0 - \widehat{\beta}_1 \text{BMI}_j - \widehat{\beta}_2 \text{income}_j$, $j = 1, \ldots, n_x$, and $V_{kl} = A_{(n_x+k)l} - \widehat{\beta}_0 - \widehat{\beta}_1 \text{BMI}_{n_x+k} - \widehat{\beta}_2 \text{income}_{n_x+k}$, for $k = 1, \ldots, n_y$, where $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$ are the estimated regression coefficients. Now, I apply the proposed test and the two naive tests on the residuals. As in the previous application, for the proposed test I consider the $\text{unif}_{0.99}$ weight function, where $t_1$ and $t_2$ are calculated from the deconvoluted distribution functions $\widehat{F}_x$ and $\widehat{F}_y$.

The resulting $p$-values are given in the second row of Table 2.3. For the proposed test, we get $p$-value 0.008 so that we conclude that smoking status and ACR are related. At the $5\%$ level, the A-D fails to reject $H_0$ while the $p$-value for the K-S test is barely below the nominal level. Therefore, as a whole the naive test could be misleading. The test result based on the proposed method is consistent with the finding of Hogan et al. (2007) who considered a similar issue with different smoking groups and have used the data from the NHANES III survey (1988-1994).

Table 2.4: The entries of the table show the proportion of the rejection of $H_0$ at the $5\%$ level for the simulation study where simulated datasets mimicked the blood pressure dataset ($m_x = m_y = 3$) given in Section 2.5. Here K-S, A-D, and C-F refer to the Kolmogorov-Smirnov, Anderson-Darling, and the proposed characteristic function based test, respectively.

| | | $n_x = 200, n_y = 120$ | | | $n_x = 400, n_y = 240$ | | |
|---|---|---|---|---|---|---|---|
| | | K-S | A-D | C-F | K-S | A-D | C-F |
| Type-I error rate | D3 | 0.039 | 0.049 | 0.046 | 0.043 | 0.051 | 0.048 |
| | D4 | 0.050 | 0.060 | 0.039 | 0.064 | 0.083 | 0.043 |
| | D11 | 0.053 | 0.060 | 0.038 | 0.065 | 0.073 | 0.046 |
| Power | D6 | 0.125 | 0.121 | 0.398 | 0.261 | 0.273 | 0.716 |
| | D8 | 0.282 | 0.220 | 0.716 | 0.609 | 0.587 | 0.969 |
| | D12 | 0.575 | 0.814 | 0.821 | 0.882 | 0.986 | 0.984 |

**Simulation study that mimics the NHANES data:** To show the effectiveness of the confounding variable adjustment method, mimicking the real dataset on the systolic blood pressure example, another simulation study is conducted. I generated two covariates $T_1$ and $T_2$ by mimicking the distributions of BMI and income. Specifically, $T_1$ was generated from the Gamma distribution with shape 26.7 and rate 0.9, $T_2$ was generated from the multinomial distribution with the cell probability same as the observed relative frequency from the data. Next, I defined $B_j^X = \widehat{\beta}_0 + \widehat{\beta}_1 T_{1j} + \widehat{\beta}_2 T_{2j} + X_j$, $B_k^Y = \widehat{\beta}_0 + \widehat{\beta}_1 T_{1k} + \widehat{\beta}_2 T_{2k} + Y_k$, $B_{jl} = B_j^X + U_{x,jl}$ and $B_{kl} = B_k^Y + U_{y,kl}$, for $n_x = 200$, $n_y = 120$ and $n_x = 400$, $n_y = 240$, where $X_j, Y_k, U_{x,jl}$ and $U_{y,kl}$ were specified by some designs given in Section 2.4. Here $\widehat{\boldsymbol{\beta}}$ denotes the estimated $\boldsymbol{\beta}$ in the first data example. For checking the type-I error rate, I considered designs D3 and D4, and a new design

D11 $X_j, Y_k \sim \widehat{F}, U_{x,jl} \sim \widehat{F}_{u_x}, U_{y,kl} \sim \widehat{F}_{u_y}$,

where $\widehat{F}$ is the estimator of the common distribution of $X$ and $Y$ in the first data example, and $\widehat{F}_{u_x}$ and $\widehat{F}_{u_y}$are the corresponding estimator of the measurement error distributions. For checking power, I considered designs D6, D8, and a new design

D12 $X_j \sim \widehat{F}_x, Y_k \sim \widehat{F}_y, U_{x,jl} \sim \widehat{F}_{u_x}, U_{y,kl} \sim \widehat{F}_{u_y}$,

where $\widehat{F}_x$ and $\widehat{F}_y$ are the deconvolution estimator of $X$ and $Y$, respectively, for the first data example. Each dataset was analyzed using the adjustment approach described in the first data example. Table 2.4 contains this simulation results. I find the patterns are similar to those in Tables 3.1 and 2.2. One remarkable result in this simulation is that naive approaches cannot control the nominal level even when $X, Y \sim \widehat{F}$ as in the case D4. Overall the proposed method shows consistent behavior, and much superior performance than the other approaches.

## 2.6 Conclusions

In this chapter, I have investigated the test of homogeneity of two distributions when observed data are contaminated with the classical measurement error. To extract the true signals from the error contaminated data I have applied a non-parametric method that does not make any assumption regarding the true signal. Also, other than symmetry and non-vanishing characteristic function over the entire real line, no other assumption was used for the measurement errors. A valid Bootstrap approach to calculate the $p$-value of the test has been proposed.

The benefit of the proposed approach is shown through simulation studies. The simulation studies also show that the power of the proposed approach changes with the weight function. I have applied the proposed method to analyze two real datasets obtained from the NHANES 2009-2010 study. Since this data was collected from the nationally representative sample, the results of the data analysis is applicable to a broader section of the population. All computations were done using `R`.

Finally, the proposed method can be extended to the scenario where the number of replications ($m_x$ or $m_y$) is varying by subjects. Also, any further research in this area can focus on relaxing the real valued and non-zero characteristic function assumption on the measurement error distribution.

# 3. A BAYESIAN GOODNESS-OF-FIT TEST OF TECHNICAL INEFFICIENCY IN STOCHASTIC FRONTIER ANALYSIS

## 3.1 Background and literature review

Stochastic frontier (SF) models are used to assess deviations of the observed production from the optimal production. That deviation, commonly known as technical inefficiency or simply inefficiency, may arise due to technological drawback, or lack of proper allocation of resources to the production process. Typically, after fitting a SF model to a dataset, one intends to predict the inefficiency of a production unit that may refer to a farm or a geographical region. The SF model for the $i$th production unit is (Aigner et al., 1977; Meeusen and van den Broeck, 1977)

$$y_i = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}_i + \epsilon_i, \;\; \epsilon_i = v_i - u_i, \tag{3.1}$$

where $y_i$ is an output, $\boldsymbol{x}_i$ is a vector of input covariates, $v_i$ is a stochastic noise, and $u_i \geq 0$ is called technical inefficiency for the $i$th unit. Here $\epsilon$ is termed as the composite error. I assume that $v$ and $u$ are independent and they are independent of $\boldsymbol{x}$. The stochastic noise $v$ is assumed to have zero mean normal distribution with variance $\sigma_v^2$ in the literature (Chen and Wang, 2012). Commonly a half-normal distribution, exponential distribution, or a Gamma distribution is assumed for the distribution of the inefficiency term. A good review of the current state-of-the-art methods on stochastic frontier models can be found in Kumbhakar and Lovell (2003).

For consistent estimation of the model parameters, and *good* prediction of the inefficiency of production units, one needs to correctly specify the distribution of $u$. Schmidt and Lin (1984), Kopp and Mullahy (1990), and Coelli (1995) developed tests for the presence of technical inefficiency. However, these tests fail to provide any information on the underlying distributions of $v$ and $u$. For this purpose, Chen and Wang (2012) proposed consistent estimators of the parameters involving the distribution of $v$ and $u$ using the generalized moment method of the centered residuals calculated from the least square estimators for $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ . Then under several moment conditions, they developed a test for the distributional assumption on $u$ and $v$ together. On the other hand, Wang et al. (2011) proposed chi-squared tests and the Kolmogorov-Smirnov (KS) test for testing the distributional assumption of $u$ assuming that the distribution $v$ is known. They concluded that the KS test performs the best among the other approaches they considered.

Like Wang et al. (2011), in this paper we assume that the distribution of $v$ is a mean zero normal distribution, and develop a test for testing the distributional assumption of $u$ only. I develop a Bayesian test using the Bayes factor. In order to formulate the Bayes factor, one needs to verify the specific form of alternative hypothesis. Under the null hypothesis we assume that $u$ follows a given family of parametric distributions that result in a composite null hypothesis. Under the alternative hypothesis, I model the distribution of $u$ via a flexible semiparametric class of distributions that is capable of approximating the true unknown density of $u$. Under the stated conditions, I shall prove that the proposed test is consistent, and the details are given in Section 3.3. The Bayes factor involves computation of the marginal likelihood, and computation of the marginal likelihood is not straight forward in our case. Particularly, our likelihood function under the null or alternative hypotheses involves with multiple parameters, and the marginal likelihoods do not have closed form. There are several methods on computing Bayes factors and each of them have some advantages and disadvantages (Lewis and Raftery, 1997; Meng and Schilling, 2002; Mira and Nicholls, 2004; Chib and Jeliazkov, 2005; Weinberg, 2012). However, I shall apply the power posterior approach due to Friel and Pettitt (2008) to compute the marginal likelihoods, and the details are given in Section 3.3. Through, simulation studies I compare the performance of the proposed test with the existing test in the literature. The proposed test remarkably outperforms the existing test in terms of the rejection probability of the null hypothesis in both scenarios, when it is false and when it is true. The simulation study is given in Section 3.5 while the analysis of a real data is given in Section 3.6, followed by conclusions given in Section 3.7.

### 3.2 Notation and existing method

Suppose that observed data are $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, and the data follow model (3.1). Assume that $v \sim \mathrm{Normal}(0, \sigma_v^2)$. We want to test if $H_0 : F_u(\cdot) = F_{0u}(\cdot, \boldsymbol{\lambda})$, where $F_{0u}(\cdot, \boldsymbol{\lambda})$ is a parametric family of distributions known up to the parameter $\boldsymbol{\lambda}$.

To test $H_0$, Wang et al. (2011) used the following test statistic:

$$KS_1 = \sup_r |F_{0\epsilon}(r, \widehat{\boldsymbol{\theta}}_{0\epsilon}) - F_n(r)|,$$

where $F_{0\epsilon}(\cdot, \boldsymbol{\theta}_{0\epsilon})$ is the CDF of $\epsilon$ under the null hypothesis, and $\boldsymbol{\theta}_{0\epsilon}$ denotes the set of parameters that include $\sigma_v$ and $\boldsymbol{\lambda}$. The MLE of $\boldsymbol{\theta}_0 \equiv (\boldsymbol{\beta}^T, \boldsymbol{\theta}_{0\epsilon}^T)^T$ under $H_0$ is denoted by $\widehat{\boldsymbol{\theta}}_0 = (\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\theta}}_{0\epsilon}^T)^T$, and $F_n(\cdot)$ denotes the empirical distribution based on the residual $\widehat{\epsilon} = y - \widehat{\beta}_0 - \widehat{\boldsymbol{\beta}}_1^T \boldsymbol{x}$, where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}_1^T)^T$ is the MLE for $\boldsymbol{\beta}$ under the null hypothesis. They estimated the null distribution of the test statistic by a bootstrap method. To be specific, for $b = 1, \ldots, B$, they generated the bootstrap data $y_i^{(b)} = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1^T \boldsymbol{x}_i + \epsilon_i^{(b)}$, $i = 1, \ldots, n$, where $\epsilon_1^{(b)}, \ldots, \epsilon_n^{(b)}$ is a random sample drawn from $F_{0\epsilon}(\cdot, \widehat{\boldsymbol{\theta}}_{0\epsilon})$, and computed the test statistic $KS_1^{(b)} = \sup_r |F_{0\epsilon}(r, \widehat{\boldsymbol{\theta}}_{0\epsilon}^{(b)}) - F_n^*(r)|$, where $\widehat{\boldsymbol{\theta}}_{0\epsilon}^{(b)}$ is the MLE from the $b$th bootstrap data, $F_n^*(r)$ is the empirical distribution based on the residual $\widehat{\epsilon}_i^{(b)} \equiv y_i^{(b)} - \widehat{\beta}_0^{(b)} - \widehat{\boldsymbol{\beta}}_1^{(b)T} \boldsymbol{x}_i$, and $\widehat{\boldsymbol{\beta}}^{(b)}$ is the MLE of $\boldsymbol{\beta}^{(b)}$ obtained from the $b$th bootstrap dataset under $H_0$. Then the $p$-value was estimated by $\sum_{b=1}^{B} I(KS_1^{(b)} > KS_1)/B$.

## 3.3 Testing methodology

### 3.3.1 Models and priors

We assume that model (3.1) along with the normality assumption on $v$ hold. With the parametric model assumption on the distribution of $u$ and the independence assumption between $u$ and $v$, the density function of the composite error $\epsilon$ is

$$f_{0\epsilon}(\epsilon; \sigma_v, \boldsymbol{\lambda}) = \int_0^\infty \frac{1}{\sigma_v} \phi\left(\frac{\epsilon + \eta}{\sigma_v}\right) f_{0u}(\eta; \boldsymbol{\lambda}) d\eta,$$

where $\phi$ is the density of the standard normal distribution and $f_u(\cdot; \boldsymbol{\lambda})$ is a density function of $u$ with parameter $\boldsymbol{\lambda}$. Typically, one tests $H_0 : u$ follows a half-normal distribution, $H_0 : u$ follows an Exponential distribution, or $H_0 : u$ follows a Gamma distribution (Kumbhakar and Lovell, 2003; van den Broeck et al., 1994). Then the conditional distribution of $y$ given the covariates $\boldsymbol{x}$ with parameter $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}^T, \sigma_v, \boldsymbol{\lambda})^T$ can be written as $f_{y|\boldsymbol{x}}(y; \boldsymbol{x}, \boldsymbol{\theta}_0) = f_\epsilon(y - \beta_0 - \boldsymbol{\beta}_1^T \boldsymbol{x}; \sigma_v, \boldsymbol{\lambda})$. Suppose that $\Pi_0(\boldsymbol{\theta}_0)$ is the prior density for $\boldsymbol{\theta}_0$. Then the posterior distribution of $\boldsymbol{\theta}_0$ is

$$\Pi_0(\boldsymbol{\theta}_0|y, \boldsymbol{x}) \propto \left\{ \prod_{i=1}^{n} f_{0\epsilon}(y_i - \beta_0 - \boldsymbol{\beta}_1^T \boldsymbol{x}_i; \sigma_v, \boldsymbol{\lambda}) \right\} \Pi_0(\boldsymbol{\theta}_0).$$

We call this the null model.

For the alternative hypothesis, I consider a broad semiparametric family of distributions which contains the true distribution of $u$ or there is a distribution in this class that approximates the true distribution of $u$. Define $w = e^{-u}$, and model the density of $w$ using splines (Kooperberg and Stone, 1991, 1992). Note that $0 \leq w \leq 1$. We fix the degree of the splines as $q$ and knots as $\{k/K : k = 1, \ldots, K\}$ for large K and we allow $K \to \infty$ for consistency. Then, the density of $w$ is

$$f_w(w; \boldsymbol{\gamma}) = \frac{\exp\{\sum_{k=1}^{L} B_k(w)\gamma_k\}}{\int_0^1 \exp\{\sum_{k=1}^{L} B_k(s)\gamma_k\} ds},$$

where $B_k(\cdot)$ is the $k^{th}$ B-spline basis function with fixed degree $q$, and $(\gamma_1, \ldots, \gamma_L)^T$ denotes $L$ spline coefficients with $L = K + q$. Since $\sum_{k=1}^{L} B_k(w) = 1$, $\gamma_1 + \cdots + \gamma_L = 1$, hence there are $L - 1$ free gamma-parameters. Now, we write the density of $u$ as

$$f_{1u}(u; \boldsymbol{\gamma}) = \frac{\exp\{-u + \sum_{k=1}^{L} B_k(e^{-u})\gamma_k}{\int_0^1 \exp\{\sum_{k=1}^{L} B_k(s)\gamma_k\} ds},$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{L-1})^T$. Therefore, the probability density function of $\epsilon$ under the alternative hypothesis is

$$\begin{aligned} f_{1\epsilon}(\epsilon; \sigma_v, \boldsymbol{\gamma}) &= \int_0^\infty \frac{1}{\sigma_v} \phi\left(\frac{\epsilon + \eta}{\sigma_v}\right) f_{1u}(\eta; \boldsymbol{\gamma}) d\eta \\ &= \int_0^\infty \frac{1}{\sigma_v} \phi\left(\frac{\epsilon + \eta}{\sigma_v}\right) \times \frac{\exp\{-\eta + \sum_{k=1}^{L} B_k(e^{-\eta})\gamma_k}{\int_0^1 \exp\{\sum_{k=1}^{L} B_k(s)\gamma_k\} ds} d\eta \\ &= \int_0^1 \frac{1}{\sigma_v} \phi\left\{\frac{\epsilon - \log(t)}{\sigma_v}\right\} \frac{\exp\{\sum_{k=1}^{L} B_k(t)\gamma_k\}}{\int_0^1 \exp\{\sum_{k=1}^{L} B_k(s)\gamma_k\} ds} dt. \end{aligned}$$

Finally, approximating the integrals by the Gauss-Legendre (GL) quadrature formula we obtain

$$f_{1\epsilon}(\epsilon;\sigma_v,\boldsymbol{\gamma}) \approx \frac{\sum_{j=1}^J c_j\{\phi([\epsilon-\log\{(s_j+1)/2\}]/\sigma_v)/\sigma_v\}\exp[\sum_{k=1}^L B_k\{(s_j+1)/2\}\gamma_k]/2}{\sum_{j=1}^J c_j\exp[\sum_{k=1}^L B_k\{(s_j+1)/2\}\gamma_k]/2},$$

where $c_j$'s are the weights and $s_j$'s are nodes for GL quadrature. Suppose that $\Pi_1(\boldsymbol{\theta}_1)$ is the prior distribution for $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}^T,\sigma_v,\boldsymbol{\gamma})^T$, then the posterior distribution of $\boldsymbol{\theta}_1$ is

$$\Pi_1(\boldsymbol{\theta}_1|y,\boldsymbol{x}) \propto \left\{\prod_{i=1}^n f_{1\epsilon}(y_i-\beta_0-\boldsymbol{\beta}_1^T\boldsymbol{x}_i;\sigma_v,\boldsymbol{\gamma})\right\}\Pi_1(\boldsymbol{\theta}_1).$$

For each component of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ we set the prior distribution on a compact support and the density function is positive and finite valued at every point of the compact support, and specific prior choices are discussed in Section 3.5. With this choice of prior, the posterior contraction rate attains the minimax rate of estimation (Theorem 9.1, Ghosal and van der Vaart, 2017).

### 3.3.2 Calculation of Bayes factor

Let $f_0$ denote the true but unknown distribution for $y$ given $\boldsymbol{x}$. Given that the distribution of $v$ follows $\mathrm{Normal}(0,\sigma_v^2)$, testing whether the distribution of $u$ follows a specified parametric family of distributions or not is equivalent to the following test:

$$H_0 : f_0 \in \mathcal{F}_0 \text{ vs. } H_1 : f_0 \notin \mathcal{F}_0, \tag{3.2}$$

where $\mathcal{F}_0 = \{f_{y|\boldsymbol{x}}(\cdot;\boldsymbol{\theta}_0) : \boldsymbol{\theta}_0 \in \Theta_0\}$ is a class of fixed parametric models defined in previous section. Specifically, for testing $H_0 : u$ follows $f_{0u}$,

$$f_{y|\boldsymbol{x}}(\cdot;\boldsymbol{x},\boldsymbol{\theta}_0) = \int_0^\infty \frac{1}{\sigma_v}\phi\left(\frac{\cdot-\beta_0-\boldsymbol{\beta}_1^T\boldsymbol{x}+\eta}{\sigma_v}\right) f_{0u}(\eta;\boldsymbol{\lambda})d\eta.$$

**Remark 2.** *In order to define Bayes factor, we need to specify models for the alternative hypothesis. I define $\mathcal{F}_1 = \{f_{y|\boldsymbol{x}}(\cdot;\boldsymbol{\theta}_1) : \boldsymbol{\theta}_1 \in \Theta_{1n}\}$ as the alternative model space as described in Section 3.3.1. One advantage of considering this infinite dimensional model is that we can consider the* *case where the true density $f_0$ is not necessarily in $\mathcal{F}_1$. However, as $K \to \infty$ in $f_0$ is approximated by some distribution in $\mathcal{F}_1$. Therefore, it is enough to consider whether $f_0 \in \mathcal{F}_0$ or not.*

Under the same prior probabilities for the models $\mathcal{F}_0$ and $\mathcal{F}_1$, testing (3.2) can be conducted via the Bayes factor defined by

$$B_{01} = \frac{\mathrm{pr}_0(y, \boldsymbol{x})}{\mathrm{pr}_1(y, \boldsymbol{x})} = \frac{\int \{\prod_{i=1}^{n} f_{y|\boldsymbol{x}}(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_0)\}\Pi_0(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0}{\int \{\prod_{i=1}^{n} f_{y|\boldsymbol{x}}(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_1)\}\Pi_1(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1}, \tag{3.3}$$

where $p_0(y, \boldsymbol{x})(p_1(y, \boldsymbol{x}))$ denotes the marginal likelihood under the null (alternative) model. Then the Bayes factor defined in (3.3) can be estimated by $\log(B_{01}) = \log\{\mathrm{pr}_0(y, \boldsymbol{x})\} - \log\{\mathrm{pr}_1(y, \boldsymbol{x})\}$.

In order to evaluate/compute marginal likelihoods, I adapt the power posterior approach. For the notational convenient, we omit the subscript in what follows. The computation of the marginal likelihood is based on the identity (Friel and Pettitt, 2008)

$$\log\{\mathrm{pr}(y, \boldsymbol{x})\} = \int_0^1 E_{\theta|y,\boldsymbol{x},t}[\log\{f_{y|\boldsymbol{x}}(y; \boldsymbol{x}, \boldsymbol{\theta})\}]dt, \tag{3.4}$$

where the expectation is with respect to the power posterior distribution defined as

$$\mathrm{pr}(\boldsymbol{\theta}|y, \boldsymbol{x}, t) = \frac{\{f_{y|\boldsymbol{x}}(y; \boldsymbol{x}, \boldsymbol{\theta})\}^t \Pi(\boldsymbol{\theta})}{\int_0^1 \{f_{y|\boldsymbol{x}}(y; \boldsymbol{x}, \boldsymbol{\theta})\}^t \Pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \tag{3.5}$$

for $t \in [0, 1]$. We assume that $\Pi(\boldsymbol{\theta})$ is proper so that the denominator of (3.5) is finite. Note that $\mathrm{pr}(\boldsymbol{\theta}|y, \boldsymbol{x}, t = 1)$ is the posterior distribution and $\mathrm{pr}(\boldsymbol{\theta}|y, \boldsymbol{x}, t = 0)$ is the prior distribution of $\boldsymbol{\theta}$.

We discretize $t \in [0, 1]$, say, $0 = t_0 < t_1 < \cdots < t_r = 1$. For each $t_i$, posterior samples from (3.5) are generated using MCMC sampling to compute $E_{\theta|y,\boldsymbol{x},t_i}[\log\{f_{y|\boldsymbol{x}}(y; \boldsymbol{x}, \boldsymbol{\theta})\}]$ via Monte Carlo integration. Finally the marginal likelihood is estimated via a trapezoidal rule

$$\log\{\mathrm{pr}(y, \boldsymbol{x})\} \approx \sum_{i=0}^{r-1} (t_{i+1} - t_i) \frac{E_{\theta|y,\boldsymbol{x},t_{i+1}}[\log\{f_{y|\boldsymbol{x}}(y; \boldsymbol{x}, \boldsymbol{\theta})\}] + E_{\theta|y,\boldsymbol{x},t_i}[\log\{f_{y|\boldsymbol{x}}(y; \boldsymbol{x}, \boldsymbol{\theta})\}]}{2}.$$

In the next section I show that the Bayes factor is consistent for testing $H_0$. Following that we define the test function

$$\Psi_n = \begin{cases} 1 & \text{if } B_{01} < t_{\text{cutoff}} \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

I took $t_{\text{cutoff}} = 1/3$. This choice of $t_{\text{cutoff}}$ was guided by the recommendation given in Kass and Raftery (1995), where $1/3 \leq BF_{01} < 1$ implies that there is *not worth more than a bare mention* against $H_0$, $1/20 \leq BF_{01} < 1/3$ indicates that there is a *positive evidence* against $H_0$, $1/150 \leq BF_{01} < 1/20$ implies a *strong evidence* and $BF_{01} < 1/150$ shows *very strong evidence* against $H_0$.

### 3.4 Large sample properties of the test

The test will be called consistent if $E_{f_0}(\Psi_n) \to 1$ as $n \to \infty$ for any $f_0 \notin \mathcal{F}_0$. A consistent test is called Chernoff-consistent if the probability of Type-I error goes to zero as $n \to \infty$ (Shao, 1999, p.111). Our proposed test (3.6) based on Bayes fact $B_{01}$ is Chernoff-consistent if we can show that $B_{01}$ is consistent according to the following definition.

**Definition 1.** *The Bayes factor defined in (3.3) under the hypotheses (3.2) is said to be consistent if*

*(a)* $B_{01} \to \infty$ *if* $f_0 \in \mathcal{F}_0$ *in probability*

*(b)* $B_{01} \to 0$ *if* $f_0 \notin \mathcal{F}_0$ *in probability,*

*with respect to an appropriate measure whose density is* $f_0$.

**Proposition 1.** *Under the conditions stated in Appendix,* $B_{01}$ *defined in (3.3) is consistent.*

Proof of this result is given in the Appendix.

### 3.5 Simulation studies

**Simulation design:** In this section, I present the finite sample performance of the proposed test via Monte-Carlo simulations. I simulated datasets with sample sizes $n = 50, 100$ and 200, which consist of a scalar covariate $x$ and a response $y$ following model (3.1), where $x \sim \mathrm{Normal}(0,1)$, $\beta_0 = 1$ and $\beta_1 = 1.5$. I set the composed error $\epsilon = v - u$, where $v$ was simulated from $\mathrm{Normal}(0,(1/\sqrt{5})^2)$ and $u$ from the following distributions:

**Scenario 1.** (NHN) $u \sim \mathrm{Normal}^{+}(0,(2/\sqrt{5})^2)$,

**Scenario 2.** (NEX) $u \sim \mathrm{EXP}(2/\sqrt{5})$,

**Scenario 3.** (NGM) $u \sim \mathrm{Gamma}(1.75, 0.41)$.

Here $\mathrm{Normal}^{+}(0,\sigma_u^2)$ denotes the half-normal distribution whose density function is $(2/\sigma_u)\phi(u/\sigma_u)$ $\times I(u > 0)$, where $\phi$ is the density of the standard normal distribution, $\mathrm{EXP}(\sigma_\mathrm{u})$ represents the exponential distribution with variance $\sigma_u^2$, and $\mathrm{Gamma(a,b)}$ denotes the gamma distribution with mean $ab$ and variance $ab^2$. Note that scenarios 1 and 2 are similar to the cases considered in the simulation study of Wang et al. (2011). Note that when $v \sim \mathrm{Normal}(0,\sigma_v^2)$ and $u \sim \mathrm{Normal}^{+}(0,\sigma_u^2)$, $\sigma^2 \equiv \text{var}(\epsilon) = \sigma_v^2 + \sigma_u^2$. Wang et al. (2011) considered several combinations of $(\sigma_v^2, \sigma_u^2)$ while constraining $\sigma^2 = 1$. In scenario 3, $u$ has the same mean and variance as in the first case (NHN).

**Method of analysis:** For each simulated dataset I test $H_0$ : $u$ follows $\mathrm{Normal}^{+}(0,\sigma_u^2)$ against $H_a$ : $u$ does not follow $\mathrm{Normal}^{+}(0,\sigma_u^2)$. For analyzing data under the null hypothesis, I used independent $\mathrm{Normal}(0,1)$ prior distribution truncated between $-4$ and $4$ for the regression parameters as well as for $\log(\sigma_v^2)$ and $\log(\sigma_u^2)$. Also, for the alternative hypothesis, I used independent $\mathrm{Normal}(0,1)$ prior distribution truncated between $-4$ and $4$ for the regression parameters as well as for $\log(\sigma_v^2)$ and all $L-1$ components of $\boldsymbol{\gamma}$-parameter. The results are fairly insensitive towards the variance of the prior distribution when it was varied between 0.5 and 5. For the spline estimation, I considered $0.25, 0.5, 0.75$ as internal knot points. To evaluate log-marginal likelihood, I used $t_r = (r/20)^5$ for $r = 0, \ldots, 20$ as the grid points for the trapezoid rule. I also applied the method of Wang et al. (2011) described in Section 3.2.

**Results:** Table 3.1 contains the simulation results. Under each scenario and for different sample sizes I simulated $1,000$ datasets, and I computed proportion of times we reject $H_0$ (given in column 6), also columns 3, 4 and 5 contains the proportion of times the Bayes factor falls $[1/20, 1/3)$,

Table 3.1: The first column shows the true data generating process. The entries from column 3 to 5 are the proportion of cases where $BF_{01}$ falls into $[1/20, 1/3)$, $[1/150, 1/20)$ and $[0, 1/150)$, respectively. The entries in the sixth column presents the proportion of cases where $BF_{01} < 1/3$. The records in the last column show the proportion of simulated datasets where the $p$-value of the bootstrap KS test (Wang et al., 2011) is less than $0.05$.

| Model | $n$ | Evidence against $H_0$ | | | Reject $H_0$ | Bootstrap KS |
|---|---|---|---|---|---|---|
| | | Positive | Strong | Very strong | | |
| NHN | 50 | 0.031 | 0.001 | 0 | 0.032 | 0.023 |
| | 100 | 0.027 | 0.002 | 0 | 0.029 | 0.019 |
| | 200 | 0.017 | 0.000 | 0 | 0.017 | 0.039 |
| NEX | 50 | 0.307 | 0.131 | 0.082 | 0.520 | 0.127 |
| | 100 | 0.299 | 0.215 | 0.225 | 0.739 | 0.240 |
| | 200 | 0.213 | 0.228 | 0.451 | 0.892 | 0.448 |
| NGM | 50 | 0.118 | 0.018 | 0.008 | 0.144 | 0.046 |
| | 100 | 0.122 | 0.033 | 0.010 | 0.165 | 0.054 |
| | 200 | 0.126 | 0.052 | 0.019 | 0.197 | 0.065 |

$[1/150, 1/20)$, and $[0, 1/150]$, respectively. In scenario 1 (NHN) data follow the null hypothesis. The proportion of rejection of $H_0$ (column 6) is decreasing towards 0 as the sample size increases. For scenarios 2 (NEX) and 3 (NGM) where data do not follow $H_0$, the rejection proportions are increasing with the sample size. Column 7 of Table 3.1 shows the proportion of rejection of $H_0$ when Wang et al. (2011)'s test is used. The proposed test clearly outperforms the existing test in terms of both Type-I and Type-II error rates.

### 3.6 Analysis of the U.S. electricity data

XI apply the proposed Bayesian approach to test the distributional assumption of the technical inefficiency of the SF model for analyzing the U.S. electricity data. The dataset contains information on $Q$, a function of labor, capital, and fuel, $P_l$, the price of labor, $P_k$, the price of capital, $P_f$, the price of fuel, and the cost of production from $n = 123$ companies (Greene, 1990). Because of linear homogeneity for $P_l$, $P_k$ and $P_f$, *a priori* restriction was adopted to the model by dividing each of these quantities by $P_f$ and the resulting model was (van den Broeck et al., 1994):

$$\begin{aligned}-\log(Cost/P_f) &= \beta_0+\beta_1\{-\log(Q)\}+\beta_2[-\{\log(Q)\}^2]+\beta_3\{-\log(P_l/P_f)\} \\ &\quad +\beta_4\{-\log(P_k/P_f)\}+v-u.\end{aligned}$$

Thus in this case, higher cost is likely to be caused by inefficiency $u > 0$.

Previously, Greene (1990) and van den Broeck et al. (1994) analyzed this dataset. Greene (1990) considered MLE based on normal $v$ and half normal $u$, normal $v$ and gamma $u$, and normal $v$ and exponential $u$ assumption. On the other hand, van den Broeck et al. (1994) used normal $v$ and gamma $u$ to analyze this dataset using a Bayesian approach but they restricted the gamma shape parameter to 1, 2 and 3. The estimated regression parameters were more or less similar across different models. However, resulting $E(u)$, var$(u)$ and var$(v)$ were varying by models. This empirically explains that the distributional assumption of $u$ matters, specifically for predicting inefficiency.

Assuming the regression structure given in (3.7) is true and $v$ follows mean zero normal distribution $\mathrm{Normal}(0, \sigma_{\mathrm{v}}^2)$, I tested $H_0 : u \sim \mathrm{Normal}^+(0, \sigma_{\mathrm{u}}^2)$. Similar to simulation study, here also I used mean-zero normal truncated distribution as the prior for each parameter, and I took three different prior variances, 0.064, 0.5, 1. Particularly, 0.064 was the maximum of all the square of standard errors of the parameters when MLE was calculated by fitting the null model to the data. The resulting Bayes factor was $4.227e-06$, $4.346e-08$, $1.359e-05$, for three different prior variances. These values indicate strong evidence against the null hypothesis, and using our test (3.6) we rejected $H_0$. For the comparison purpose, I calculated the $p$-value of the bootstrap KS test (Wang et al., 2011), and that was 0.0015, and at the $5\%$ level, we also rejected the $H_0$. Based on both tests, we concluded that half-normal assumption was an unrealistic assumption on $u$.

## 3.7 Conclusions

In this chapter, a test of the distributional assumption of technical inefficiency is developed in a Bayesian context. In order to construct the Bayes factor I have considered a flexible semiparametric family of distributions as an alternative model specification which is capable of approximating any distribution with reasonable accuracy. The consistency property of the proposed test is established. The advantage of the proposed test is shown via Monte-Carlo simulation studies. The results are fairly robust when the prior variances are varied within a reasonable range. The proposed idea of Bayesian test can be extended to test the presence of inefficiency term in the SF model.

# 4. FREQUENTIST STANDARD ERRORS OF BAYES ESTIMATORS [*]

## 4.1 Background and literature review

Suppose that $f(\bullet|\theta)$ is the data generating density and $\pi(\theta)$ is the prior distribution for the parameter $\theta$. Let $\boldsymbol{D}$ be the observed data. Then the posterior distribution of $\theta$ is

$$\pi(\theta|\boldsymbol{D}) = K_\pi f(\boldsymbol{D}|\theta)\pi(\theta),$$

where $K_\pi$ denotes the normalizing constant. There are several posterior summaries, such as the mean, $m(\boldsymbol{D}) = E(\theta|\boldsymbol{D}) = \int \theta\pi(\theta|\boldsymbol{D})d\theta$, the posterior median $\widetilde{m}(\boldsymbol{D})$, which satisfies $\int_{-\infty}^{\widetilde{m}(\boldsymbol{D})} \pi(\theta|\boldsymbol{D})d\theta = 0.5$, the $\alpha^{\text{th}}$ quantile $q_\alpha(\boldsymbol{D})$, that satisfies $\int_{-\infty}^{q_\alpha(\boldsymbol{D})} \pi(\theta|\boldsymbol{D})d\theta = \alpha$ for any $\alpha \in (0,1)$, and the posterior mode $m_o(\boldsymbol{D}) = \arg\max_\theta \pi(\theta|\boldsymbol{D})$. By $s(\boldsymbol{D})$ I refer to any summary of the posterior distribution. Throughout this article I assume that $\boldsymbol{D}$ consists of $(X_1, \ldots, X_n)$ iid observations. The goal of this chapter is to discuss approaches of computing the frequentist standard error of $s(\boldsymbol{D})$.

Under a large sample, the observed data dominates the prior information in a Bayesian framework, and under standard regularity conditions, the posterior distribution of finite dimensional model parameters converges to the Gaussian distribution with the maximum likelihood estimator and the inverse of the Fisher Information matrix as the asymptotic mean and asymptotic variance, respectively. This asymptotic connection indicates that the Bayesian philosophy of integrating the observed data and the prior knowledge can be seen as a general procedure that encompasses the frequentist procedure as a special case. Therefore frequentist standard error of a Bayes estimator is a way of assessing uncertainty of the general procedure. Particularly, for a large sample, the frequentist variance of the posterior mean converges to the inverse of the Fisher's information matrix. From the Bayesian perspective, frequentist standard errors can be used for comparing un-

[*] Reprinted with permission from "Frequentist Standard Errors of Bayes Estimators" by DongHyuk Lee, Raymond J. Carroll, Samiran Sinha, 2017. *Computational Statistics*, 32, 867–888, Copyright [2017] by Springer Nature.

certainty of estimators under different priors (Efron, 2015). Although the posterior standard error (the standard deviation of the posterior distribution) is a measure of uncertainty of the posterior distribution, Efron (2015) argued that in a Bayesian paradigm, accuracy of a Bayes estimator, such as posterior mean could be judged based on the posterior distribution given that the prior distribution of the parameter reflects the truth in some degree. Therefore, finding accuracy of a Bayes estimator in an objective way among different subjective and objective priors is important. In fact, Berger (2006) discussed that the "pseudo-Bayes procedures" where subjective, objective, or a mixture of subjective and objective priors are used, often fail to provide any guidance on the performance of true subjective or objective Bayesian analysis. He then pointed out the necessity of validating these Bayesian approaches, and frequentist standard error of a posterior summary can be seen as a measure of such validation. Although Bayes factor is a way of comparing Bayesian procedures, many practitioners still want to compare estimators based on a frequentist uncertainty measure. Therefore, despite an apparent lack of coherence for incorporating a frequentist comparisons among Bayes procedures, it provides a measure of comparing uncertainties of the estimators. Of course, we should not use this measure solely to elicit the optimal prior for a Bayes procedure as for a proper comparison one should consider consistency, posterior convergence rate, along with the uncertainty of the estimator.

Efron (2015) proposed methods for computing frequentist standard errors of the posterior mean of a function of a parameter. In particular, he derived the approximate frequentist standard deviation of the posterior mean of a parameter based on the delta method. Suppose that $T$ is the sufficient statistic. Following our notations, his formula for the approximate standard deviation of $\widehat{t} = E\{t(\theta)|\boldsymbol{D}\} = E\{t(\theta)|T\}$, the posterior mean of $t(\theta)$, a function of $\theta$, is $[\text{cov}\{t(\theta), \alpha_T(\theta)|T\}^T V_\theta \text{cov}\{t(\theta), \alpha_T(\theta)|T\}]^{1/2}$, where $\alpha_T(\theta) = \partial\text{log}\{f_\theta(T)\}/\partial T$ denotes the gradient of $\log\{f_\theta(T)\}$ with respect to $T$, the sufficient statistic for $\theta$, $f_\theta(T)$ is the density for the sufficient statistic $T$, and $V_\theta$ denotes the variance of the sufficient statistic. For application of this method it is critical that $V_\theta$ is readily available. Secondly, one key component of the delta method is the gradient of $\widehat{t}$ with respect to $T$, and here this gradient is expressed as the posterior covariance

$\text{cov}\{t(\theta), \alpha_T(\theta)|T\}$. Expressing $\partial\widehat{t}/\partial T$ as $\text{cov}\{t(\theta), \alpha_T(\theta)|T\}$ critically relies on the fact that $\widehat{t}$ is a posterior expectation. This posterior covariance is easy to estimate from a sample from the posterior distribution of $\theta$. Therefore, when $V_\theta$ is available and $\widehat{t}$ is a posterior expectation, then Efron's formula is easy to apply and it is computationally fast.

In a special case with the exponential family of distributions where $\theta$ is considered to be the natural or canonical parameter vector, along with an uninformative prior for $\theta$, he showed that the standard error of the posterior mean of $t(\theta) = \theta$ can be computed without running the MCMC step to generate posterior samples for computing $\text{cov}\{\theta, \alpha_T(\theta)|T\}$. In lieu of the MCMC sampling, he used a parametric bootstrap resampling technique (Efron, 2012) to compute the posterior covariance term. Although the proposed method is applicable to only posterior means and when $V_\theta$ is easily available, the main advantage is that this method, when it is applicable, is much faster than the regular bootstrap procedure.

Inspired by this work I propose a general method of efficiently computing the frequentist standard error not only of the posterior mean but also of any posterior summary, $s(\boldsymbol{D})$. Our method is applicable for data generated from any parametric model, not necessarily from an exponential family of distributions. The proposed method relies on the bootstrap idea. Usually, the standard error of an estimator can be computed by the bootstrap method (Efron and Tibshirani, 1986), where the standard error is estimated by the standard deviation of the Bayes estimators obtained from a large number of bootstrap samples. On the other hand, the Bayes estimator for a bootstrap sample is usually calculated by drawing a large number of Markov chain Monte Carlo (MCMC) samples, which is often time consuming, and consequently drawing posterior samples for each of the bootstrap data can be a prohibitively time consuming task.

The main aim of Chapter 4 is to reduce this computation time. To do so, the MCMC method will be used once to draw samples from the posterior distribution of the parameters given the original data. Then use these posterior samples along with the importance sampling idea to compute the posterior summary for each bootstrap data. The details are discussed in the following sections. To make it clearer, we want to re-state that in the proposed method, we do need bootstrap sampling, but we bypass the MCMC sampling for each bootstrap data by a clever use of the importance sampling method. Here is a brief description of the importance sampling method in a few words. Suppose that we are interested in estimating $\theta = \int g(x)f(x)dx$, where $f(x)$ is a density. With another density $h(x)$, we can re-write $\theta = \int g(x)\omega(x)h(x)dx$, where $\omega(x) = f(x)/h(x)$ is called the importance weight. Then the importance sampling estimator of $\theta$ is $\widehat{\theta} = m^{-1}\sum_{i=1}^{m} g(x_i)\omega(x_i)$, where $x_1, \ldots, x_m$ are iid from $h(x)$. This technique is quite useful for efficient estimation of tail probabilities, and is used for drawing bootstrap samples, specially for estimating standard error of small probabilities (pp. 349, Efron and Tibshirani, 1994). However, I use importance sampling technique to compute estimators based on a bootstrap re-sampled data. Basically in the proposed approach bootstrap samples are drawn using standard bootstrap resampling technique and then importance sampling is used to compute the Bayes estimators. Although importance sampling idea has been used in many other contexts, including but not limited to the simulated maximum likelihood estimation, computer graphics, modelling stock market data, modelling linear and non-linear dynamic processes (Liang, 2002), the use of this technique in the present context seems to be novel.

A brief outline of the chapter is as follows. In Section 4.2 I provide the widely used examples. The main idea related to the posterior mean is discussed in Section 4.3, while Section 4.4 considers posterior quantiles and the posterior mode. Section 4.5 describes the results of two simulation studies and a real data examples. Section 4.6 contains conclusions.

## 4.2 Motivating examples

To motivate this research first I consider three commonly used models.

**Logistic regression model:** Suppose that $Y_1, \ldots, Y_n$ are independently drawn from the Bernoulli($p_i$) distribution, where $p_i = \mathrm{pr}(Y_i = 1|X_i) = \{1 + \exp(-\alpha - \beta X_i)\}^{-1}$ with a scalar covariate $X_i$. Assume priors $\alpha \sim \mathrm{Normal}(a, \sigma^2)$ and $\beta \sim \mathrm{Normal}(b, \tau^2)$, and let $\boldsymbol{D}$ denote the observed data $\{(X_i, Y_i), i = 1, \ldots, n\}$. Then the posterior distribution of $\alpha$ and $\beta$ is

$$\pi(\alpha, \beta | \boldsymbol{D}) \propto \prod_{i=1}^{n} \left\{ \frac{1}{1 + \exp(-\alpha - \beta X_i)} \right\}^{Y_i} \left\{ \frac{\exp(-\alpha - \beta X_i)}{1 + \exp(-\alpha - \beta X_i)} \right\}^{(1-Y_i)} \times \frac{\exp\{-(\alpha - a)^2/2\sigma^2\}}{\sqrt{2\pi\sigma^2}} \times \frac{\exp\{-(\beta - b)^2/2\tau^2\}}{\sqrt{2\pi\tau^2}}.$$

For computing any posterior summary for $\pi(\alpha, \beta|\boldsymbol{D})$, usually we draw posterior samples from $\pi(\alpha, \beta|\boldsymbol{D})$ using the MCMC method. So, a numerical method is must for computing frequentist standard errors of any summary of the posterior distribution. In the simulation section, for illustration, we apply the proposed method on this model.

**Linear measurement error model:** Now, we consider the following simple linear regression problem, where using the observed data $\boldsymbol{D} = \{(Y_i, W_i), i = 1, \ldots, n\}$, we want to fit $Y_i = \alpha + X_i\beta + \epsilon_i$, where $X_i$ is unobserved but we observed its surrogate variable $W_i$, and $\epsilon_i \sim \text{Normal}(0, \sigma_\epsilon^2)$. The observed surrogate $W_i$ is associated with the true $X_i$ through the classical additive measurement error model $W_i = X_i + U_i$, where $U_i \sim \text{Normal}(0, \sigma_u^2)$ and $\sigma_u^2$ is considered to be known for simplicity. We further assume that measurement error is nondifferential such that $Y_i$ is conditionally independent of $W_i$ given the true $X_i$ (pp. 36, Carroll et al., 2006), and $X_i \sim \text{Normal}(\mu_x, \sigma_x^2)$.

It is well-known that the simple linear regression of $Y$ on $W$ will cause an attenuation towards 0 by the multiplicative factor $\sigma_x^2/(\sigma_x^2 + \sigma_u^2)$. One of the corrections for attenuation is the method of moments. That is, the resulting estimator $\widehat{\beta} = \widehat{\beta}_w \widehat{\sigma}_w^2/(\widehat{\sigma}_w^2 - \sigma_u^2)$, where $\widehat{\beta}_w$ is the OLS estimator ignoring measurement error, $\widehat{\sigma}_w^2$ is the sample variance of the observed $W$, and $\sigma_u^2$ is the variance of $U$ (Section 2.5, Fuller, 1987; Section 3.4.1, Carroll et al., 2006). In addition, it is well-known that $\widehat{\beta}$ has no finite moments, because the denominator term $\widehat{\sigma}_w^2 - \sigma_u^2$ can get arbitrarily close to zero (Fuller, 1987). Therefore, Bayesian calculations are an attractive alternative.

We attempt to use a Bayesian inference for the parameters $\theta = (\alpha, \beta, \mu_x, \sigma_x^2, \sigma_\epsilon^2)$ in which $\alpha$ and $\beta$ are the main parameters of interest. Assigning normal priors, $\text{Normal}(0, \sigma_\alpha^2)$, $\text{Normal}(0, \sigma_\beta^2)$, $\text{Normal}(0, \sigma_\mu^2)$ for $\alpha$, $\beta$, $\mu_x$, respectively and inverse gamma priors $\text{IG}(\delta_x, \lambda_x)$, $\text{IG}(\delta_\epsilon, \lambda_\epsilon)$ for $\sigma_x^2$,

$\sigma_\epsilon^2$, respectively (Section 9.4, Carroll et al., 2006), the joint posterior distribution of $\theta$ and the latent variable $X = (X_1, \ldots, X_n)$ is

$$\begin{aligned}\pi(\theta, X|\boldsymbol{D}) \;\propto\; & (\sigma_\epsilon^2)^{-n/2-\delta_\epsilon-1}(\sigma_x^2)^{-n/2-\delta_x-1}\exp\bigg\{-\frac{\sum_{i=1}^n (Y_i-\alpha-X_i\beta)^2/2+\lambda_\epsilon}{\sigma_\epsilon^2}\\ & -\frac{\sum_{i=1}^n (W_i-X_i)^2/2+\lambda_u}{\sigma_u^2}-\frac{\sum_{i=1}^n (X_i-\mu_x)^2/2+\lambda_x}{\sigma_x^2}-\frac{\alpha^2}{\sigma_\alpha^2}-\frac{\beta^2}{\sigma_\beta^2}-\frac{\mu_x^2}{\sigma_\mu^2}\bigg\}.\end{aligned}$$

Due to the conjugacy of the prior distributions, it is easy to apply the Gibbs sampler to draw posterior samples from $\pi(\theta, X|\boldsymbol{D})$. Specifically, the conditional posterior distributions of $\alpha$ and $\beta$ given other parameters and the latent variable $X$ are normal distributions so that we can easily obtain their posterior summaries. However, it is not an easy problem to find the variances of their posterior summaries mainly because they are dependent on the unobserved $X$. Thus a numerical method is required.

**Weibull regression model:** Suppose that $T_1, \ldots, T_n$ are independently drawn from the $\mathrm{Weibull}(\alpha, \lambda_\mathrm{i})$ distribution whose density is $g(t|\alpha, \lambda) = \alpha t^{\alpha-1}\exp\{\lambda - \exp(\lambda)t^\alpha\}$ (eq. 2.2.1, Ibrahim et al., 2001). Let $C_1, \ldots, C_n$ be the corresponding censoring times whose distribution does not include any information about parameters $\alpha$ and $\lambda_i$ (non-informative censoring) and $\Delta_1, \ldots, \Delta_n$ be the censoring indicator where $\Delta_i = 1$ if $T_i \le C_i$ (observed) and $\Delta_i = 0$ if $T_i > C_i$ (censored). In this example, let $\boldsymbol{D} = \{Y_i, \Delta_i, X_i, i = 1, \ldots, n\}$, where $Y_i = \min(T_i, C_i)$, and $X_i$ is the covariate for the $i^{\mathrm{th}}$ individual. We regress the parameter $\lambda_i$ on covariates $X_i$, i.e., $\lambda_i = X_i'\beta$. Assigning a normal prior, $\mathrm{Normal}(\mu_0, \Sigma_0)$, for $\beta$ and a gamma prior, $\mathrm{Gamma}(\alpha_0, \kappa_0)$, for $\alpha$, the posterior distribution of $\alpha$ and $\beta$ is

$$\begin{aligned}\pi(\alpha, \beta|\boldsymbol{D}) \;\propto\; & \alpha^{\alpha_0+d-1}\exp\bigg[\sum_{i=1}^n\{\Delta_i X_i'\beta + \Delta_i(\alpha-1)\log(Y_i) - Y_i^\alpha\exp(X_i'\beta)\}\\ & -\kappa_0\alpha - \frac{1}{2}(\beta-\mu_0)\Sigma^{-1}(\beta-\mu_0)\bigg],\end{aligned}$$

where $d = \sum_{i=1}^n \Delta_i$ (eq. 2.2.4, Ibrahim et al., 2001). Likewise in the logistic regression example, we need not only to draw posterior samples from $\pi(\alpha, \beta|\boldsymbol{D})$ using the MCMC method to evaluate any posterior summary, we also necessitate a numerical procedures for computing frequentist standard errors of those posterior summaries. We use this model to analyze the Melanoma data set in Section 4.5.3, and compute uncertainty measures using the proposed approach.

These examples show that even for these well researched models, posterior summaries may not have an explicit expression that is easy to compute. Additionally, the computation of the standard error of the posterior summaries requires extra numerical work.

### 4.3 Standard errors of posterior means

In this section we concentrate only on the posterior mean and its standard error calculations. In Section 4.4, we provide recipes for efficiently calculating frequentist standard errors of other types of Bayes estimators. For any generic vector $a$, we shall use $a^{\otimes 2}$ to denote $aa^T$.

The frequentist standard error of the posterior mean of $\theta$, $\widehat{\theta} = m^*(\boldsymbol{D}) = E(\theta|\boldsymbol{D}) = E_{\pi(\cdot|\boldsymbol{D})}(\theta)$, where $\pi(\cdot|\boldsymbol{D})$ is the posterior distribution, is

$$\sqrt{\text{var}_F(\widehat{\theta})} = \sqrt{\int \{m^*(\boldsymbol{D})\}^{\otimes 2} dF(\boldsymbol{D}|\theta) - \left\{\int m^*(\boldsymbol{D}) dF(\boldsymbol{D}|\theta)\right\}^{\otimes 2}}.$$

Suppose that one draws $B$ random samples each of size $m$ from the posterior distribution $\pi(\theta|\boldsymbol{D})$. Denote the $b^{\text{th}}$ sample as $(\theta_{b1}, \ldots, \theta_{bm})$, $b = 1, \cdots, B$. Define $\widehat{\theta}_b = \sum_{j=1}^m \theta_{bj}/m$, and $\overline{\theta}. = \sum_{b=1}^B \widehat{\theta}_b/B$. It is obvious that the variance among $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$ does not estimate $\text{var}_F(\widehat{\theta})$ as $(B-1)^{-1}\sum_{b=1}^B(\widehat{\theta}_b - \overline{\theta}.)^2 \to (1/m)\text{var}_{\pi(\cdot|\boldsymbol{D})}(\theta)$ almost surely as $B \to \infty$, where $\text{var}_{\pi(\cdot|\boldsymbol{D})}(\theta)$ denotes the posterior variance of $\theta$. One obvious approach to estimate $\text{var}_F(\widehat{\theta})$ is to adopt the bootstrap idea. In the bootstrap world, instead of $\text{var}_F(\widehat{\theta})$ we target estimating $\text{var}_{\widehat{F}}(\widehat{\theta})$, where the observed data are treated as the entire population. In the bootstrap method, we draw $B$ bootstrap samples with replacement from the original data, calculate the posterior mean for each bootstrap sample, and then take the variance of the $B$ posterior means. Let $\boldsymbol{D}^{(b)}$ be the $b^{\text{th}}$ bootstrap data, and $\pi(\theta|\boldsymbol{D}^{(b)})$ be the corresponding posterior distribution. Define $\widehat{\theta}^{(b)} = E(\theta|\boldsymbol{D}^{(b)}) = E_{\pi(\cdot|\boldsymbol{D}^{(b)})}(\theta)$ as the posterior mean of $\theta$ for the $b^{\text{th}}$ bootstrap data. Further define $\overline{\theta}^{(\cdot)} = \sum_{b=1}^{B} \widehat{\theta}^{(b)}/B$. Then

$$(B-1)^{-1}\sum_{b=1}^{B}(\widehat{\theta}^{(b)} - \overline{\theta}^{(\cdot)})^2 \to \text{var}_{\widehat{F}}(\widehat{\theta}) \text{ as } B \to \infty.$$

In practice, $\widehat{\theta}^{(b)}$ is estimated by the Monte Carlo estimator $\widehat{\theta}^{(b)}_{\text{mc}} = \sum_{j=1}^{M} \theta_j^{(b)}/M$, where $\theta_1^{(b)}, \ldots, \theta_M^{(b)}$ are $M$ random draws from $\pi(\theta|\boldsymbol{D}^{(b)})$, and $\widehat{\theta}^{(b)}_{\text{mc}} \to \widehat{\theta}^{(b)}$ almost surely as $M \to \infty$. Also, define $\overline{\theta}^{(\cdot)}_{\text{mc}} = B^{-1}\sum_{b=1}^{B} \widehat{\theta}^{(b)}_{\text{mc}}$. Then as $M \to \infty$,

$$(B-1)^{-1}\sum_{b=1}^{B}(\widehat{\theta}^{(b)}_{\text{mc}} - \overline{\theta}^{(\cdot)}_{\text{mc}})^2 \to (B-1)^{-1}\sum_{b=1}^{B}(\widehat{\theta}^{(b)} - \overline{\theta}^{(\cdot)})^2.$$

Hence $\sum_{b=1}^{B}(\widehat{\theta}^{(b)}_{\text{mc}} - \overline{\theta}^{(\cdot)}_{\text{mc}})^2/(B-1)$ will be used as the estimator of $\text{var}_{\widehat{F}}(\widehat{\theta})$. In the following paragraph we describe how we estimate $\widehat{\theta}^{(1)}, \ldots, \widehat{\theta}^{(B)}$ without having numerically computing $B$ posterior distributions using $B$ MCMC chains thereby saving lots of computation time.

Suppose that using MCMC method we have drawn $\theta_1, \ldots, \theta_M$ from $\pi(\theta|\boldsymbol{D})$, the posterior distribution of $\theta$ given the entire data $\boldsymbol{D}$. Suppose that in the $b^{\text{th}}$ bootstrap sample, $X_i$ occurs $r_i^{(b)}$ times, where $0 \le r_i^{(b)} \le n$, but $\sum_{i=1}^{n} r_i^{(b)} = n$. Then the posterior distribution of $\theta$ given the $b^{\text{th}}$ bootstrap data $\boldsymbol{D}^{(b)}$ is

$$\pi(\theta|\boldsymbol{D}^{(b)}) = \frac{\prod_{i=1}^{n} f^{r_i^{(b)}}(X_i|\theta)\pi(\theta)}{\int \prod_{i=1}^{n} f^{r_i^{(b)}}(X_i|\theta)\pi(\theta)d\theta},$$

so

$$\widehat{\theta}^{(b)} = \int \theta\pi(\theta|\boldsymbol{D}^{(b)})d\theta = \frac{\int \theta \prod_{i=1}^{n} f^{r_i^{(b)}}(X_i|\theta)\pi(\theta)d\theta}{\int \prod_{i=1}^{n} f^{r_i^{(b)}}(X_i|\theta)\pi(\theta)d\theta} = \frac{G_1^{(b)}}{G_0^{(b)}},$$

where $G_s^{(b)} = \int \theta^s \prod_{i=1}^{n} f^{r_i^{(b)}}(X_i|\theta)\pi(\theta)d\theta$ for $s = 0$ and $1$. Next, we can re-write

$$G_s^{(b)} \;=\; \frac{1}{K_\pi}\int \theta^s \frac{\prod_{i=1}^{n} f^{r_i^{(b)}}(X_i|\theta)\pi(\theta)}{\prod_{i=1}^{n} f(X_i|\theta)\pi(\theta)} K_\pi \prod_{i=1}^{n} f(X_i|\theta)\pi(\theta)d\theta$$

$$= \frac{1}{K_\pi} \int \theta^s \omega^{(b)}(\theta) K_\pi \prod_{i=1}^n f(X_i|\theta)\pi(\theta) d\theta,$$

where the importance weight $\omega^{(b)}(\theta) = \prod_{i=1}^n f^{r_i^{(b)}}(X_i|\theta) / \prod_{i=1}^n f(X_i|\theta) = \prod_{i=1}^n f^{(r_i^{(b)}-1)}(X_i|\theta)$. Hence $\widehat{\theta}^{(b)}$ can be estimated by

$$\widehat{\theta}_{\text{is}}^{(b)} = \frac{\sum_{j=1}^M \theta_j \omega^{(b)}(\theta_j)}{\sum_{j=1}^M \omega^{(b)}(\theta_j)},$$

where $\theta_1, \cdots, \theta_M$ are $M$ MCMC samples drawn from $\pi(\theta|\boldsymbol{D})$, the posterior distribution of $\theta$ given the original data $\boldsymbol{D}$. Importantly, under regularity conditions, $\widehat{\theta}_{\text{is}}^{(b)} \to \widehat{\theta}^{(b)}$ almost surely as $M \to \infty$.

**Proposition 2.** *Under regularity conditions, $\widehat{\theta}_{\text{is}}^{(b)} \to \widehat{\theta}^{(b)}$ with probability 1.*

*Proof.* Suppose that $\omega^{(b)}(\theta)$ and $\theta\omega^{(b)}(\theta)$ are integrable functions of $\theta$ with respect to the posterior distribution of the original data $\pi(\theta|\boldsymbol{D})$ so that $G_s^{(b)} = \int \theta^s \omega^{(b)}(\theta)\pi(\theta|\boldsymbol{D})d\theta/K_\pi = E_{\pi(\cdot|\boldsymbol{D})}\{\theta^s\omega^{(b)}(\theta)\}/K_\pi$ is finite for all $b$ and $s = 0, 1$. Therefore, as $M \to \infty$, from the ergodic theorem (Jones, 2004; Robert and Casella, 2005), with probability 1,

$$\begin{aligned}
\frac{1}{M}\sum_{j=1}^M \omega^{(b)}(\theta_j) &\to E_{\pi(\cdot|\boldsymbol{D})}\{\omega^{(b)}(\theta)\} = K_\pi G_0^{(b)}, \\
\frac{1}{M}\sum_{j=1}^M \theta_j\omega^{(b)}(\theta_j) &\to E_{\pi(\cdot|\boldsymbol{D})}\{\theta\omega^{(b)}(\theta)\} = K_\pi G_1^{(b)}.
\end{aligned}$$

From Remark 3 in Section 4.3, $\omega^{(b)}(\theta) = \exp\{\ell^{(b)}(\theta) - \ell(\theta)\}$ implies $\omega^{(b)}(\theta)$ is positive for all $\theta$. Therefore, $\sum_{j=1}^M \omega^{(b)}(\theta_j) > 0$ and $G_0^{(b)} > 0$, and consequently

$$\widehat{\theta}_{\text{is}}^{(b)} = \frac{\sum_{j=1}^M \theta_j \omega^{(b)}(\theta_j)}{\sum_{j=1}^M \omega^{(b)}(\theta_j)} \to \frac{G_1^{(b)}}{G_0^{(b)}} = \widehat{\theta}^{(b)}$$

with probability 1 as $M \to \infty$. □

Next define $\overline{\theta}_{\text{is}}^{(\cdot)} = B^{-1}\sum_{b=1}^{B}\widehat{\theta}_{\text{is}}^{(b)}$. As $M$ gets large,

$$(B-1)^{-1}\sum_{b=1}^{B}(\widehat{\theta}_{\text{is}}^{(b)} - \overline{\theta}_{\text{is}}^{(\cdot)})^2 \to (B-1)^{-1}\sum_{b=1}^{B}(\widehat{\theta}^{(b)} - \overline{\theta}^{(\cdot)})^2.$$

Hence we use $\sum_{b=1}^{B}(\widehat{\theta}_{\text{is}}^{(b)} - \overline{\theta}_{\text{is}}^{(\cdot)})^2/(B-1)$ to estimate $\text{var}_{\widehat{F}}(\widehat{\theta})$. The above procedure can be summarized in the following steps.

Step 1. Draw $M$ MCMC samples from $\pi(\theta|\boldsymbol{D})$, and call them $(\theta_1, \cdots, \theta_M)$.

Step 2. Draw $B$ bootstrap samples with replacement from $\boldsymbol{D}$, and each bootstrap sample consists of $n$ observations. For the $b^{\text{th}}$ sample we obtain $(r_1^{(b)}, \cdots, r_n^{(b)})$, with $0 \le r_i^{(b)} \le n$ and $\sum_{i=1}^{n} r_i^{(b)} = n$, where $r_i^{(b)}$ is the number of times $X_i$ appears in the $b^{\text{th}}$ bootstrap sample, $b = 1, \ldots, B$.

Step 3. Compute $\widehat{\theta}_{\text{is}}^{(b)} = \sum_{j=1}^{M}\theta_j\omega^{(b)}(\theta_j)/\sum_{j=1}^{M}\omega^{(b)}(\theta_j)$ with $\omega^{(b)}(\theta_j) = \prod_{i=1}^{n} f^{(r_i^{(b)}-1)}(X_i|\theta_j)$ for $b = 1, \ldots, B$, and $\overline{\theta}_{\text{is}}^{(\cdot)} = \sum_{b=1}^{B}\widehat{\theta}_{\text{is}}^{(b)}/B$.

Step 4. Compute $(B-1)^{-1}\sum_{b=1}^{B}(\widehat{\theta}_{\text{is}}^{(b)} - \overline{\theta}_{\text{is}}^{(\cdot)})^2$.

One of the main concerns of importance sampling is the behavior of the importance weights that have influence on the efficiency of the estimator. The following remark gives an intuitive justification that our choice $\pi(\theta|\boldsymbol{D})$ as the trial distribution provides a bounded importance weight with high probability.

**Remark 3.** *Note that* $\omega^{(b)}(\theta) = \exp[\sum_{i=1}^{n}(r_i^{(b)} - 1)\log\{f(X_i|\theta)\}] = \exp\{\ell^{(b)}(\theta) - \ell(\theta)\}$*, where* $\ell^{(b)}(\theta) = \sum_{i=1}^{n} r_i^{(b)}\log f(X_i|\theta) + \log\{\pi(\theta)\}$ *and* $\ell(\theta) = \sum_{i=1}^{n}\log\{f(X_i|\theta)\} + \log\{\pi(\theta)\}$*, and* $\theta$ *is drawn from the posterior distribution* $\pi(\theta|\boldsymbol{D})$*. Now,*

$$\ell^{(b)}(\theta) - \ell(\widetilde{\theta}) \le \ell^{(b)}(\theta) - \ell(\theta) \le \ell^{(b)}(\widetilde{\theta}_{(b)}) - \ell(\theta),$$

*where* $\widetilde{\theta}_{(b)}$ *is the posterior mode based on the* $b^{\text{th}}$ *bootstrap data set and* $\widetilde{\theta}$ *is the posterior mode based on the original data. Then under certain regularity conditions, posterior distribution* $\pi(\theta|\boldsymbol{D})$ *has the asymptotic normal distribution having mean* $\widetilde{\theta}$ *and the variance is minus the inverse Hessian of the log posterior evaluated at* $\widetilde{\theta}$ *for large* $n$ *(Theorem 3.1 of Carlin and Louis, 2008).*

## 4.4 Other Bayes estimators

### 4.4.1 Posterior quantile

Here we broadly discuss the standard error calculation of posterior quantiles that include the posterior median and credible intervals as special cases. The $\alpha^{\text{th}}$ quantile is defined as $q_\alpha(\boldsymbol{D}) = F^{-1}_{\pi(\theta|\boldsymbol{D})}(\alpha)$, where $F_{\pi(\theta|\boldsymbol{D})}(r) = \int_{-\infty}^{r} \pi(\theta|\boldsymbol{D})d\theta$. To estimate the frequentist standard error of $q_\alpha(\boldsymbol{D})$, we may apply the regular bootstrap method by calculating the $\alpha^{\text{th}}$ quantile for each of the $B$ posterior distributions, that means one needs to draw posterior samples from $\pi(\theta|\boldsymbol{D}^{(b)})$ using MCMC technique for each $b = 1, \ldots, B$. Instead of doing this for multiple bootstrap data sets, here we can also apply the importance sampling idea. For a trial density $h(\theta)$, we have

$$\begin{aligned} F_{\pi(\theta|\boldsymbol{D}^{(b)})}(r) = \int_{-\infty}^{\infty} I(\theta \le r)\pi(\theta|\boldsymbol{D}^{(b)})d\theta &= \int_{-\infty}^{\infty} I(\theta \le r)\frac{\pi(\theta|\boldsymbol{D}^{(b)})}{h(\theta)}h(\theta)d\theta \\ &= \int_{-\infty}^{\infty} I(\theta \le r)\omega^{(b)}(\theta)h(\theta)d\theta, \end{aligned}$$

where $\omega^{(b)}(\theta) = \pi(\theta|\boldsymbol{D}^{(b)})/h(\theta)$. The distribution function can be estimated by

$$\widehat{F}_{\pi(\theta|\boldsymbol{D}^{(b)})}(r) = \frac{\sum_{j=1}^{M} I(\theta_j \le r)\omega^{(b)}(\theta_j)}{\sum_{j=1}^{M} \omega^{(b)}(\theta_j)}, \tag{4.1}$$

where $\theta_1, \ldots, \theta_M$ are drawn from $h(\theta)$. We shall evaluate $\widehat{F}_{\pi(\theta|\boldsymbol{D}^{(b)})}(r)$ for a grid of values of $r$. Next, the estimated $\alpha^{\text{th}}$ quantile is defined as $q^{(b)}_{\alpha,\text{is}} = \inf\{r : \widehat{F}_{\pi(\theta|\boldsymbol{D}^{(b)})}(r) \ge \alpha\}$. Note that we shall use the same set of $\theta_1, \ldots, \theta_M$ drawn from $h(\theta)$, for each bootstrap data set thereby saving considerable computation time.

When $\alpha$ takes a moderate value in the range of $0.2$ to $0.8$, the importance sampling estimates are reasonable if $\pi(\theta|\boldsymbol{D})$ is used as the trial distribution. For more extreme values of $\alpha$, (smaller than 0.2 or larger than 0.8), we recommend the following trial distribution for efficient estimation of the $\alpha^{\text{th}}$ quantile. To be more specific, without any loss of generality, write $\theta = (\theta_1, \theta_2^T)^T$, and suppose that we are interested in estimating the $\alpha^{\text{th}}$ quantile of $\theta_1$ based on the $b^{\text{th}}$ bootstrap data. Take $h(\theta) = h_1(\theta_1)h_2(\theta_2)$, where $h_1$ denotes the uniform density over $[l, u]$ for given values of $l$

and $u$, and $h_2$ is taken as the posterior distribution of $\theta_2$ given the data $\boldsymbol{D}$, that means, $h_2(\theta_2) = \int \pi(\theta|\boldsymbol{D})d\theta_1$. Although there is no optimum choice of $l$ or $u$, based on our computing experiences, we recommend $l = q_{0.5}(\boldsymbol{D}) - 6 \times sd_{\theta_1}(\boldsymbol{D})$ and $u = q_{0.5}(\boldsymbol{D}) + 6 \times sd_{\theta_1}(\boldsymbol{D})$, where $q_\alpha(\boldsymbol{D})$ and $sd_{\theta_1}(\boldsymbol{D})$ denote the $\alpha^{\text{th}}$ quantile and the posterior standard deviation of $\theta_1$ given the entire data $\boldsymbol{D}$.

Suppose that $(\theta_{11}, \ldots, \theta_{1M})$ are $M$ random draws from $h_1(\theta_1)$, and $(\theta_{21}, \ldots, \theta_{2M})$ are $M$ random draws from $\pi(\theta_2|\boldsymbol{D})$. The later sample is obtained by simply discarding the first component from each of the $M$ MCMC samples drawn from $\pi(\theta|\boldsymbol{D}) \equiv \pi(\theta_1, \theta_2|\boldsymbol{D})$. Computation of the importance weight $\omega^{(b)}(\theta)$ at $\theta = \theta_j = (\theta_{1j}, \theta_{2j}^T)^T$, for any $j = 1, \ldots, M$, requires $h_2(\theta_{2j}) = \int \pi(\theta_1^*, \theta_{2j}^T|\boldsymbol{D})d\theta_1^* = \kappa^{-1} \int \prod_{i=1}^n f(X_i|\theta_1^*, \theta_{2j}^T)\pi(\theta_1^*, \theta_{2j}^T)d\theta_1^*$, where $\kappa$ is the normalizing constant that does not depend on $\theta_j$. In order to save computation time, instead of targeting to evaluate $h_2(\theta_{2j})$ separately, we consider directly evaluating $\omega^{(b)}(\theta_j)$, and

$$\begin{aligned}
\omega^{(b)}(\theta_j) &= \frac{\pi(\theta_{1j}, \theta_{2j}^T|\boldsymbol{D}^{(b)})}{h_1(\theta_{1j}) \int_{\theta_{1,\min}-\varepsilon}^{\theta_{1,\max}+\varepsilon} \pi(\theta_1^*, \theta_{2j}^T|\boldsymbol{D})d\theta_1^*} \\
&= \frac{\kappa_b^{-1} \prod_{i=1}^n f^{r_i^{(b)}}(X_i|\theta_{1j}, \theta_{2j}^T)\pi(\theta_{1j}, \theta_{2j}^T)}{h_1(\theta_{1j})\kappa^{-1} \int_{\theta_{1,\min}-\varepsilon}^{\theta_{1,\max}+\varepsilon} \prod_{i=1}^n f(X_i|\theta_1^*, \theta_{2j}^T)\pi(\theta_1^*, \theta_{2j}^T)d\theta_1^*} \\
&= \left[ h_1(\theta_{1j}) \frac{\kappa^{-1}}{\kappa_b^{-1}} \int_{\theta_{1,\min}-\varepsilon}^{\theta_{1,\max}+\varepsilon} \left\{ \prod_{i=1}^n \frac{f(X_i|\theta_1^*, \theta_{2j}^T)}{f^{r_i^{(b)}}(X_i|\theta_{1j}, \theta_{2j}^T)} \right\} \left\{ \frac{\pi(\theta_1^*, \theta_{2j}^T)}{\pi(\theta_{1j}, \theta_{2j}^T)} \right\} d\theta_1^* \right]^{-1}, \quad (4.2)
\end{aligned}$$

where $\kappa_b$ is the normalizing constant for the $b^{\text{th}}$ bootstrap data $\boldsymbol{D}^{(b)}$, and $\theta_{1,\min}$ and $\theta_{1,\max}$ denote the observed minimum and maximum values of $\theta_1$ in the posterior samples drawn from $\pi(\theta_1, \theta_2|\boldsymbol{D})$. To cover the entire domain of $\theta_1$, we extend the range of the integration by adding and subtracting a small number $\varepsilon > 0$. In all our computations, we used $\varepsilon = 0.1 \times \text{IQR}$, where IQR stands for the inter quartile range of the posterior distribution of $\theta_1$ given the original data $\boldsymbol{D}$. Importantly, we do not need to evaluate $\kappa$ and $\kappa_b$ for estimating $F_{\pi(\theta|\boldsymbol{D}^{(b)})}(r)$ as they are independent of $\theta_j$, so they get canceled from the normalized weight. Finally, we recommend to use Gauss-Legendre quadrature to determine the above integral in (4.2). Also to reduce the computational burden, once $\omega^{(b)}(\theta_j)$ is calculated for some $b$, then we compute $\omega^{(b')}(\theta_j)$ using the following formula

$\omega^{(b')}(\theta_j) = \omega^{(b)}(\theta_j)\pi(\theta_{1j}, \theta_{2j}^T|\boldsymbol{D}^{(b')})/\pi(\theta_{1j}, \theta_{2j}^T|\boldsymbol{D}^{(b)})$, for any $b' \neq b$ as

$$\begin{aligned}\omega^{(b')}(\theta_j) &= \frac{\pi(\theta_{1j}, \theta_{2j}^T|\boldsymbol{D}^{(b')})}{h_1(\theta_{1j})\int_{\theta_{1,\min}-\varepsilon}^{\theta_{1,\max}+\varepsilon}\pi(\theta_1^*, \theta_{2j}^T|\boldsymbol{D})d\theta_1^*} = \underbrace{\frac{\pi(\theta_{1j}, \theta_{2j}^T|\boldsymbol{D}^{(b)})}{h_1(\theta_{1j})\int_{\theta_{1,\min}-\varepsilon}^{\theta_{1,\max}+\varepsilon}\pi(\theta_1^*, \theta_{2j}^T|\boldsymbol{D})d\theta_1^*}}_{\omega^{(b)}(\theta_j)} \\ &\times \frac{\pi(\theta_{1j}, \theta_{2j}^T|\boldsymbol{D}^{(b')})}{\pi(\theta_{1j}, \theta_{2j}^T|\boldsymbol{D}^{(b)})}.\end{aligned}$$

### 4.4.2 Posterior mode

Here we do not apply the importance sampling idea but use another approach for time efficient computation. The posterior mode is defined as $\widehat{\theta}_{\text{mode}} = \arg\max_\theta \pi(\theta|\boldsymbol{D})$. The variance of $\widehat{\theta}_{\text{mode}}$, $\text{var}_F(\widehat{\theta}_{\text{mode}})$ can be estimated by $\sum_{b=1}^B(\widehat{\theta}_{\text{mode}}^{(b)} - \overline{\theta}_{\text{mode}}^{(\cdot)})^2/(B-1)$, where $\widehat{\theta}_{\text{mode}}^{(b)}$ denotes the posterior mode for the $b^{\text{th}}$ bootstrap sample, and $\overline{\theta}_{\text{mode}}^{(\cdot)} = \sum_{b=1}^B \widehat{\theta}_{\text{mode}}^{(b)}/B$. Since this standard bootstrap method could be time consuming as it requires to solve a set of gradient equations for each of the $B$ bootstrap data sets, we propose the following alternative approach of estimating that variance.

Under sufficient smoothness conditions, $\widehat{\theta}_{\text{mode}}$ will satisfy $S(\widehat{\theta}_{\text{mode}}|\boldsymbol{D}) = 0$, where $S(\theta|\boldsymbol{D}) = \partial\text{log}\{\pi(\theta|\boldsymbol{D})\}/\partial\theta = \partial\text{log}\{f(\boldsymbol{D}|\theta)\}/\partial\theta + \partial\text{log}\{\pi(\theta)\}/\partial\theta = 0$. Suppose that as $n \to \infty$, $\widehat{\theta}_{\text{mode}} \to \theta_{\text{mode}}$. Then

$$\begin{aligned}0 = S(\widehat{\theta}_{\text{mode}}|\boldsymbol{D}) &= \frac{\partial}{\partial\theta}\text{log}\{f(\boldsymbol{D}|\widehat{\theta}_{\text{mode}})\} + \frac{\partial}{\partial\theta}\text{log}\{\pi(\widehat{\theta}_{\text{mode}})\} \\ &\approx [\frac{\partial}{\partial\theta}\text{log}\{f(\boldsymbol{D}|\theta_{\text{mode}})\} + \frac{\partial}{\partial\theta}\text{log}\{\pi(\theta_{\text{mode}})\}] + \\ &\quad [\frac{\partial^2}{\partial\theta^2}\text{log}\{f(\boldsymbol{D}|\theta_{\text{mode}})\} + \frac{\partial^2}{\partial\theta^2}\text{log}\{\pi(\theta_{\text{mode}})\}](\widehat{\theta}_{\text{mode}} - \theta_{\text{mode}}).\end{aligned}$$

Thus, with $A = E[\partial^2\text{log}\{f(\boldsymbol{D}|\theta_{\text{mode}})\}/\partial\theta^2 + \partial^2\text{log}\{\pi(\theta_{\text{mode}})\}/\partial\theta^2]$, we have $(\widehat{\theta}_{\text{mode}} - \theta_{\text{mode}}) \approx A^{-1}[\partial\text{log}\{f(\boldsymbol{D}|\theta_{\text{mode}})\}/\partial\theta + \partial\text{log}\{\pi(\theta_{\text{mode}})\}/\partial\theta]$, and consequently the variance can be obtained by the sandwich formula,

$$\text{var}_F(\widehat{\theta}_{\text{mode}}) = A^{-1}\text{var}[\frac{\partial}{\partial\theta}\text{log}\{f(\boldsymbol{D}|\theta_{\text{mode}})\} + \frac{\partial}{\partial\theta}\text{log}\{\pi(\theta_{\text{mode}})\}]A^{-T}.$$

Here $A$ can be estimated by $\widehat{A} = \partial^2\text{log}\{f(\boldsymbol{D}|\widehat{\theta}_{\text{mode}})\}/\partial\theta^2 + \partial^2\text{log}\{\pi(\widehat{\theta}_{\text{mode}})\}/\partial\theta^2$. The middle term of the variance formula is var$[\partial\text{log}\{f(\boldsymbol{D}|\theta_{\text{mode}})\}/\partial\theta]$ that can be estimated by

$$\begin{aligned}\widehat{\text{var}}[\frac{\partial}{\partial\theta}\text{log}\{f(\boldsymbol{D}|\theta_{\text{mode}})\}] &= (B-1)^{-1}\sum_{b=1}^{B}\Bigg[\frac{\partial}{\partial\theta}\text{log}\{f(\boldsymbol{D}^{(b)}|\widehat{\theta}_{\text{mode}})\} \\ &\qquad \frac{1}{B}\sum_{b'=1}^{B}\frac{\partial}{\partial\theta}\text{log}\{f(\boldsymbol{D}^{(b')}|\widehat{\theta}_{\text{mode}})\}\Bigg]^2,\end{aligned}$$

and in particular, for fast computation we use $\partial\text{log}\{f(\boldsymbol{D}^{(b)}|\widehat{\theta}_{\text{mode}})\}/\partial\theta = \sum_{i=1}^{n} r_i^{(b)}\partial\text{log}\{f(X_i$ $|\widehat{\theta}_{\text{mode}})\}/\partial\theta$. Finally, var$_F(\widehat{\theta}_{\text{mode}})$ is estimated by $\widehat{A}^{-1}\widehat{\text{var}}[\partial\text{log}\{f(\boldsymbol{D}|\theta_{\text{mode}})\}/\partial\theta]\widehat{A}^{-T}$.

### 4.5 Simulation studies

In order to assess and compare the performances of the methods, we conducted simulation studies and real data analysis for the motivating examples described in Section 4.2. Specifically, we provide simulation results for the logistic regression model. Next, the linear measurement error model is illustrated using a simulated data set. Third, we present an analysis of real data set using the Weibull regression model. Finally, I consider an application of the proposed method to a vector autoregressive (VAR) model.

#### 4.5.1 Logistic regression model

I generated $500$ data sets, and each simulated data set consists of $n = 500$ observations, denoted by $\{(X_i, Y_i), i = 1, \ldots, n\}$. We drew $X$ from $\text{Normal}(0, 1)$ distribution and the response variable $Y$ was simulated from a Bernoulli distribution with the success probability pr$(Y = 1|X) = \exp(\alpha + \beta X)/\{1 + \exp(\alpha + \beta X)\}$. The true values of $\alpha$ and $\beta$ were $-2.5$ and $1$, respectively. That makes the proportion of success around $10\%$. For the Bayesian inference of the parameters $\alpha$ and $\beta$ we used the same $\text{Normal}(0, 2)$ priors for both of them. Then for the MCMC computation, we used $15,000$ iterations with the first $5,000$ samples were used as burn-in samples.

For each data set, we estimated the posterior mean of $\alpha$ and $\beta$. We also calculated standard errors of the posterior means for each data set. Let $\widehat{\alpha}_j$ and $\widehat{\beta}_j$ be the posterior mean based on the $j^{\text{th}}$ data set, for $j = 1, \ldots, 500$. For each data set, we computed the frequentist standard error of the estimator based on 1) the regular bootstrap method and 2) the proposed importance sampling based approach. For the $j^{\text{th}}$ data set, we drew $B = 500$ bootstrap samples with replacement. Suppose that $(\widehat{\alpha}_{mcmc,j}^{(b)}, \widehat{\beta}_{mcmc,j}^{(b)})$ denotes the posterior means for the $b^{\text{th}}$ bootstrap data, for $b = 1, \ldots, 500$, and these posterior means were calculated by applying the MCMC method to each bootstrap data separately. The regular bootstrap standard error for $\widehat{\alpha}_j$ and $\widehat{\beta}_j$ are now expressed as $sd_{1,j}(\alpha) = \sqrt{(1/499)\sum_{b=1}^{500}(\widehat{\alpha}_{mcmc,j}^{(b)} - \sum_{b'=1}^{500}\widehat{\alpha}_{mcmc,j}^{(b')}/500)^2}$ and $sd_{1,j}(\beta) = \sqrt{(1/499)\sum_{b=1}^{500}(\widehat{\beta}_{mcmc,j}^{(b)} - \sum_{b'=1}^{500}\widehat{\beta}_{mcmc,j}^{(b')}/500)^2}$, respectively. Next, we computed the proposed importance sampling based standard error, $sd_{2,j}(\alpha) = \sqrt{(1/499)\sum_{b=1}^{500}(\widehat{\alpha}_{is,j}^{(b)} - \sum_{b'=1}^{500}\widehat{\alpha}_{is,j}^{(b')}/500)^2}$ and $sd_{2,j}(\beta) = \sqrt{(1/499)\sum_{b=1}^{500}(\widehat{\beta}_{is,j}^{(b)} - \sum_{b'=1}^{500}\widehat{\beta}_{is,j}^{(b')}/500)^2}$, where $(\widehat{\alpha}_{is,j}^{(b)}, \widehat{\beta}_{is,j}^{(b)})$ denotes the posterior means for the $b^{\text{th}}$ bootstrap data based on the importance sampling idea. Our goal is to illustrate that instead of using the regular bootstrap idea that is way more time consuming, one can simply use the importance sampling based method to estimate the frequentist standard error of the Bayes estimators. We wanted to show that proposed method is computationally far more time efficient, and on the other hand, the standard error calculated using the proposed method is close to the standard error calculated based on the regular bootstrap method. We, once again, point out that the regular bootstrap approach requires enumeration of $B$ MCMC chains, one for each of the $B$ bootstrap data sets, while the proposed approach requires enumeration of only one MCMC chain. In the appendix, we compare the computational complexity of the two approaches.

Figure 4.1 shows a scatter plot of two standard errors ($sd_1$ and $sd_2$) for $500$ data sets for the intercept and slope parameter. The figure reveals that the two estimates of the standard error are in good agreement as the points are well dispersed around the $45$ degree line. Table 1 shows the computation time (in sec) for the two methods, and clearly the proposed importance sampling based approach is computationally far more superior than the regular bootstrap method.

Since the logistic regression belongs to the class of the generalized linear models, we are able to apply Efron (2015)'s method to evaluate the standard deviation of the posterior mean for the intercept and slope parameters. Let $\theta = (\alpha, \beta)^T$. From Equation (3.1) of Efron (2015),

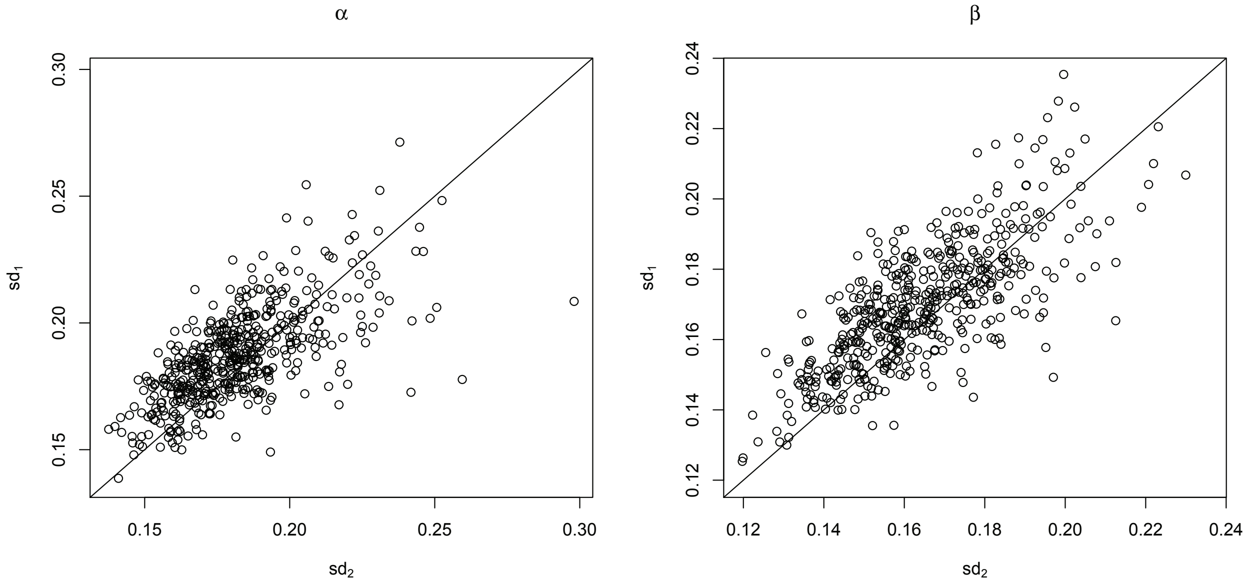

Figure 4.1: Frequentist standard errors of posterior means of the intercept ($\alpha$) and the slope ($\beta$) of the logistic regression model from the 500 simulated data sets in Section 4.5.1 based on the regular bootstrap method (Y-axis) and the proposed importance sampling based method (X-axis).

Table 4.1: Average computing time (± standard deviation of 500 simulated data sets) measured in seconds for calculating standard errors of posterior summaries in logistic regression model from Section 4.5.1 based on the 1) regular bootstrap method, 2) the importance sampling based approach, and 3) the method proposed in Efron (2015). Here $Q_1$ and $Q_3$ denote the first and third quartiles, and 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ denote the 2.5$^{\text{th}}$ percentile and 97.5$^{\text{th}}$ percentile of the posterior distribution, respectively.

| Method | Time to calculate | | | Computational |
|---|---|---|---|---|
| | Mean | $Q_1$ ($Q_3$) | 2.5$^{\text{th}}$ (97.5$^{\text{th}}$) | complexity |
| 1 | 247.78 ± 6.70 | 247.78 ± 6.70 | 247.78 ± 6.70 | $O(BMn)$ |
| 2 | 46.65 ± 0.41 | 51.66 ± 0.40 | 120.44 ± 0.93 | $O(BMn)$ |
| 3 | 4.18 ± 0.6 | | | $O(Mn)$ |

$f_\theta(T) = \exp[\theta^T T - \sum_{j=1}^n \log\{1 + \exp(\alpha + \beta X_j)\}]$, where $T = (\sum_{j=1}^n Y_j, \sum_{j=1}^n X_j Y_j)^T$ is the sufficient statistic for $\theta$. Then, $E(T) = (\sum_{j=1}^n p_j, \sum_{j=1}^n X_j p_j)^T$ and $\text{var}(T) = V_\theta = \sum_{j=1}^n p_j(1-p_j)(1, X_j)^T(1, X_j)$, where $p_j = \text{P}(Y = 1|X_j) = \exp(\alpha+\beta X_j)/\{1+\exp(\alpha+\beta X_j)\}$,

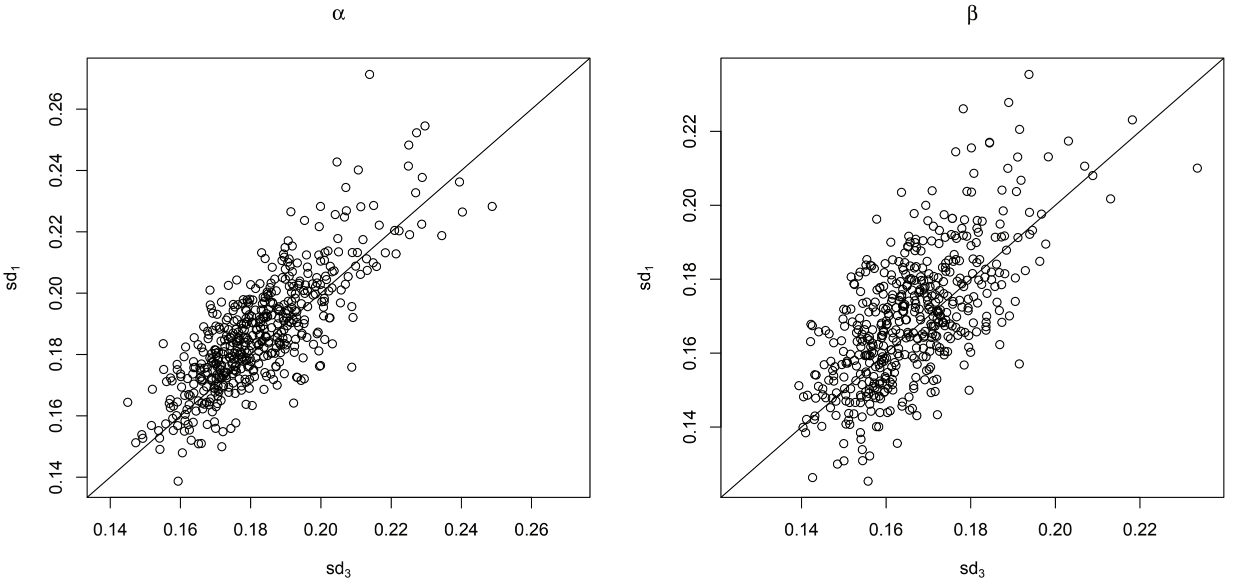

Figure 4.2: Frequentist standard errors of posterior means of the intercept ($\alpha$) and the slope ($\beta$) of the logistic regression model from the 500 simulated data sets in Section 4.5.1 based on the regular bootstrap method (Y-axis) and the approach proposed in (Efron, 2015) (X-axis).

the success probability given $X = X_j$. Due to the numerical instability of the "conversion factors" we are not able to apply his method that completely avoids MCMC sampling, and this issue has been acknowledged in Efron (2015). However, we apply his general approach for calculating the standard deviation of the posterior mean that is summarized in the following steps.

Step 1. Draw $M$ MCMC samples $(\theta_1, \ldots, \theta_M)$ from $\pi(\theta|\boldsymbol{D})$.

Step 2. Estimate cov$(\theta, \theta|T)$ by $\widehat{\text{cov}} = \sum_{j=1}^{M}(\theta_j - \bar{\theta})(\theta_j - \bar{\theta})^T/M$, where $\bar{\theta} = \sum_{j=1}^{M}\theta_j/M$. Then we obtain $sd_3 = [\widehat{\text{cov}}^T V_{\widehat{\theta}} \widehat{\text{cov}}]^{(1/2)}$.

Now we compare $sd_3$ with the gold standard approach, $sd_1$, in Figure 4.2. In terms of computation time, Efron's approach is much much faster than any other procedure (Table 1). However, Efron's approach is applicable when $V_\theta$ is easily available, and his method can compute standard error for posterior mean only, not for any quantiles.

Next we calculated the standard error of the first quartile, third quartile, the $2.5^{\text{th}}$ percentile, and the $97.5^{\text{th}}$ percentile of the posterior distribution of $\alpha$ and $\beta$ based on 1) the regular bootstrap

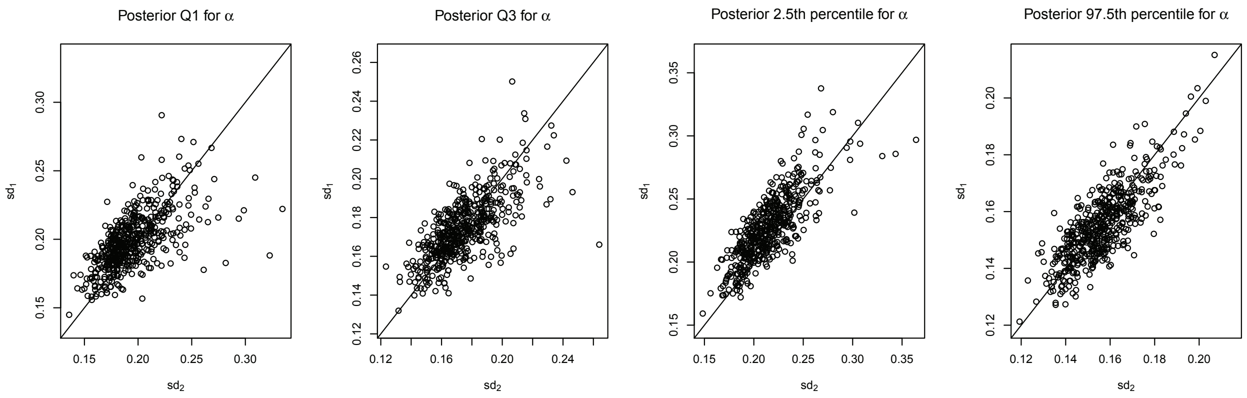

Figure 4.3: Frequentist standard errors of $Q_1$, $Q_3$, $2.5^{th}$ percentile, and $97.5^{th}$ percentile of the posterior distribution of $\alpha$ in the logistic regression model from the 500 simulated data set in Section 4.5.1. Regular bootstrap standard errors are presented along the Y-axis while importance sampling based standard errors are presented along the X-axis.

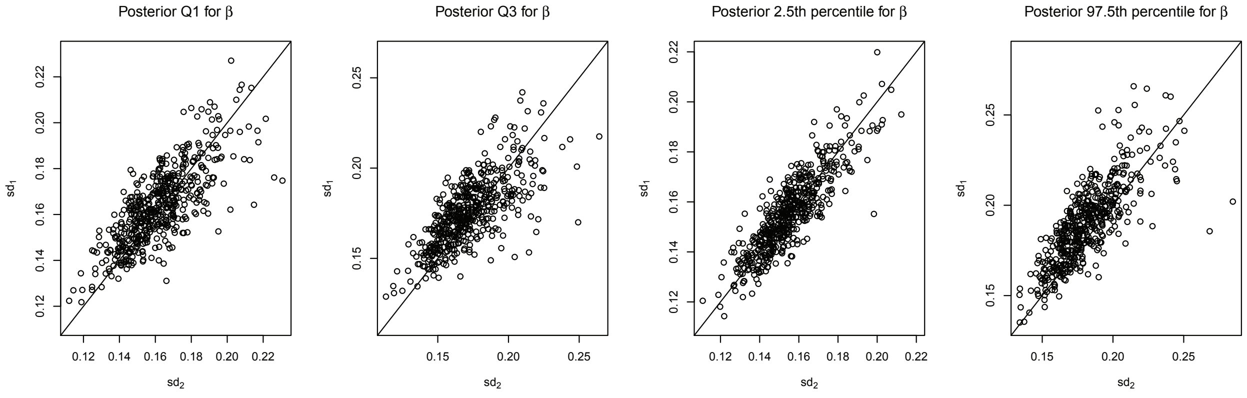

Figure 4.4: Frequentist standard errors of $Q_1$, $Q_3$, $2.5^{th}$ percentile, and $97.5^{th}$ percentile of the posterior distribution of $\beta$ in the logistic regression model from the 500 simulated data set in Section 4.5.1. Regular bootstrap standard errors are presented along the Y-axis while importance sampling based standard errors are presented along the X-axis.

method and 2) the importance sampling based method. We particularly considered $2.5^{\text{th}}$ and $97.5^{\text{th}}$ percentiles as they are often used for constructing credible intervals. Figures 4.3 and 4.4 show the standard errors computed using the two approaches for each of these summary statistics for the simulated data sets. We want to point out that for the $2.5^{\text{th}}$ and $97.5^{\text{th}}$ percentiles we used the trial distribution that is described in Section 4.4.1 and it involves with a slightly more computation than the scenario where $\pi(\alpha, \beta|\boldsymbol{D})$ is used as a trial distribution (see Table 4.1). However, despite of being more computationally involved, overall this approach is more time efficient (see Table 4.1) than the regular bootstrap method where one needs to run MCMC method on each bootstrap data set separately. We also need to keep in mind that this time comparison is heavily depended on the number of MCMC iterations used in the computation, and the time gain will be more if more MCMC iterations are used for the posterior inference. For a fair comparison, every core computation was conducted using FORTRAN 90 within an R script. That is, generation of random samples from the posterior distribution $\pi(\theta|\boldsymbol{D})$ and evaluation of the importance weight $\omega^{(b)}(\theta)$ in Sections 4.3, 4.4.1 were programmed in FORTRAN. Although there are a number of presumably optimized programs or R packages for Bayesian computing, we decide to write our own code for fair comparison across the methods.

The computational complexity of the proposed method and the regular bootstrap method using MCMC simulations are of the same order, and according to the Bachman-Landau notation it is $O(BMn)$, where $B$, $M$, $n$ denote the number of bootstrap samples, the number of MCMC iterations, and the sample size, respectively. In Appendix B, we have explained the computational complexity for this example through algorithms, and similar algorithms can be written for other examples. Although the computational complexity of the regular bootstrap method and the proposed method are of the same order, by avoiding MCMC simulations the computation of posterior summary is much faster in the latter method than the former approach.

### 4.5.2 Linear measurement error model

Next, we revisit the linear measurement error model. We first note that the joint distribution of the observed $Y$ and $W$, $f_{Y,W}(y, w)$ is an exponential family. Since $f_{Y,W}(y, w) = \int f(w, x, y)dx$, where $f(y, w, x)$ is the joint density of $W, X, Y$,

$$\begin{aligned} f_{Y,W}(Y, W) &= h(Y, W)c(\theta)\exp\Bigg[ -\frac{1}{2}\Big\{\frac{1}{\sigma_\epsilon^2} - \frac{\beta^2/\sigma_\epsilon^4}{\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2}\Big\}Y^2 \\ &\qquad -\Big\{\frac{\alpha\beta^2/\sigma_\epsilon^4 + \beta\mu_x/(\sigma_\epsilon^2\sigma_x^2)}{\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2} - \frac{\alpha}{\sigma_\epsilon^2}\Big\}Y + \frac{1/\sigma_u^4}{2(\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2)}W^2 \\ &\qquad -\frac{\alpha\beta/(\sigma_\epsilon^2\sigma_u^2) - \mu_x/(\sigma_u^2\sigma_x^2)}{\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2}W + \frac{\beta/(\sigma_\epsilon^2\sigma_u^2)}{\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2}YW\Bigg], \end{aligned}$$

where $h(Y, W) = \exp(-W^2/2\sigma_u^2)$ does not depend on $\theta$ since $\sigma_u^2$ is known and $c(\theta) = (2\pi)^{-1}\{\sigma_\epsilon^2\sigma_u^2\sigma_x^2(\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2)\}^{-1/2}\exp\{-\alpha^2/2\sigma_\epsilon^2 - \mu_x^2/2\sigma_x^2 + (\alpha^2\beta^2/2\sigma_\epsilon^4 + \mu_x^2/2\sigma_x^4 - \alpha\beta\mu_x/2\sigma_x^2\sigma_\epsilon^2)/(\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2)\}$ is a function of $\theta$. Therefore, $T = (Y^2, Y, W^2, W, YW)$ is a sufficient statistic for the natural parameter $\eta = (\eta_1, \ldots, \eta_5)$, where $\eta_1 = 1/\sigma_\epsilon^2 - (\beta^2/\sigma_\epsilon^4)/(\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2)$, $\eta_2 = \{\alpha\beta^2/\sigma_\epsilon^4 + \beta\mu_x/(\sigma_\epsilon^2\sigma_x^2)\}/(\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2) - \alpha/\sigma_\epsilon^2$, $\eta_3 = 1/\sigma_u^2 - (1/\sigma_u^4)/(\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2)$ $\eta_4 = \{\alpha\beta/(\sigma_\epsilon^2\sigma_u^2) - \mu_x/(\sigma_u^2\sigma_x^2)\}/(\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2)$, and $\eta_5 = \{\beta/(\sigma_\epsilon^2\sigma_u^2)\}/(\beta^2/\sigma_\epsilon^2 + 1/\sigma_u^2 + 1/\sigma_x^2)$. In order to apply Efron (2015)'s method, we need to find the variance covariance matrix $V_\eta$ of $T$, which is a very difficult if not impossible task. Therefore, we applied our approach to compute the frequentist standard error for the posterior summaries of $\alpha$ and $\beta$.

We generated a single data set comprising of $\boldsymbol{D} = \{(Y_i, W_i, X_i), i = 1, \ldots, n = 1,000\}$ under the true model $Y_i = \alpha + \beta X_i + \epsilon_i$, $\alpha = 0.23$, $\beta = 0.47$, and $W_i = X_i + U_i$, where $\epsilon_i \sim \text{Normal}[0, (\sqrt{0.5})^2]$, $U_i \sim \text{Normal}[0, (\sqrt{0.5})^2]$ and $X_i \sim \text{Normal}(0.5, 1)$. We analyzed the data according to the method described in Sections 4.3 and 4.4, without using $X$ in the analysis. We applied Gibbs sampling to draw samples from the posterior distribution of the parameters, and used $M = 10,000$ iterations after the first $5,000$ samples as burn-in samples. For the prior distributions, we set $\sigma_\alpha^2 = \sigma_\beta^2 = \sigma_\mu^2 = 10,000$ and $\delta_x = \delta_\epsilon = \lambda_x = \lambda_\epsilon = 1$. Then we drew $B = 500$ bootstrap samples with replacement and we evaluated $sd_1$ and $sd_2$ as described in Section 4.5.1. Table 4.2 shows the frequentist standard errors corresponding to the posterior summaries of $\alpha$ and $\beta$, along with the computation time. The results show the advantages of the proposed method over the regular bootstrap method in terms of computational time.

Table 4.2: The frequentist standard errors and computing times for $\alpha$ and $\beta$ of the linear measurement error model in Section 4.5.2. Here $sd_1$ and $sd_2$ denote the standard errors based on the regular bootstrap method and the importance sampling based approach.

| | | Posterior | | | |
|---|---|---|---|---|---|
| Parameter | | Mean | $Q_2$ | $2.5^{\text{th}}$ | $97.5^{\text{th}}$ |
| $\alpha$ | $sd_1$ | 0.028 | 0.027 | 0.028 | 0.027 |
| | $sd_2$ | 0.027 | 0.029 | 0.028 | 0.030 |
| $\beta$ | $sd_1$ | 0.034 | 0.034 | 0.033 | 0.036 |
| | $sd_2$ | 0.030 | 0.032 | 0.035 | 0.031 |
| Computation | $sd_1$ | 233.06 | 233.06 | 233.06 | 233.06 |
| time in second | $sd_2$ | 22.99 | 26.36 | 24.16 | 24.16 |
| Computational | $sd_1$ | $O(BMn)$ | $O(BMn)$ | $O(BMn)$ | $O(BMn)$ |
| complexity | $sd_2$ | $O(BMn)$ | $O(BMn)$ | $O(BMn)$ | $O(BMn)$ |

### 4.5.3 Weibull regression model

We now analyze a subset of the E1684 melanoma clinical trial data (Example 1.2 and 2.2 of Ibrahim et al., 2001) to determine the frequentist standard errors of posterior summaries from the Weibull model. This was a phase III clinical trial conducted by Eastern Cooperative Oncology Group (ECOG) with chemotherapy of interferon alpha-2b in melanoma patient and can be found at “`http://merlot.stat.uconn.edu/~mhchen/survbook/`". The data set contains observed time measured in year, (right) censoring indicator and chemotherapy treatment indicator for each of 255 patients. The purpose of this clinical study was to examine the treatment effect on the survival times ($Y$). Among the possible models for this objective, we fit a Weibull regression model on the survival times ($Y$) using chemotherapy as a covariate ($X$) according to Example 2.2 in Ibrahim et al. (2001). Following Ibrahim et al. (2001), we used a Gamma(1, 0.001) prior for $\alpha$ and a Normal($(0, 0)^T$, $10^4 I_2$) prior for $\beta$, where $I_2$ denotes the $2 \times 2$ identity matrix, for the Weibull regression model described in Section 4.2. Here we also generated $B = 500$ bootstrap data sets to calculate standard errors for the posterior summaries of parameters.

Table 4.3 shows the posterior estimates of $\beta_0$, $\beta_1$ and $\alpha$, corresponding frequentist standard errors, and computing times. Instead of presenting only posterior means as done in Table 2.2 of

Table 4.3: Posterior summaries and the corresponding frequentist standard errors of $\beta_0$, $\beta_1$, and $\alpha$ used in the Weibull model for analyzing the E1684 melanoma data given in Section 4.5.3. Here $sd_1$ and $sd_2$ denote the standard errors based on the regular bootstrap method and the importance sampling based approach.

| | | Posterior | | | |
|---|---|---|---|---|---|
| Parameter | | Mean | $Q_2$ | 2.5$^{\text{th}}$ | 97.5$^{\text{th}}$ |
| $\beta_0$ | | $-1.103$ | $-1.101$ | $-1.710$ | $-0.586$ |
| | $sd_1$ | 0.278 | 0.278 | 0.295 | 0.266 |
| | $sd_2$ | 0.255 | 0.265 | 0.252 | 0.261 |
| $\beta_1$ | | $-0.256$ | $-0.256$ | $-0.585$ | 0.090 |
| | $sd_1$ | 0.177 | 0.178 | 0.180 | 0.179 |
| | $sd_2$ | 0.169 | 0.176 | 0.163 | 0.183 |
| $\alpha$ | | 0.791 | 0.793 | 0.688 | 0.891 |
| | $sd_1$ | 0.038 | 0.039 | 0.035 | 0.043 |
| | $sd_2$ | 0.037 | 0.038 | 0.034 | 0.038 |
| Computation time in sec | $sd_1$ | 151.77 | 151.77 | 151.77 | 151.77 |
| | $sd_2$ | 37.31 | 43.59 | 85.75 | 85.75 |
| Computational complexity | $sd_1$ | $O(BMn)$ | $O(BMn)$ | $O(BMn)$ | $O(BMn)$ |
| | $sd_2$ | $O(BMn)$ | $O(BMn)$ | $O(BMn)$ | $O(BMn)$ |

Ibrahim et al. (2001), we extend that table to include other posterior summaries and the frequentist uncertainty of the estimates. Moreover, following the method described in Section 4.5, we are able to calculate the standard errors more time efficiently.

Furthermore, it is worth to note that it is difficult to apply Efron (2015)'s approach for calculating frequentist standard deviation of posterior mean to the Weibull model because it is not an exponential family of distributions. Secondly, the joint density of the above model is $f(\boldsymbol{D}|\alpha,\beta_0,\beta_1) = \exp[\sum_{i=1}^{n}\{\Delta_i\log\alpha + \Delta_i(\beta_0 + X_i\beta_1) + \Delta_i(\alpha-1)\log(Y_i) - Y_i^{\alpha}\exp(\beta_0 + X_i\beta_1)\}]$ so that it is also hard to calculate $V_\theta$ the variance of the sufficient statistic, where $\theta = (\alpha, \beta_0, \beta_1)$. Hence, we are not able to apply his method in this context.

### 4.5.4 Vector autoregressive model (VAR)

In the previous examples, we discussed the frequentist standard errors of posterior summaries for parameters themselves. We now discuss a more complicated case where the main interest is a function of parameters. Suppose that we have a $p$-dimensional time series data $\boldsymbol{y}_s, s = 1, \ldots, S$, and assume that the data follows a vector autoregression (VAR) model. The VAR model with lag $L$ is $\boldsymbol{y}_s' = \boldsymbol{\mu} + \sum_{j=1}^{L} \boldsymbol{y}_{s-j}' \boldsymbol{B}_j + \boldsymbol{\epsilon}_s'$, where $\boldsymbol{\mu}$ is an $1 \times p$ vector, $\boldsymbol{B}_j$ is a $p \times p$ coefficient matrix, $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_S$ are iid $N(0, \boldsymbol{\Sigma})$, and the covariance $\boldsymbol{\Sigma}$ is an unknown $p \times p$ positive definite matrix. Instead of focusing our attention on the elements of parameter matrices $\boldsymbol{B} = (\boldsymbol{B}_1', \ldots, \boldsymbol{B}_L')'$ and $\boldsymbol{\Sigma}$, it is more of interest to estimate the impact of changing an element of $\boldsymbol{y}_s$ on the future value $\boldsymbol{y}_{s+k}$. These effects are called impulse responses (Stock and Watson, 2001), and they are defined as nonlinear functions of the parameter matrices $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$.

The likelihood function of $(\boldsymbol{\mu}, \boldsymbol{B}, \boldsymbol{\Sigma})$ is

$$L(\boldsymbol{\Phi}, \boldsymbol{\Sigma}) = (2\pi)^{-Sp/2} |\boldsymbol{\Sigma}|^{-S/2} \exp[-tr\{(\boldsymbol{Y} - \boldsymbol{X\Phi})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X\Phi})'\}/2]$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S)'$, $\boldsymbol{\Phi} = (\boldsymbol{\mu}', \boldsymbol{B}')'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_S)'$, and $\boldsymbol{x}_s = (1, \boldsymbol{y}_{s-1}', \ldots, \boldsymbol{y}_{s-L}')'$. Note that $\boldsymbol{Y}$ is $S \times p$ matrix, $\boldsymbol{X}$ is $S \times (Lp + 1)$ matrix, and $\boldsymbol{\Phi}$ is $(Lp + 1) \times p$ matrix. Here we consider the impulse response to orthogonalized errors $U = \boldsymbol{\epsilon}'\boldsymbol{\Psi}^{-1}$, where $\boldsymbol{\Psi}$ is the Cholesky matrix for $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{\Sigma} = \boldsymbol{\Psi}'\boldsymbol{\Psi}$. That is, the impulse responses $\boldsymbol{Z}_k$ of $\boldsymbol{y}_{s+k}$ based on the structural shock $\boldsymbol{\epsilon}_s'\boldsymbol{\Psi}^{-1}$ is $\boldsymbol{Z}_k = \boldsymbol{\Psi H}_k$, where $\boldsymbol{H}_j = \sum_{i=1}^{j} \boldsymbol{B}_j \boldsymbol{H}_{j-i}$, and $\boldsymbol{B}_i = 0$ for $i$ larger than lag $L$ and $\boldsymbol{B}_0 = I$ (Sims, 1980; Ni et al., 2007).

For the computational purpose, we consider conjugate priors for $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$. That is, $\boldsymbol{\pi}(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(p+1)/2}$, the Jeffreys prior, and $\boldsymbol{\pi}(\boldsymbol{\phi}) \propto |M_0|^{-1/2} \exp\{-(\boldsymbol{\phi} - \boldsymbol{\phi}_0)M_0^{-1}(\boldsymbol{\phi} - \boldsymbol{\phi}_0)'/2\}$, where $\boldsymbol{\phi} = \text{vec}(\boldsymbol{\Phi})$. Next, following Ni et al. (2007), the conditional density of $\boldsymbol{\phi}$ given $\boldsymbol{\Sigma}, \boldsymbol{D}$ is $N(\boldsymbol{m}, \boldsymbol{V})$ and the conditional density of $\boldsymbol{\Sigma}$ given $\boldsymbol{\Phi}, \boldsymbol{D}$ is inverse Wishart $(\boldsymbol{S}(\boldsymbol{\Phi}), M)$, where $\boldsymbol{m} = \widehat{\boldsymbol{\phi}}_{mle} + \{M_0^{-1} + \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\}^{-1} M_0^{-1}(\boldsymbol{\phi}_0 - \widehat{\boldsymbol{\phi}}_{mle})$, $\boldsymbol{V} = \{M_0^{-1} + \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})\}^{-1}$, $\widehat{\boldsymbol{\phi}}_{mle} = \text{vec}(\widehat{\boldsymbol{\Phi}}_{mle})$, $\widehat{\boldsymbol{\Phi}}_{mle} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, $\boldsymbol{S}(\boldsymbol{\Phi}) = (\boldsymbol{Y} - \boldsymbol{X\Phi})'(\boldsymbol{Y} - \boldsymbol{X\Phi})$, and $\boldsymbol{D} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S\}$ is the observed data. Since the impulse response is a function of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, we rewrite $\boldsymbol{Z}_k = \boldsymbol{Z}(\boldsymbol{B}, \boldsymbol{\Sigma}, k)$. We take the posterior mean as a Bayes estimator of the impulse response, and it is $(\widehat{\boldsymbol{Z}}_k)_{(i,j)} = \int \{\boldsymbol{Z}(\theta, k)\}_{(i,j)} \pi(\theta|\boldsymbol{D}) d\theta$, where $\theta = (\boldsymbol{B}, \boldsymbol{\Sigma})$, and $(\boldsymbol{Z}_s)_{(i,j)}$ is the

$(i, j)$ element of $\boldsymbol{Z}_s$. Now, we apply the proposed method to calculate the frequentist standard error of $\widehat{\boldsymbol{Z}}_k$. In this complex example, it is nearly impossible to find the variance-covariance matrix of the sufficient statistics of $\theta$, therefore it is not possible to apply Efron's approach.

For illustration purpose, we generated a data set from the following VAR(1) model with $p = 2$,

$$\boldsymbol{y}_s' = \begin{bmatrix} -0.7 & 1.3 \end{bmatrix} + \begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.6 \end{bmatrix} \boldsymbol{y}_{s-1}' + \boldsymbol{\epsilon}_s,\ \boldsymbol{\epsilon}_s \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}),\ \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix},\ s = 1, \ldots, S,$$

with $S = 1,000$. Since the time series data is no more independent, we used the moving block bootstrap (MBB), where we divided the series into $N$ overlapping blocks of length $\ell$ to preserve the dependence structure of the original dataset (Kreiss and Lahiri, 2012). Then we chose $b$ blocks out of $N$ blocks to make the bootstrap observations $\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_S^*$.

We fit a VAR(2) model to the simulated dataset. As in previous examples, we used $M = 10000$ iterations after the burn-in samples. I imposed noninformative priors for $\boldsymbol{\Phi}$, where $\boldsymbol{\phi}_0 = \boldsymbol{0}$ and $M_0 = 20\boldsymbol{I}$. Then we drew $B = 500$ MBB samples with 15% of the total dataset as a block length $(\ell)$.

Figure 4.5 show the point estimate (posterior mean) and the 95% confidence band based on the frequentist standard error of the posterior mean for the impulse responses of $\boldsymbol{y}_2$ to $\boldsymbol{y}_1$ and $\boldsymbol{y}_1$ to $\boldsymbol{y}_2$, respectively. The confidence bands based on $sd_1$ and $sd_2$ are similar, but computationally the second approach $(sd_2)$ was about 5.6 times faster than the first approach $(sd_1)$. In Table 4.4, we also report the numerical values of the standard errors at each time lag, and the results do not show any appreciable difference between $sd_1$ and $sd_2$.

## 4.6 Conclusions

In this chapter we have discussed numerical approaches for efficient computation of standard errors for posterior summaries. The main theme of the chapter is to use bootstrap samples but avoid using full blown MCMC based inference for each of the bootstrap data. The methods rely on the importance sampling idea, and are broadly applicable. The R code for our computation is

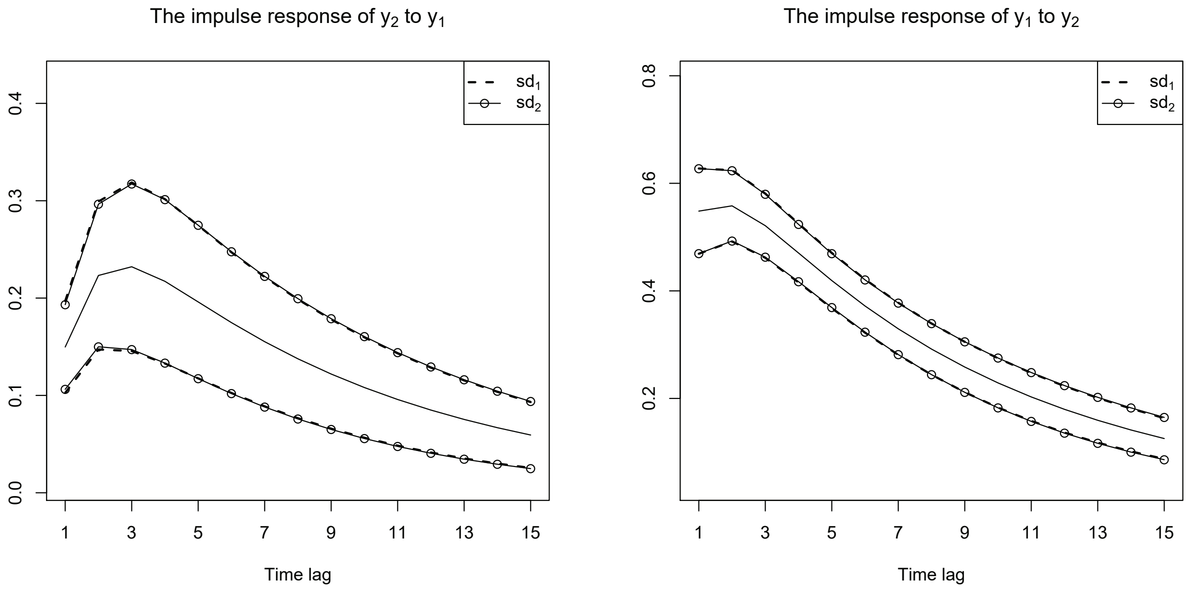

Figure 4.5: The estimated impulse responses (solid line) and its 95% (pointwise) confidence band of $\boldsymbol{y}_2$ to a shock in $\boldsymbol{y}_1$ (left panel) and vice versa (right panel) referenced in Section 5.4. The bold dotted line is based on regular bootstrap approach ($sd_1$) while the circled solid line is based on importance sampling based approach ($sd_2$).

available at `https://stat.tamu.edu/~sinha/research.html`.

It is well-known that the presence of outliers results in a poor performance in a bootstrap approach because they are more frequent in bootstrap samples than the original dataset if we consider the classical nonparametric bootstrap (Salibian-Barrera and Zamar, 2002; Willems and Van Aelst, 2005; Huber and Ronchetti, 2009). Therefore, the performance of our proposed method can be affected by the outliers in the data as we have used the classical nonparametric bootstrap with replacement. However, this may be overcome by considering robust bootstrap methods for drawing samples (Singh, 1998; Hu and Hu, 2000; Salibian-Barrera and Zamar, 2002), or a combination of a robust bootstrap method and a robust Bayesian method, possibly with a flat-tailed prior (Berger et al., 1994; Marín, 2000).

Table 4.4: The frequentist standard errors of the estimated the impulse responses at each time lag. Here $sd_1$ and $sd_2$ denote the standard errors based on the regular bootstrap method and the importance sampling based approach.

| Time lag | $\boldsymbol{y}_2$ to $\boldsymbol{y}_1$ | | $\boldsymbol{y}_1$ to $\boldsymbol{y}_2$ | |
|---|---|---|---|---|
| | $sd_1$ | $sd_2$ | $sd_1$ | $sd_2$ |
| 1 | 0.0241 | 0.0222 | 0.0404 | 0.0403 |
| 2 | 0.0388 | 0.0374 | 0.0338 | 0.0334 |
| 3 | 0.0442 | 0.0434 | 0.0304 | 0.0298 |
| 4 | 0.0431 | 0.0428 | 0.0279 | 0.0272 |
| 5 | 0.0401 | 0.0402 | 0.0262 | 0.0256 |
| 6 | 0.0369 | 0.0372 | 0.0253 | 0.0248 |
| 7 | 0.0339 | 0.0342 | 0.0247 | 0.0244 |
| 8 | 0.0311 | 0.0315 | 0.0243 | 0.0242 |
| 9 | 0.0286 | 0.0290 | 0.0239 | 0.0240 |
| 10 | 0.0263 | 0.0267 | 0.0234 | 0.0236 |
| 11 | 0.0242 | 0.0246 | 0.0227 | 0.0232 |
| 12 | 0.0222 | 0.0226 | 0.0220 | 0.0225 |
| 13 | 0.0204 | 0.0208 | 0.0212 | 0.0218 |
| 14 | 0.0188 | 0.0191 | 0.0203 | 0.0210 |
| 15 | 0.0173 | 0.0176 | 0.0194 | 0.0201 |

# 5. IDENTIFIABILITY AND BIAS REDUCTION IN THE SKEW-PROBIT MODEL FOR A BINARY RESPONSE

## 5.1 Background and literature review

Logistic or probit model is widely used for modelling the success probability of a binary variable in terms of covariates. Under the logistic model pr$(Y = 1|\boldsymbol{X}) = H(\boldsymbol{\gamma}^T\boldsymbol{Z})$ with $H(u) = \exp(u)/\{1 + \exp(u)\}$, and under the probit model pr$(Y = 1|\boldsymbol{X}) = \Phi(\boldsymbol{\gamma}^T\boldsymbol{Z})$ with $\Phi(u)$ being the cumulative distribution function (CDF) of the standard normal distribution, and $\boldsymbol{Z} = (1, \boldsymbol{X}^T)^T$. Both link functions, $H$ and $\Phi$, are considered to be symmetric link functions as they approach to zero and one at the same rate. For a flexible regression model, practitioners may wish to use an asymmetric link that accommodates different convergence rates towards zero and one. Failure to fit a flexible model to the data may result in biased estimates of regression parameters, odds ratios, or risk differences. As discussed in Section 1.2.2, binary regression with the *skew-probit* link is a good alternative to ones with symmetric link function. Particularly, for the *skew-probit* link,

$$\text{pr}(Y = 1|\boldsymbol{X}) = F(\eta, \delta) = \int_{-\infty}^{\eta} 2\phi(u)\Phi(\delta u)du, \tag{5.1}$$

where $\eta = \boldsymbol{Z}^T\boldsymbol{\beta} = \beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}_1$ with $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$, and $\phi(u) = d\Phi(u)/du$. Note that the integrand in (5.1) represents the density of the standard skew-normal distribution with the skewness parameter $\delta$, that is denoted by Skew-Normal$(\mu = 0, \omega = 1, \delta)$. Here $F$ denotes the CDF of Skew-Normal$(\mu = 0, \omega = 1, \delta)$.

The skew-normal distribution and its properties are well studied in the literature (Azzalini, 1985; Genton et al., 2001; Ma and Genton, 2004). Regarding the exact use of the skew-probit link, Bazán et al. (2006) used this skew-probit model to analyze a Rasch-model for the item response theory. Stingo et al. (2011) considered an extension of the skew-probit link to model a binary response variable in the presence of selectivity bias (Bhattacharya et al., 2006). A decent review of some recent applications of the skew-probit link can be found in Bazán et al. (2014).

In this chapter I address two important issues, identifiability of the model parameters and the bias of the maximum likelihood estimator (MLE) of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \delta)^T$. A clear knowledge on the identifiability of parameters is necessary for proposing any method of estimation. Secondly, biased estimates may lead to incorrect inference regarding the model parameters, the association between the response and covariates, and the marginal effect of the covariate. Although these issues are important for model formulations and deciding on the appropriate method of analysis, these issues have not been investigated till date.

Now let me briefly mention some existing literature on these issues. Genton and Zhang (2012) investigated identifiability for some non-Gaussian spatial random fields that include multivariate skew-normal distributions. Castro et al. (2013) studied parameter identifiability for multivariate skew-normal distributions. Otiniano et al. (2015) investigated parameter identifiability for a finite mixture of skew-normal distributions and a finite mixture of skew-t distributions. Although these approaches considered the important case of a continuous response variable, parameter identifiability has not been investigated for a binary response variable that follows the skew-probit link.

The bias in the MLE of the skew-normal model where the response $Y$ is continuous and follows Skew-Normal$(\mu, \omega, \delta)$, is a well-researched topic. Following Firth (1993)'s general recommendation to reducing finite sample bias, Sartori (2006) proposed to estimate the skewness parameter $\delta$ of the Skew-Normal$(\mu = 0, \omega = 1, \delta)$ model by maximizing the penalized log-likelihood, $\ell + 0.5\log\{\mathrm{determinant}(\mathcal{I})\}$, where $\ell$ stands for the log-likelihood while $\mathcal{I}$ stands for the Fisher information matrix. Sartori (2006) also considered estimation of $\delta$ in the presence of unknown $\mu$ and $\omega$, where only $\delta$ was estimated by maximizing a penalized profiled log-likelihood function and the other parameters were estimated by maximizing the likelihood function for a given $\delta$. Later on, Azzalini and Arellano-Valle (2013) applied the penalized likelihood idea in the general case of three-parameter Skew Normal$(\mu, \omega, \delta)$ model, where all three parameters were estimated by maximizing the penalized log-likelihood function. To reduce the finite sample bias, researchers considered Bayesian inference of the skew-normal model under various priors including default and proper priors (Liseo and Loperfido, 2006; Bayes and Branco, 2007).

In this chapter, I shall consider model (5.1) for a binary response variable $Y$. Hence, our model is distinct from the existing papers discussed in the previous two paragraphs where the response $Y$ was considered to be a continuous variable. Furthermore, we are considering the issue in the presence of a regressor variable $\boldsymbol{X}$ that no one has considered before even when a continuous $Y$ followed a skew-normal distribution. As a general strategy to reduce the first order bias in the MLE of $\boldsymbol{\beta}$ and $\delta$, one may consider the bootstrap bias correction approach or the bias correction approach of Cox and Snell (1968). These two approaches require the MLE to be finite that may not happen in small samples. Therefore, as an alternative, I consider estimation of the parameters by maximizing a penalized likelihood function. In this penalized likelihood method, first I apply Firth (1993)'s method to prevent the bias where the likelihood function is penalized by the Jeffrey's prior. Additionally, we consider two more penalization approaches one by using the generalized information matrix prior (Gupta and Ibrahim, 2009) and two by using the Cauchy prior (Gelman et al., 2008). Finally, all these methods are compared through extensive simulation studies.

This research was partly motivated by a dataset on heart-disease (Detrano et al., 1989), where the interest is in finding association between the occurrence of artery blockage and several clinical variables. A standard probit analysis of this data indicates a lack-of-fit at the 5% level of significance and that led us to consider the skew-probit model. As we will see in the data analysis section that there is a significant improvement in the goodness-of-fit statistic after considering a bias correction approach in the skew-probit model.

Before concluding this section I would like to highlight the novelties of this work. To the best of knowledge, this is the first work that investigates parameter identifiability and the bias in the MLE of the binary model with the skew-probit link function. To reduce finite sample bias, we apply general bias reduction strategies to this particular problem, and compare and assess the effectiveness of the approaches through simulation studies. Simulation results indicate that the bias reduction strategies need to be used judiciously.

### 5.2 Parameter identifiability

In general, model parameters are identifiable if the parameter values uniquely identify the underlying probability model. Now, following Rothenberg (1971)'s general concept of identifiability, I present a formal definition of identifiability in our context of the skew-probit model.

Identifiability. The parameter set $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \delta)^T$ is said to be identifiable if $F(\boldsymbol{Z}^T\boldsymbol{\beta}, \delta) = F(\boldsymbol{Z}^T\boldsymbol{\beta}^{'}, \delta^{'})$ for every $\boldsymbol{Z}$ implies $(\boldsymbol{\beta}^{'}, \delta^{'}) = (\boldsymbol{\beta}, \delta)$. A parameter set $\boldsymbol{\theta}$ is said to be locally identifiable if within a neighborhood $\mathcal{N}$ there does not exist a $(\boldsymbol{\beta}^{'}, \delta^{'}) \in \mathcal{N}\backslash\{(\boldsymbol{\beta}, \delta)\}$ such that $F(\boldsymbol{Z}^T\boldsymbol{\beta}, \delta) = F(\boldsymbol{Z}^T\boldsymbol{\beta}^{'}, \delta^{'})$ for every $\boldsymbol{Z}$. A necessary and sufficient condition for local identifiability is the non-singularity of the Fisher information matrix. Now, we investigate identifiability of three different cases.

**No covariate:** In the absence of any covariate, the intercept $\beta_0$ and the skewness parameter $\delta$ are not identifiable in the skew-probit model pr$(Y = 1) = F(\beta_0, \delta) = \int_{-\infty}^{\beta_0} 2\phi(u)\Phi(\delta u)du$. In other words, for a given value of $(\beta_0, \delta)$ we can find another $(\beta_0^{'}, \delta^{'})$ such that $F(\beta_0, \delta) = F(\beta_0^{'}, \delta^{'})$. This fact is illustrated in Figure 5.1. This figure contains two CDFs for Skew-Normal$(\mu = 0, \omega = 1, \delta)$ and Skew-Normal$(\mu = 0, \omega = 1, \delta^{'})$ distributions. At the abscissa $\beta_0$, the height of the dotted vertical line up to the CDF for the Skew-Normal$(\mu = 0, \omega = 1, \delta)$ distribution is $F(\beta_0, \delta)$. For the same value of the CDF, $F(\beta_0, \delta)$, there is another $\beta_0^{'}$ and $\delta^{'}$, such that $F(\beta_0, \delta) = F(\beta_0^{'}, \delta^{'})$. Particularly, the abscissa of the point where the horizontal line at $F(\beta_0, \delta)$ hits the CDF for the Skew-Normal$(\mu = 0, \omega = 1, \delta^{'})$ distribution is $\beta_0^{'}$. This signifies that the CDF of the Skew-Normal$(\mu = 0, \omega = 1, \delta^{'})$ distribution at $\beta_0^{'}$ is the same as $F(\beta_0, \delta)$. If $\ell$ stands for the log-likelihood, then analytical calculations show that $E(\partial^2\ell/\partial\beta_0\partial\beta_0) = -4\phi^2(\beta_0)\Phi^2(\beta_0\delta)/F(\beta_0, \delta)\{1 - F(\beta_0, \delta)\}$, $E(\partial^2\ell/\partial\delta\partial\delta) = -\exp\{-\beta_0^2(1 + \delta^2)\}/\pi^2(1+\delta^2)^2 F(\beta_0, \delta)\{1-F(\beta_0, \delta)\}$, $E(\partial^2\ell/\partial\beta_0\partial\delta) = 2\phi(\beta_0)\Phi(\beta_0\delta)\exp\{-\beta_0^2(1+\delta^2)/2\}/\pi(1+\delta^2)F(\beta_0, \delta)\{1 - F(\beta_0, \delta)\}$, and the determinant of the Fisher information matrix $E(\partial^2\ell/\partial\beta_0\partial\beta_0)$ $E(\partial^2\ell/\partial\delta\partial\delta) - E^2(\partial^2\ell/\partial\beta_0\partial\delta) = 0$.

**Binary covariate:** Now suppose that there is a binary covariate $X$, and the model is pr$(Y = 1|X) = F(\beta_0 + \beta_1 X, \delta)$. If the parameter $(\beta_0, \beta_1, \delta)$ is non-identifiable, then we can find a

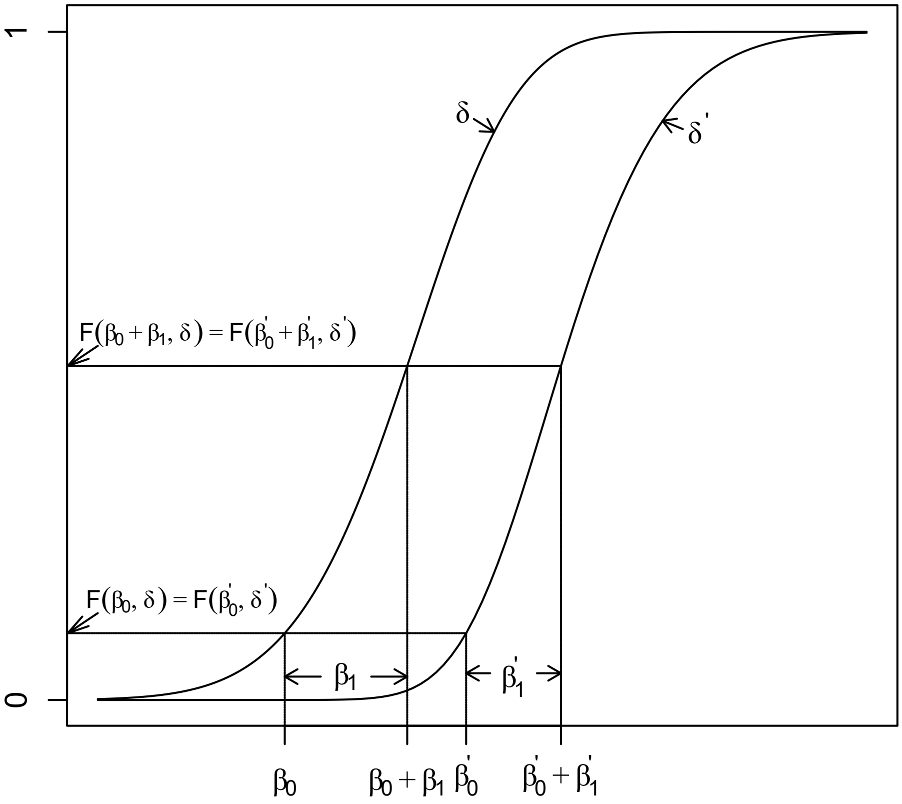

Figure 5.1: Illustration of parameter identifiability in the skew-probit model with a binary covariate.

$(\beta_0^{'}, \beta_1^{'}, \delta^{'}) \neq (\beta_0, \beta_1, \delta)$, such that $F(\beta_0 + \beta_1 X, \delta) = F(\beta_0^{'} + \beta_1^{'} X, \delta^{'})$ for every $X$. Now consider the two probabilities, pr$(Y = 1|X = 1) = F(\beta_0 + \beta_1, \delta)$ and pr$(Y = 1|X = 0) = F(\beta_0, \delta)$. From the discussion in the previous paragraph, we know that for a given $(\beta_0, \delta)$ we can find a $(\beta_0^{'}, \delta^{'}) \neq (\beta_0, \delta)$ such that $F(\beta_0, \delta) = F(\beta_0^{'}, \delta^{'})$. Now, it turns out that given these two sets, $(\beta_0, \delta)$ and $(\beta_0^{'}, \delta^{'})$, for every $\beta_1$ we can find a $\beta_1^{'}$, such that $F(\beta_0 + \beta_1, \delta) = F(\beta_0^{'} + \beta_1^{'}, \delta^{'})$. In Figure 5.1, at the abscissa $(\beta_0 + \beta_1)$ the height of the dotted vertical line up to the CDF for the Skew-Normal$(\mu = 0, \omega = 1, \delta)$ distribution is $F(\beta_0 + \beta_1, \delta)$. Now, the abscissa of the intersection point of the horizontal line at $F(\beta_0 + \beta_1, \delta)$ with the CDF for the Skew-Normal$(\mu = 0, \omega = 1, \delta^{'})$ distribution is $\beta_0^{'} + \beta_1^{'}$. That means, $F(\beta_0 + \beta_1, \delta) = F(\beta_0^{'} + \beta_1^{'}, \delta^{'})$. Hence, the model parameters are not identifiable. Using similar arguments we conclude that for a categorical covariate $X$, the model parameters of a skew-probit model are not identifiable.

**Continuous covariate:** Here I show that if the covariate $X$ is a continuous variable, the model parameters are identifiable. We assume that $\beta_1$ is non-zero, otherwise it will be the same as the case where there is no covariate. Suppose that $\boldsymbol{\theta} = (\beta_0, \beta_1, \delta)^T$ involved in the skew-probit model is not identifiable. In the following discussion we shall be using the fact that for a fixed $\delta$, $F(\cdot, \delta)$ is a strictly increasing function so that its inverse function $F_\delta^{-1}(\cdot)$ exists. If the parameters are not identifiable, then there exists a $\boldsymbol{\theta}^{'} = (\beta_0^{'}, \beta_1^{'}, \delta^{'}) \neq \boldsymbol{\theta}$ such that

$$F(\beta_0 + \beta_1 X, \delta) = F(\beta_0^{'} + \beta_1^{'} X, \delta^{'}) \text{ for all } X, \tag{5.2}$$

and particularly for $X = 0$, non-identifiability implies

$$F(\beta_0, \delta) = F(\beta_0^{'}, \delta^{'}). \tag{5.3}$$

Now, using the inverse operation on (5.3) and (5.2) we obtain

$$F_\delta^{-1}\{F(\beta_0, \delta)\} = \beta_0 = F_\delta^{-1}\{F(\beta_0^{'}, \delta^{'})\}, \tag{5.4}$$

$$\beta_0 + \beta_1 X = F_\delta^{-1}\{F(\beta_0^{'} + \beta_1^{'} X, \delta^{'})\}. \tag{5.5}$$

When $\delta = \delta^{'}$, $\beta_0 = F_\delta^{-1}\{F(\beta_0^{'}, \delta^{'})\} = F_\delta^{-1}\{F(\beta_0^{\prime}, \delta)\} = \beta_0^{'}$, and similarly we obtain $\beta_1 = \beta_1^{'}$. Thus, when $\delta = \delta^{'}$, we cannot have two different sets $(\beta_0, \beta_1, \delta) \neq (\beta_0^{\prime}, \beta_1^{\prime}, \delta)$ such that $F(\beta_0 + \beta_1 X, \delta) = F(\beta_0^{\prime} + \beta_1^{\prime} X, \delta)$ for all $X$.

When $\delta \neq \delta^{'}$, subtracting (5.4) from (5.5) we obtain

$$\beta_1 X = F_\delta^{-1}\{F(\beta_0^{'} + \beta_1^{'} X, \delta^{\prime})\} - F_\delta^{-1}\{F(\beta_0^{'}, \delta^{'})\} \tag{5.6}$$

for $X \neq 0$. Differentiating both sides of Equation (5.6) with respect to $X$ we get

$$\beta_1 = \frac{\phi(\beta_0^{'} + \beta_1^{'} X)\Phi\{\delta^{'}(\beta_0^{'} + \beta_1^{'} X)\}\beta_1^{'}}{\phi[F_\delta^{-1}\{F(\beta_0^{'} + \beta_1^{'} X, \delta^{'})\}]\Phi[\delta F_\delta^{-1}\{F(\beta_0^{'} + \beta_1^{'} X, \delta^{'})\}]}. \tag{5.7}$$

Since $\delta' \neq \delta$, $F_\delta^{-1}\{F(\beta_0^{'} + \beta_1^{'} X, \delta^{'})\} \neq \beta_0^{'} + \beta_1^{'} X$ for all $X$, which means that the right-hand side of (5.7) is a non-linear function of $X$ while the left-hand side is a constant. Therefore, our assumption that $\boldsymbol{\theta}$ is not identifiable is wrong.

## 5.3 Bias reduction

### 5.3.1 Maximum likelihood and bootstrap

Suppose that the observed data $\boldsymbol{D} = (D_1, \ldots, D_n)$ with $D_i = (Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$ are collected from $n$ subjects that are randomly drawn from the underlying population. At least one component of the covariate vector is assumed to be continuous. We want to fit the regression model $(5.1)$ to the data. The logarithm of the likelihood is

$$\ell = \sum_{i=1}^{n} Y_i \text{log}\{F(\eta_i, \delta)\} + (1 - Y_i)\text{log}\{1 - F(\eta_i, \delta)\},$$

where $\eta_i = \boldsymbol{Z}_i^T \boldsymbol{\beta}$ and $\boldsymbol{Z}_i = (1, \boldsymbol{X}_i^T)^T$. The maximum likelihood estimators (MLE) of $\boldsymbol{\beta}$ and $\delta$ are obtained by solving $\partial\ell/\partial\boldsymbol{\theta} = (\partial\ell/\partial\boldsymbol{\beta}^T, \partial\ell/\partial\delta)^T = \mathbf{0}$, where

$$\begin{aligned} \frac{\partial\ell}{\partial\boldsymbol{\beta}} &= 2\sum_{i=1}^{n}\left\{\frac{Y_i}{F(\eta_i, \delta)} - \frac{(1 - Y_i)}{1 - F(\eta_i, \delta)}\right\}\phi(\eta_i)\Phi(\delta\eta_i)\boldsymbol{Z}_i, \\ \frac{\partial\ell}{\partial\delta} &= \sum_{i=1}^{n}\left\{-\frac{Y_i}{F(\eta_i, \delta)} + \frac{(1 - Y_i)}{1 - F(\eta_i, \delta)}\right\}\frac{\exp\{-\eta_i^2(1 + \delta^2)/2\}}{\pi(1 + \delta^2)}. \end{aligned}$$

In principle, the parameter estimates can be obtained by solving the above equations using the scoring method. Let $\boldsymbol{\theta}^{(t)}$ be the parameter value at the $t$th iteration of the scoring method. Then at the $(t + 1)$th iteration we obtain

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \mathcal{I}^{-1}(\boldsymbol{\theta}^{(t)})\left(\frac{\partial\ell}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}},$$

where the information matrix $\mathcal{I}(\boldsymbol{\theta}) = -E(\partial^2\ell/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)$ with

$$
\begin{aligned}
E\left(\frac{\partial^2\ell}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T}\right) &= -4\sum_{i=1}^{n}\frac{\phi^2(\eta_i)\Phi^2(\delta\eta_i)}{F(\eta_i,\delta)\{1-F(\eta_i,\delta)\}}\boldsymbol{Z}_i\boldsymbol{Z}_i^T,\\
E\left(\frac{\partial^2\ell}{\partial\delta^2}\right) &= -\sum_{i=1}^{n}\frac{\exp\{-\eta_i^2(1+\delta^2)\}}{\pi^2(1+\delta^2)^2F(\eta_i,\delta)\{1-F(\eta_i,\delta)\}},\\
E\left(\frac{\partial^2\ell}{\partial\delta\partial\boldsymbol{\beta}}\right) &= 2\sum_{i=1}^{n}\frac{\phi(\eta_i)\Phi(\delta\eta_i)\exp\{-\eta_i^2(1+\delta^2)/2\}}{\pi(1+\delta^2)F(\eta_i,\delta)\{1-F(\eta_i,\delta)\}}\boldsymbol{Z}_i.
\end{aligned}
$$

We note that the information matrix can be written as $\mathcal{I}(\boldsymbol{\theta}) = \boldsymbol{W}(\boldsymbol{\theta})^T\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})$, where $\boldsymbol{A}(\boldsymbol{\theta}) = \text{diag}[F(\eta_i,\delta)\{1-F(\eta_i,\delta)\}]^{-1}$, $\boldsymbol{W}(\boldsymbol{\theta})^T = [\boldsymbol{W}_1(\boldsymbol{\theta}),\ldots,\boldsymbol{W}_n(\boldsymbol{\theta})]$, $\boldsymbol{W}_i^T(\boldsymbol{\theta}) = [2\phi(\eta_i)\Phi(\delta\eta_i)\boldsymbol{Z}_i^T, -\exp\{-\eta_i^2(1+\delta^2)/2\}/\pi(1+\delta^2)] = 2\phi(\eta_i)\Phi(\delta\eta_i)[\boldsymbol{Z}_i^T, -\phi(\delta\eta_i)/(1+\delta^2)\Phi(\delta\eta_i)]$. Since $\eta_i = \boldsymbol{Z}_i^T\boldsymbol{\beta}$, the last element in $\boldsymbol{W}_i^T(\theta)$ is non-linearly related to $\boldsymbol{Z}_i$, which means $\boldsymbol{W}(\boldsymbol{\theta})$ has full rank unless the original design matrix is singular. Therefore the above estimation technique can be nicely expressed in terms of the iteratively re-weighted least square (IWLS) method, where $\boldsymbol{\theta}^{(t+1)} = \{\boldsymbol{W}(\boldsymbol{\theta}^{(t)})^T\boldsymbol{A}(\boldsymbol{\theta}^{(t)})\boldsymbol{W}(\boldsymbol{\theta}^{(t)})\}^{-1}\boldsymbol{W}(\boldsymbol{\theta}^{(t)})^T\boldsymbol{A}(\boldsymbol{\theta}^{(t)})\boldsymbol{Y}^*(\boldsymbol{\theta}^{(t)})$, where $\boldsymbol{Y}^*(\boldsymbol{\theta}^{(t)}) = \boldsymbol{W}(\boldsymbol{\theta}^{(t)})\boldsymbol{\theta}^{(t)} + \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}^{(t)})\}, \boldsymbol{Y} = (Y_1,\ldots,Y_n)^T, \boldsymbol{\mu}(\boldsymbol{\theta}^{(t)}) = (F(\eta_1^{(t)},\delta^{(t)}),\ldots,F(\eta_n^{(t)},\delta^{(t)}))^T$, and $\eta_i^{(t)} = \boldsymbol{Z}_i^T\boldsymbol{\beta}^{(t)}$. This approach is referred to as method N.

For larger values of $\delta$, the curvature $E(-\partial^2\ell/\partial\ell^2)$ tends to be small, resulting in highly biased MLE of $\delta$. Additionally, if there is no covariate, and the model for $Y$ is pr$(Y = 1) = 2\int_{-\infty}^{0}\phi(u)\Phi(\delta u)du$ that involves with only one parameter $\delta$, the probability that the MLE of $\delta$ diverges to $+\infty$ or $-\infty$ is $p_n(\delta) = \text{pr}(Y_1 = \cdots = Y_n = 0) + \text{pr}(Y_1 = \cdots = Y_n = 1) = \{\pi + 2\tan^{-1}(\delta)/(2\pi)\}^n + \{\pi - 2\tan^{-1}(\delta)/(2\pi)\}^n$. Although this probability goes to zero as $n \to \infty$, this may not be negligible for a moderate value of $n$. This $p_n(\delta)$ is also the probability of diverging MLE of $\delta$ when a continuous response follows skew-normal $(\mu = 0, \omega = 1, \delta)$ (Azzalini and Arellano-Valle, 2013).

In order to reduce the finite sample bias of the MLE that is of the order $O(n^{-1})$, we consider the following strategies. First, we apply the bootstrap method to reduce the bias of the MLE. Suppose that $b(\widehat{\boldsymbol{\theta}}_{MLE})$ denotes the bias of $\widehat{\boldsymbol{\theta}}_{MLE}$, the MLE of $\boldsymbol{\theta}$. Based on $B$ bootstrap samples, we estimate

$b(\widehat{\boldsymbol{\theta}}_{MLE})$, and denote this estimator of bias by $\widehat{b}_{\mathrm{boot}}(\widehat{\boldsymbol{\theta}}_{MLE})$. The bias corrected estimator is then defined as $\widehat{\boldsymbol{\theta}}_{MLE} - \widehat{b}_{\mathrm{boot}}(\widehat{\boldsymbol{\theta}}_{MLE})$. This approach is referred to as method B.

### 5.3.2 Penalized maximum likelihood

Next, I propose to estimate the parameters by maximizing a penalized likelihood,

$$\ell_p = \ell + M(\boldsymbol{\theta}),$$

where $M(\boldsymbol{\theta})$ is the penalty function. The estimator obtained by maximizing $\ell_p$ can be seen as the posterior mode where the prior distribution $\pi(\boldsymbol{\theta}) \propto \exp\{M(\boldsymbol{\theta})\}$. Unlike the other bias correction approaches that require the estimator to be finite, this approach does not require the MLE to be finite. Rather penalization helps to add a curvature in a otherwise flat likelihood surface, and thereby the penalized likelihood method prevents the estimate to be infinite or unrealistically large and also reduces finite sample bias. Following the general strategy of Firth (1993), we replace $M(\boldsymbol{\theta})$ by $0.5\log[\det\{\mathcal{I}(\boldsymbol{\theta})\}]$, where $\det$ stands for matrix determinant. Thus, the maximum penalized likelihood estimator, denoted by $\widehat{\boldsymbol{\theta}}_{\mathrm{pj}}$, is obtained by solving

$$\begin{aligned}
\frac{\partial \ell_p}{\partial \boldsymbol{\beta}} &= 2\sum_{i=1}^{n}\left\{\frac{Y_i}{F(\eta_i,\delta)} - \frac{(1-Y_i)}{1-F(\eta_i,\delta)}\right\}\phi(\eta_i)\Phi(\delta\eta_i)\boldsymbol{Z}_i + \frac{1}{2}\mathrm{trace}\left\{\mathcal{I}^{-1}(\boldsymbol{\theta})\frac{\partial \mathcal{I}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right\} = \mathbf{0}, \\
\frac{\partial \ell_p}{\partial \delta} &= \sum_{i=1}^{n}\left\{-\frac{Y_i}{F(\eta_i,\delta)} + \frac{(1-Y_i)}{1-F(\eta_i,\delta)}\right\}\frac{\exp\{-\eta_i^2(1+\delta^2)/2\}}{\pi(1+\delta^2)} \\
&+ \frac{1}{2}\mathrm{trace}\left\{\mathcal{I}^{-1}(\boldsymbol{\theta})\frac{\partial \mathcal{I}(\boldsymbol{\theta})}{\partial \delta}\right\} = 0.
\end{aligned}$$

This approach is referred to as method J. This estimator can be seen as the posterior mode when the Jeffrey's prior is used on the parameters as $e^{M(\boldsymbol{\theta})} = \det\{\mathcal{I}(\boldsymbol{\theta})\}^{1/2}$. Although this approach of bias reduction has been extensively used in various contexts including when a continuous response follows skew-normal $(\mu = 0, \omega = 1, \delta)$ (Azzalini and Arellano-Valle, 2013), the approach has never been applied to the case where the binary response variable $Y$ is modeled via the skew-probit link.

Next, we consider a generalization of the Jeffrey's prior (Gupta and Ibrahim, 2009), where the prior $\pi_{GI}(\boldsymbol{\theta}) \propto |\det\{\mathcal{I}(\boldsymbol{\theta})\}|^{1/2}\exp\{-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathcal{I}(\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)/2c_0\}$. For large $c_0$, $\pi_{GI}(\boldsymbol{\theta})$ converges $|\det\{\mathcal{I}(\boldsymbol{\theta})\}|^{1/2}$, that is Jeffery's prior. Gupta and Ibrahim (2009) showed that under a logistic model, $\pi_{GI}$ has lower mass around the center and heavier tail than the normal distribution resulting in a relatively non-informative prior. Adopting their prior distribution with $c_0 = 1$ and $\boldsymbol{\theta}_0 = \mathbf{1}$, and setting $M(\boldsymbol{\theta}) = \log\{\pi_{GI}(\boldsymbol{\theta})\}$ in our penalized likelihood $\ell_p$, we obtain the following estimating equations to estimate $(\boldsymbol{\beta}^T, \delta)^T$

$$\begin{aligned}
\frac{\partial \ell_p}{\partial \boldsymbol{\beta}} &= 2\sum_{i=1}^{n}\left\{\frac{Y_i}{F(\eta_i,\delta)} - \frac{(1-Y_i)}{1-F(\eta_i,\delta)}\right\}\phi(\eta_i)\Phi(\delta\eta_i)\boldsymbol{Z}_i + \frac{1}{2}\text{trace}\left\{\mathcal{I}^{-1}(\boldsymbol{\theta})\frac{\partial \mathcal{I}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right\} \\
&- \frac{1}{2}\frac{\partial \boldsymbol{\theta}^T\mathcal{I}(\boldsymbol{\theta})\boldsymbol{\theta}}{\partial \boldsymbol{\beta}} = \mathbf{0}, \\
\frac{\partial \ell_p}{\partial \delta} &= \sum_{i=1}^{n}\left\{-\frac{Y_i}{F(\eta_i,\delta)} + \frac{(1-Y_i)}{1-F(\eta_i,\delta)}\right\}\frac{\exp\{-\eta_i^2(1+\delta^2)/2\}}{\pi(1+\delta^2)} + \frac{1}{2}\text{trace}\left\{\mathcal{I}^{-1}(\boldsymbol{\theta})\frac{\partial \mathcal{I}(\boldsymbol{\theta})}{\partial \delta}\right\} \\
&- \frac{1}{2}\frac{\partial \boldsymbol{\theta}^T\mathcal{I}(\boldsymbol{\theta})\boldsymbol{\theta}}{\partial \delta} = 0.
\end{aligned}$$

This method is referred to as method G.

Gelman et al. (2008) pointed out use of Jeffrey's prior distribution might produce unreliable computation and be difficult to interpret in the Bayesian context. To avoid these potential issues, they proposed weakly informative Cauchy distribution prior for estimating logistic model parameters which results in stable and regularized estimates. Adopting their recommendation in our setup we consider $e^{M(\boldsymbol{\theta})} = \Pi_k\{\pi(1+\theta_k^2/2.5^2)\}^{-1}$, i.e., $M(\boldsymbol{\theta}) = -\sum_k \log(1+\theta_k^2/2.5^2)$. This implies independent $\text{Cauchy}(0, 2.5)$ prior for each component of $\boldsymbol{\theta}$. Corresponding estimators are obtained by solving

$$\begin{aligned}
\frac{\partial \ell_p}{\partial \boldsymbol{\beta}} &= 2\sum_{i=1}^{n}\left\{\frac{Y_i}{F(\eta_i,\delta)} - \frac{(1-Y_i)}{1-F(\eta_i,\delta)}\right\}\phi(\eta_i)\Phi(\delta\eta_i)\boldsymbol{Z}_i \\
&- \mathbf{1}^T\text{Diag}\left(\frac{2\beta_0}{2.5^2+\beta_0^2}, \frac{2\beta_1}{2.5^2+\beta_1^2}, \ldots, \frac{2\beta_q}{2.5^2+\beta_q^2}\right) = \mathbf{0} \\
\frac{\partial \ell_p}{\partial \delta} &= \sum_{i=1}^{n}\left\{-\frac{Y_i}{F(\eta_i,\delta)} + \frac{(1-Y_i)}{1-F(\eta_i,\delta)}\right\}\frac{\exp\{-\eta_i^2(1+\delta^2)/2\}}{\pi(1+\delta^2)} - \frac{2\delta}{2.5^2+\delta^2} = 0,
\end{aligned}$$

where $q$ is the number of covariates. This approach is referred to as method C.

Similar to naive MLE, the parameter estimates can be obtained using the Fisher scoring method with modified score function $(\partial \ell_p / \partial \boldsymbol{\theta})$ instead of $(\partial \ell / \partial \boldsymbol{\theta})$ (Heinze and Schemper, 2002). Note that the penalty function $M(\boldsymbol{\theta})$ is a $O_p(1)$ order term while the log-likelihood $\ell$ is $O_p(n)$ order term. Therefore, the asymptotic standard error calculation using the Fisher information matrix is still valid. That is, under certain regularity conditions, we may apply the standard likelihood theory to test hypotheses regarding parameters.

We consider two other penalized estimators. First, where the Jeffrey's prior for $\delta$ is constructed assuming $\beta_0 = 0$ and $\boldsymbol{\beta}_1 = \mathbf{0}$, and the logarithm of the prior density is used as the penalty function $M(\boldsymbol{\theta})$. Second, we take $M(\boldsymbol{\theta})$ to be the logarithm of the density function of the $t$ distribution with degrees of freedom 2, location 0 and scale parameter 0.5 on the skewness parameter $\delta$. This $t$ density for $\delta$ arises due to a non-informative prior on $\kappa$ when a standard skew-normal variable $U$ with the skewness parameter $\delta$ is expressed as $U = \sqrt{1-\kappa^2}Z + \kappa Z^*$, with $Z \sim \mathrm{Normal}(0,1)$, and $Z^*$ follows a half-normal density with the density function $f(Z^*) = 2(2\pi)^{-1/2}\exp\{-(Z^*)^2/2\}$, $Z^* > 0$ (Henze, 1986). However, in our initial numerical studies the performance of these penalized estimators is much worse than the other penalized estimators, so we have omitted them from further consideration.

### 5.4 Simulation studies

Design: I simulated datasets of different sizes, $n = 100, 500, 1000$ and $2000$. Each simulated dataset consists of a scalar covariate $X$ and a binary response $Y$. Given $X$, $Y$ was generated using the $\mathrm{Bernoulli}$ distribution with success probability pr$(Y = 1|X) = F(\beta_0 + \beta_1 X, \delta)$, and define $p_m =$ pr$(Y = 1) = \int$ pr$(Y = 1|x)g(x)dx$ as the marginal success probability. By varying $\delta$, $p_m$ and the distribution of $X$, we obtained the following 8 scenarios:

**Scenario 1.** $X \sim \mathrm{Uniform}(-2,2)$, $\beta_1 = 1$, $\delta = 4$, $\beta_0 = -0.87$, $p_m = 12\%$;

**Scenario 2.** $X \sim \mathrm{Uniform}(-2,2)$, $\beta_1 = 1$, $\delta = 4$, $\beta_0 = 0.37$, $p_m = 40\%$;

**Scenario 3.** $X \sim \mathrm{Uniform}(-2,2)$, $\beta_1 = 1$, $\delta = 8$, $\beta_0 = -0.85$, $p_m = 12\%$;

**Scenario 4.** $X \sim \text{Uniform}(-2, 2)$, $\beta_1 = 1$, $\delta = 8$, $\beta_0 = 0.38$, $p_m = 40\%$;

**Scenario 5.** $X \sim \text{Normal}(0, (\sqrt{4/3})^2)$, $\beta_1 = 1$, $\delta = 4$, $\beta_0 = -0.77$, $p_m = 12\%$;

**Scenario 6.** $X \sim \text{Normal}(0, (\sqrt{4/3})^2)$, $\beta_1 = 1$, $\delta = 4$, $\beta_0 = 0.42$, $p_m = 40\%$;

**Scenario 7.** $X \sim \text{Normal}(0, (\sqrt{4/3})^2)$, $\beta_1 = 1$, $\delta = 8$, $\beta_0 = -0.73$, $p_m = 12\%$;

**Scenario 8.** $X \sim \text{Normal}(0, (\sqrt{4/3})^2)$, $\beta_1 = 1$, $\delta = 8$, $\beta_0 = 0.44$, $p_m = 40\%$.

In all scenarios, the variance of $X$ remains the same, and we consider small and moderate values for $p_m$. For each simulated dataset, $\boldsymbol{\theta} = (\beta_0, \beta_1, \delta)^T$ were estimated by the five methods discussed in the previous section.

Results: Simulation results for scenarios $1-4$ are presented in Figures 5.2-5.5, respectively. We do not present the results for scenarios $5-8$ as their comparative performance was similar to that of scenarios $1-4$. I shall present the boxplots of estimates for each parameter ($\beta_0 \equiv$ intercept, $\beta_1 \equiv$ slope, $\delta \equiv$ skewness) with the empirical coverage probability for the 95% nominal level of significance. We note that the scales of the y-axis might be different so that direct comparisons needs to be done with caution. All results are based on $1,000$ replications. The empirical coverage probability was calculated using Wald-type confidence intervals, where the standard errors were calculated by inverting the Fisher information matrix. For the bootstrap approach (method B), I have used 200 bootstrap samples.

For the estimation of the intercept ($\beta_0$) and the slope ($\beta_1$) parameters, under large sample size (when $n = 1000$ or 2000), method N performs the best across all the scenarios in terms of the bias and variability. Under small sample size (when $n = 200$ or 500), however, method J is comparable or better than method N, in the sense that method J shows less variability with similar of less bias. The bias and variability of methods B and C are poor when the sample size is small, while they get better as the sample size increases. The performance of method G is poor as its bias does not decrease with the sample size.

For the skewness parameter ($\delta$) estimation, method J outperforms all methods across almost all the scenarios. Under small sample size, boxplots corresponding to method N do not fit in the

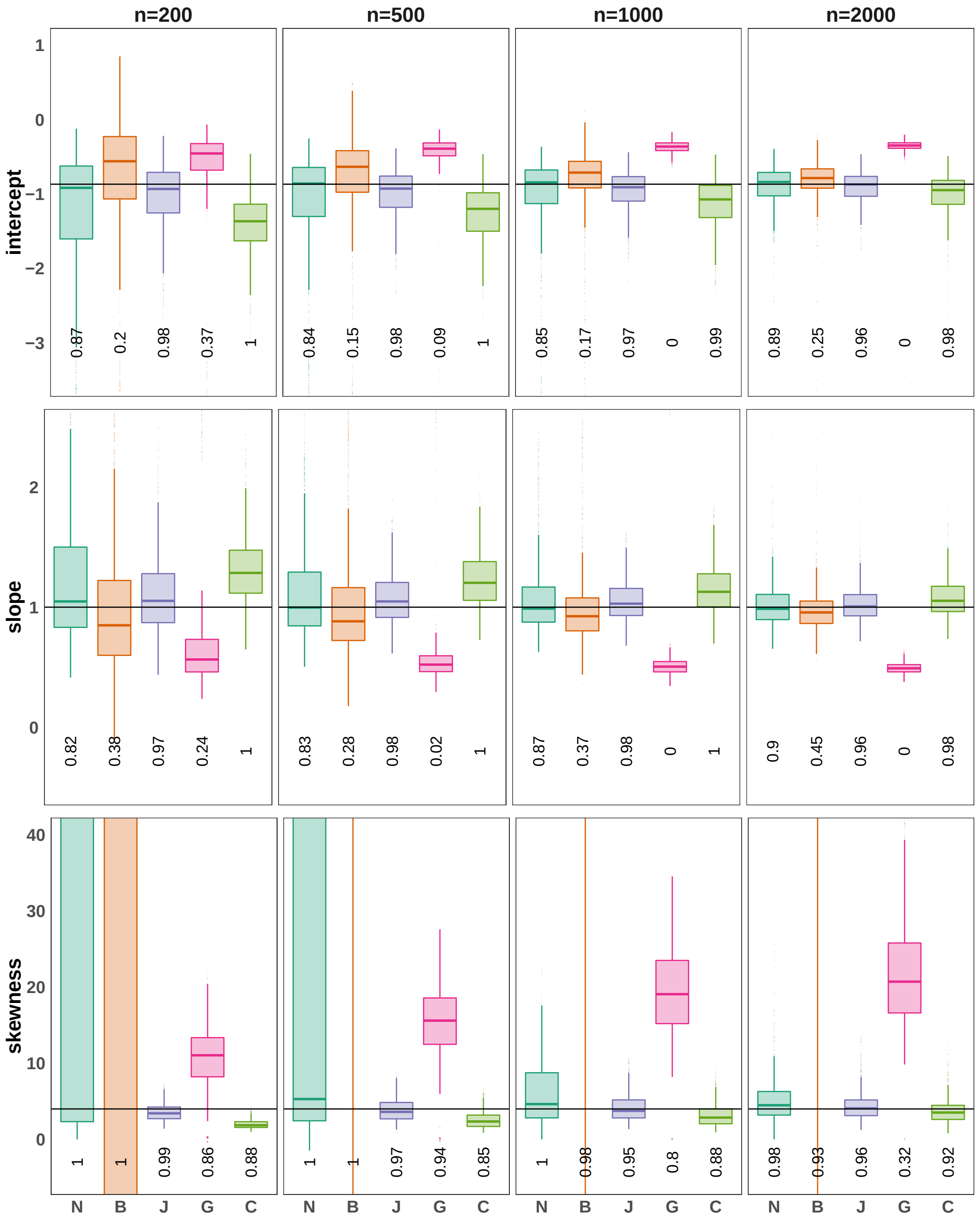

Figure 5.2: Simulation results based on 1000 replications when $X \sim \text{Uniform}(-2, 2)$, $\delta = 4$, $\beta_0 = -0.87$, $\beta_1 = 1$, and $p_m = 12\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey's prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

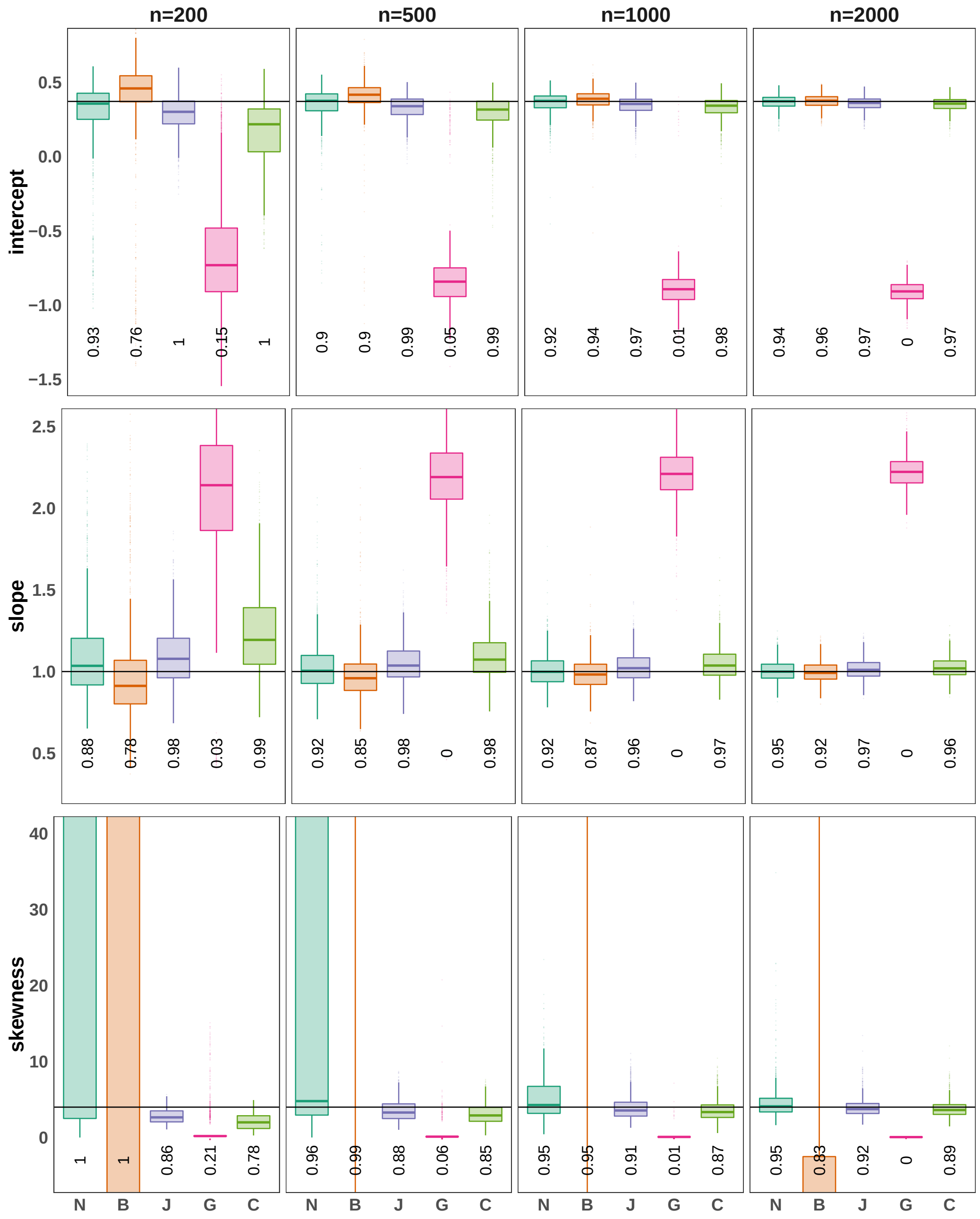

Figure 5.3: Simulation results based on 1000 replications when $X \sim \text{Uniform}(-2, 2)$, $\delta = 4$, $\beta_0 = 0.37$, $\beta_1 = 1$, and $p_m = 40\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey's prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

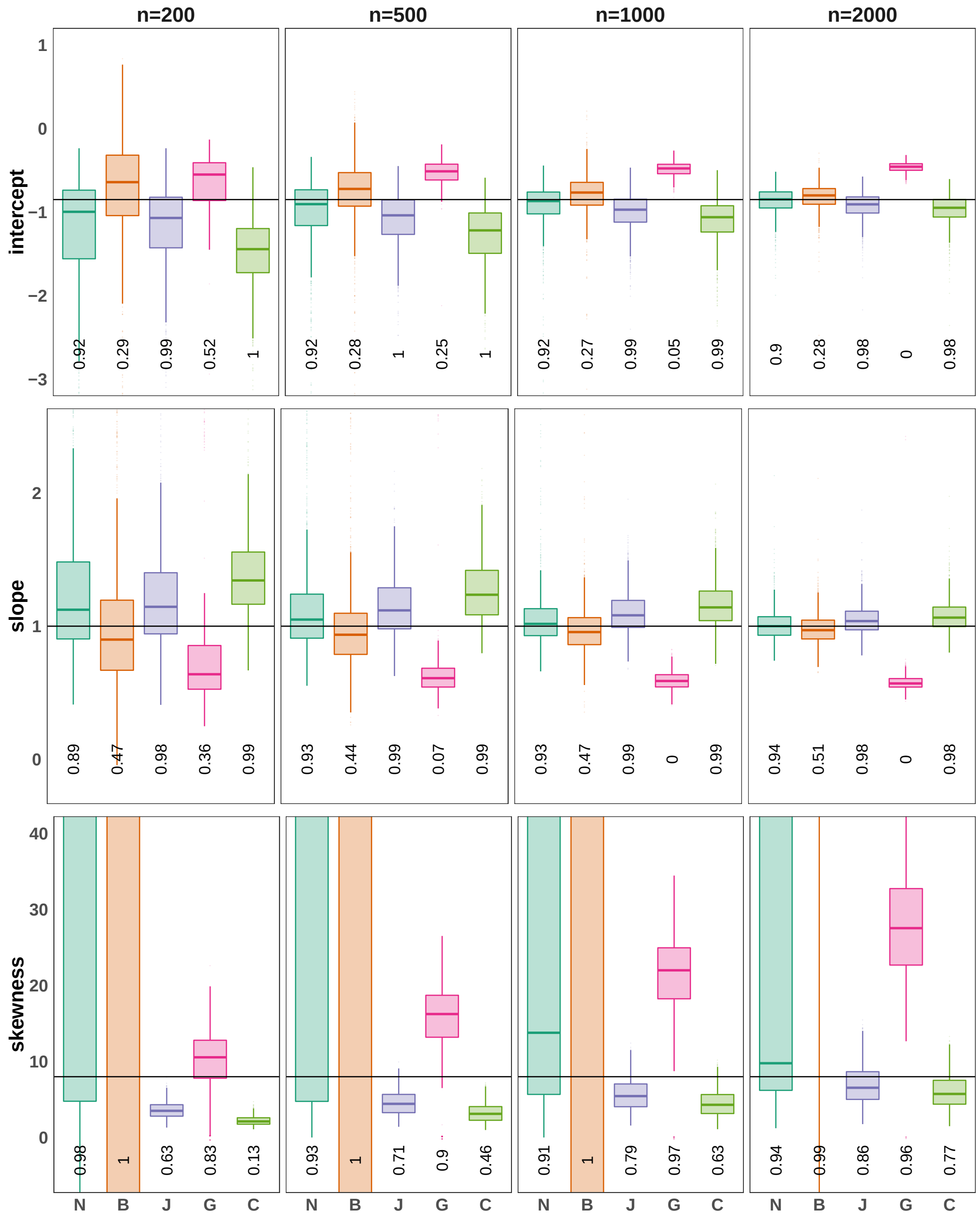

Figure 5.4: Simulation results based on 1000 replications when $X \sim \text{Uniform}(-2, 2)$, $\delta = 8$, $\beta_0 = -0.85$, $\beta_1 = 1$, and $p_m = 12\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey's prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

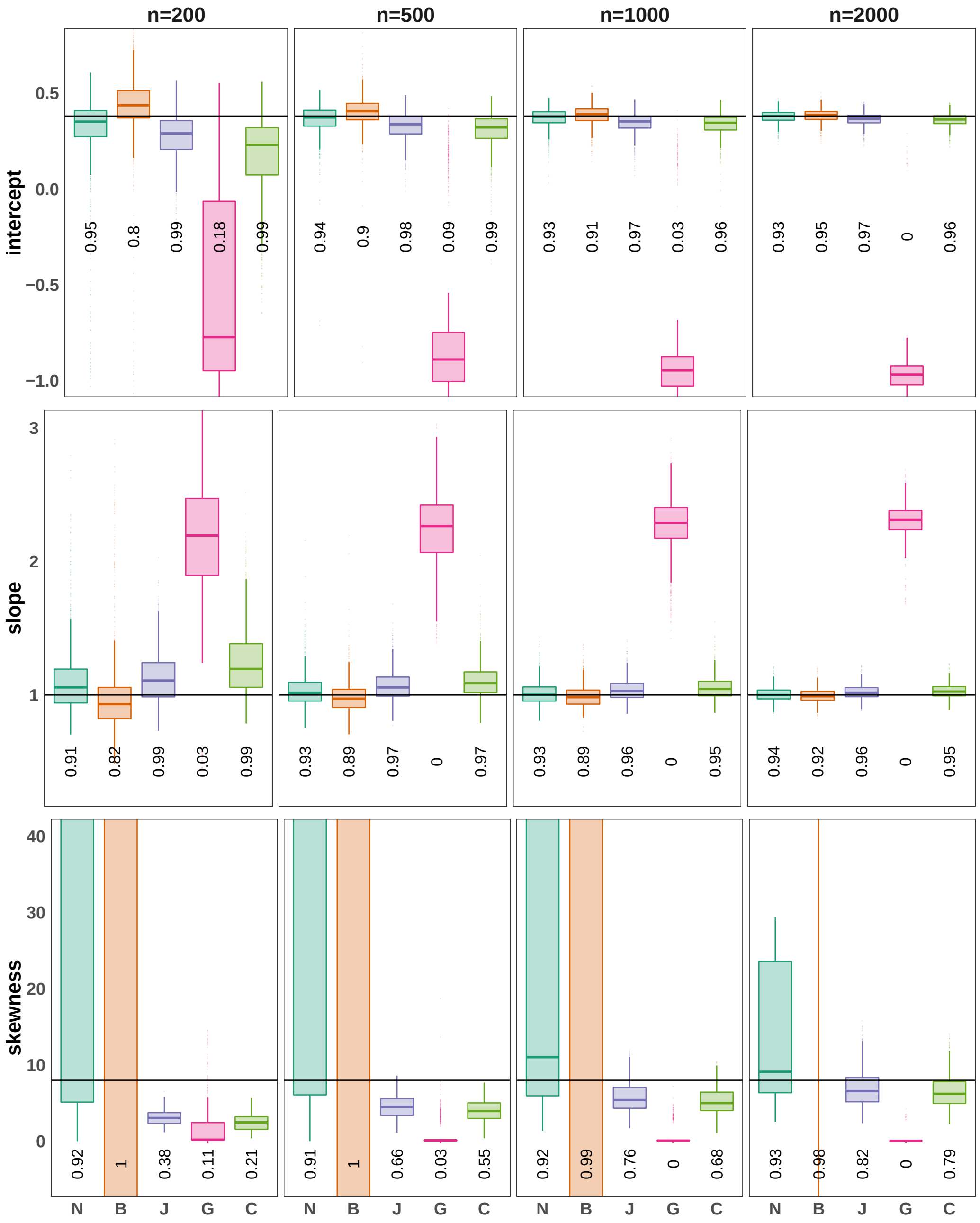

Figure 5.5: Simulation results based on 1000 replications when $X \sim \text{Uniform}(-2, 2)$, $\delta = 8$, $\beta_0 = 0.38$, $\beta_1 = 1$, and $p_m = 40\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey's prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

extended y-axis scale. The bias of method C is larger than that of method J for small sample sizes, but they become closer for larger sample sizes. On the contrary, methods B and G seem unreliable for $\delta$ estimation.

For parameters $\beta_0$ and $\beta_1$, the empirical coverage probabilities seem to be close to their nominal levels for methods N, J, and C. When $\delta = 8$, the empirical coverage probabilities for $\delta$ based on methods J and C are somewhat smaller than their nominal level. However, as the sample size increases, they become closer to the nominal level. However, with the smaller $\delta$, method J produces coverage probabilities close to the nominal level for different $n$.

In summary we can make the following conclusions. The maximum likelihood estimator has a skewed distribution, especially for small to moderate sample sizes. In general, the bootstrap bias corrected MLE (method B) does not show any better performance than method N. Rather, in some cases method B was worse than method N. Method J seems to be the best performing method for reducing the bias and variability of the MLE for all parameters regardless of the marginal success probability. For a large sample size, the performance of method C becomes similar to that of method J. The results indicate that generally the variability of the estimator decreases as $p_m$ increases.

### 5.5 Application to heart-disease data

For the illustration purpose, we analyze the heart-disease data from the Cleveland database (Detrano et al., 1989). The dataset can be found in UCI database (Dua and Karra Taniskidou, 2017). The goal of this analysis is to fit a model that explains the association between $Y$, the occurrence of a $> 50\%$ diameter narrowing in an angiography, and other clinical and test variables. In our analysis we consider subjects who have complete observations without any missing values. With this definition we have a total of 297 subjects out of 303 subjects in our analysis and 137 (46.13%) of them experienced the primary event. Among 13 available covariates, we choose the following 6 covariates which are statistically significant at the 5% level from a probit model: gender (Gender), chest pain type (CP), resting blood pressure (BP), the slope of the peak exercise ST segment (Slope), number of major vessels colored by flourosopy (CF), and thallium heart scan

Table 5.1: Results of the analysis of the heart-disease data. Est: Estimate, CI: the Wald confidence interval where standard errors are calculated by inverting the Fisher information matrix, P: probit model with MLE, N: skew-probit model with MLE, B: skew-probit model with bootstrap bias correction, J, G, C: skew-probit model with Jeffrey's prior, generalized information matrix, and Cauchy prior penalization, respectively.

| | | Method | | | | | |
|---|---|---|---|---|---|---|---|
| Covariates | | P | N | B | J | G | C |
| $\delta$ | Est | – | 1.540 | 0.954 | 2.730 | 0.139 | 1.468 |
| | CI | – | (−0.353, 3.433) | (−2.833, 4.741) | (0.566, 4.893) | (−0.002, 0.279) | (−0.166, 3.103) |
| Intercept | Est | −0.356 | 0.382 | 0.569 | 0.481 | −0.305 | 0.364 |
| | CI | (−0.816, 0.104) | (−0.062, 0.827) | (−0.569, 1.708) | (0.191, 0.771) | (−0.801, 0.190) | (−0.073, 0.801) |
| Gender | Est | 0.815 | 0.608 | 0.515 | 0.501 | 1.101 | 0.597 |
| | CI | (0.315, 1.315) | (0.197, 1.018) | (−0.107, 1.137) | (0.197, 0.806) | (0.507, 1.695) | (0.200, 0.993) |
| $\text{CP}_{\text{TA}}$ | Est | −1.355 | −0.985 | −0.791 | −0.794 | −1.561 | −0.959 |
| | CI | (−2.055, −0.654) | (−1.604, −0.366) | (−1.643, 0.060) | (−1.237, −0.350) | (−2.306, −0.816) | (−1.541, −0.378) |
| $\text{CP}_{\text{AA}}$ | Est | −0.917 | −0.680 | −0.597 | −0.582 | −0.604 | −0.673 |
| | CI | (−1.481, −0.353) | (−1.132, −0.228) | (−1.276, 0.083) | (−0.905, −0.259) | (−1.206, −0.003) | (−1.115, −0.230) |
| $\text{CP}_{\text{NA}}$ | Est | −1.272 | −0.911 | −0.804 | −0.728 | −1.542 | −0.904 |
| | CI | (−1.754, −0.790) | (−1.413, −0.409) | (−1.630, 0.021) | (−1.068, −0.389) | (−2.113, −0.972) | (−1.373, −0.435) |
| BP | Est | 1.959 | 1.420 | 1.167 | 1.154 | 2.148 | 1.316 |
| | CI | (0.458, 3.459) | (0.210, 2.631) | (−0.391, 2.725) | (0.212, 2.095) | (0.573, 3.723) | (0.177, 2.455) |
| $\text{Slope}_{\text{U}}$ | Est | −0.963 | −0.697 | −0.604 | −0.551 | −1.206 | −0.695 |
| | CI | (−1.398, −0.528) | (−1.114, −0.281) | (−1.246, 0.039) | (−0.854, −0.248) | (−1.680, −0.731) | (−1.089, −0.301) |
| $\text{Slope}_{\text{D}}$ | Est | −0.230 | −0.204 | −0.184 | −0.190 | −0.379 | −0.192 |
| | CI | (−0.976, 0.515) | (−0.775, 0.367) | (−0.795, 0.428) | (−0.665, 0.285) | (−1.311, 0.554) | (−0.760, 0.376) |
| CF | Est | 0.666 | 0.514 | 0.445 | 0.433 | 0.839 | 0.516 |
| | CI | (0.416, 0.917) | (0.283, 0.746) | (0.041, 0.848) | (0.259, 0.607) | (0.595, 1.084) | (0.295, 0.738) |
| $\text{Thal}_{\text{F}}$ | Est | 0.051 | 0.009 | 0.026 | −0.029 | −0.105 | 0.024 |
| | CI | (−0.752, 0.855) | (−0.602, 0.620) | (−0.635, 0.686) | (−0.546, 0.488) | (−0.878, 0.668) | (−0.582, 0.630) |
| $\text{Thal}_{\text{R}}$ | Est | 0.820 | 0.602 | 0.526 | 0.492 | 0.791 | 0.613 |
| | CI | (0.383, 1.257) | (0.210, 0.993) | (−0.034, 1.086) | (0.196, 0.788) | (0.344, 1.237) | (0.230, 0.995) |

results (Thal). We create relevant dummy variables for the categorical covariates, $\text{Gender} = 1$ for male and $0$ for female; $\text{CP}_{\text{TA}}$, $\text{CP}_{\text{AA}}$ and $\text{CP}_{\text{NA}}$ are dummies for chest pain types, typical angina, atypical angina, and non-anginal pain, respectively with asymptomatic being the reference; $\text{Slope}_{\text{U}}$ and $\text{Slope}_{\text{D}}$ are dummies for upsloping and downsloping of ST segment with flatness as the reference; and $\text{Thal}_{\text{F}}$ and $\text{Thal}_{\text{R}}$ are dummies for fixed detect and reversible detect while normal is considered as the reference category for Thal. Here BP and CF are continuous. We first fit the probit regression model to this dataset, however, the Hosmer and Lemeshow (Hosmer and Lemesbow, 1980) goodness-of-fit test $\widehat{C}$ based upon separations of predicted probabilities indicates a lack of fit of the assumed model at the $5\%$ level ($p$-value = 0.005).

Table 5.2: The $p$-values from Hosmer and Lemeshow (HL) goodness-of-fit test statistics ($\widehat{C}$) of the heart-disease data. P: probit model with MLE, N: skew-probit model with MLE, B: skew-probit model with bootstrap bias correction, J, G, C: skew-probit model with Jeffrey's prior, generalized information matrix, and Cauchy prior penalization, respectively.

| | P | N | B | J | G | C |
|---|---|---|---|---|---|---|
| $\widehat{C}$ | 0.011 | 0.028 | 0.000 | 0.207 | 0.000 | 0.024 |

Next, I consider the skew-probit model and estimate the model parameters using methods N, B, J, G, and C. In Table 5.1 I provide the estimates and 95% Wald-type confidence interval for each parameter based on the standard error calculated from the Fisher information matrix (CI). The $p$-values of Hosmer and Lemeshow test (Table 5.2) indicate that method J fits the data well at the $5\%$ level. The 95% CI for $\delta$ based on method J indicates that $\delta$ is significantly different from 0 ($\widehat{\delta}$: 2.73 and $95\%$ CI: 0.566, 4.893). On the other hand, we note that the $95\%$ CIs for $\delta$ based on the other approaches indicate $\delta$ is not statistically significant. In terms of estimates for other covariates, male subjects have higher risk for heart-disease than female subjects while any kind of chest pain has a lower probability of heart-disease compared to the asymptomatic pain. Also, based on method J, BP, CF and $\mathrm{That}_{\mathrm{F}}$ turn out to be positively associated with the probability of $Y = 1$. Although, the statistical significance of regression parameters $\boldsymbol{\beta}$ (except the intercept) do not change across methods P, N, J, G, and C, method J yields narrower confidence intervals compared to other methods.

## 5.6 Conclusions

In this chapter, I have investigated parameter identifiability and bias of the MLE for the skew-probit model for a binary response variable. The identifiability results will guide researchers to craft their model more carefully for the skew-probit link function. Several bias reduction strategies have been considered, and through simulation studies I have compared the performance of different approaches. The simulation results indicate that the bias reduction strategies should not be used blindly without considering the marginal success probability of the response variable and the sample size. Finally, I have applied the proposed strategies to analyze a real dataset on heart-disease, and the results show that without a proper bias correction the asymmetry in the link function may turn out to be statistically non-significant. Overall this research and the simulation results will help to develop a unique and robust method of analaysis to analyze models involving the skew-probit model.

# 6. CONCLUSION AND FUTURE WORK

In this dissertation, I construct a nonparametric testing method for homogeneity of distributions when we have multiple surrogates for the true signal. Also, I propose a Bayesian test for the goodness-of-fit of the distributional assumption of the technical inefficiency in the stochastic frontier analysis. In computational aspects, I develop fast but reliable computation of the standard error of Bayes estimator. Finally, I investigate the identifiability and bias reduction of a binary regression model with skew-probit link function.

The proposed testing method in Chapter 2 can be applied to check the homogeneity of the underlying distributions for contaminated data from various areas. When it is impossible to measure the true signal correctly, we tend to gather data multiple times. Under the classical measurement error assumption with symmetry of errors, deconvoluted characteristic function is a key building block to establish a test statistic. Bootstrap approximation of the test statistic is proposed and validated. Although the test is robust the symmetric error assumption, we can relieve the assumption. In addition, we can extend to the multivariate version of test.

Although the stochastic frontier model is widely used to analyze economic data, it should also be applicable to examine biomedical or epidemiological data due to the flexibility of technical inefficiency. In Chapter 3, I inspect the distributional assumption of the inefficiency in the model and develop a Bayesian test as a diagnostic tool. In the literature, however, the other component of the composed error is assumed to be the Normal distribution and it is also necessary to check this assumption.

Frequentist standard error is an important measure of variability of estimators and a key part in statistical inference. The derivation of standard errors in the Bayesian context is considered in Chapter 4. Because of bootstrap procedure and MCMC steps, the computation time was a big problem. By using the importance sampling approach, we can reduce the computation time significantly. This topic can also be tackled by various ways of bootstrap and MCMC scheme.

Finally, two problems of a binary regression with skew-normal link function are addressed in

Chapter 5. This link deviates from the probit link function with respect to a flexible skewness parameter. However, naive estimate for the skewness parameter is likely to be biased. It is recommended to use penalization approach with Jeffrey's prior type penalty function. Additionally, skew normal distribution is generalized to skew-*t* distribution or skew-elliptical distributions. They are another potential candidates to take into account asymmetric link function in a binary regression.

# REFERENCES

Aigner, D., Lovell, C. K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of econometrics*, 6(1):21–37.

Alba, M., Barrera, D., and Jiménez, M. (2001). A homogeneity test based on empirical characteristic functions. *Computational Statistics*, 16(2):255–270.

Anderson, T. (1962). On the distribution of the two-sample cramer-von mises criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159.

Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory (Springer Texts in Statistics)*. Springer-Verlag New York, Inc.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 12(2):171–178.

Azzalini, A. and Arellano-Valle, R. B. (2013). Maximum penalized likelihood estimation for skew-normal and skew-t distributions. *Journal of Statistical Planning and Inference*, 143(2):419–433.

Bayes, C. L. and Branco, M. D. (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics*, 21:141–163.

Bazán, J. L., Branco, M. D., Bolfarine, H., et al. (2006). A skew item response model. *Bayesian analysis*, 1(4):861–892.

Bazán, J. L., Romeo, J. S., Rodrigues, J., et al. (2014). Bayesian skew-probit regression for binary response data. *Brazilian Journal of Probability and Statistics*, 28(4):467–482.

Berger, J. (2006). The case for objective bayesian analysis. *Bayesian analysis*, 1(3):385–402.

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., et al. (1994). An overview of robust bayesian analysis. *Test*, 3(1):5–124.

Bhattacharya, J., Goldman, D., and McCaffrey, D. (2006). Estimating probit models with self-selected treatments. *Statistics in medicine*, 25(3):389–413.

Carlin, B. P. and Louis, T. A. (2008). *Bayesian Methods for Data Analysis*. CRC Press.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186.

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.

Castro, L. M., San Martín, E., and Arellano-Valle, R. B. (2013). A note on the parameterization of multivariate skewed-normal distributions. *Brazilian Journal of Probability and Statistics*, 27(1):110–115.

Chen, M.-H., Dey, D. K., and Shao, Q.-M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, 94(448):1172–1186.

Chen, Y.-T. and Wang, H.-J. (2012). Centered-residuals-based moment estimator and test for stochastic frontier models. *Econometric Reviews*, 31(6):625–653.

Chib, S. and Jeliazkov, I. (2005). Accept–reject metropolis–hastings sampling and marginal likelihood estimation. *Statistica Neerlandica*, 59(1):30–44.

Coelli, T. (1995). Estimators and hypothesis tests for a stochastic frontier function: A monte carlo analysis. *Journal of productivity analysis*, 6(3):247–268.

Cox, D. and Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 30(2):248–275.

Cramér, H. (1928). On the composition of elementary errors. *Skandinavisk Aktuarietidskrift*, 11:13–74 and 141–180.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 46(3):440–464.

Darling, D. (1957). The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838.

Delaigle, A. and Hall, P. (2016). Methodology for non-parametric deconvolution when the error distribution is unknown. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):231–252.

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *The*

*Annals of Statistics*, 36(2):665–685.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., and Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310.

Dua, D. and Karra Taniskidou, E. (2017). UCI machine learning repository [http://archive.ics.uci.edu/ml]. University of California, Irvine, School of Information and Computer Sciences.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *Annals of Applied Statistics*, 6(4):1971–1997.

Efron, B. (2015). Frequentist accuracy of bayesian estimates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):617–646.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1:54–75.

Efron, B. and Tibshirani, R. (1994). *An introduction to the bootstrap*. CRC press.

Epps, T. W. and Pulley, L. B. (1983). A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19(3):1257–1272.

Fan, Y. (1997). Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function. *Journal of Multivariate Analysis*, 62(1):36–63.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Fisz, M. (1960). On a result by m. rosenblatt concerning the von mises-smirnov test. *The Annals of Mathematical Statistics*, 31(2):427–429.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607.

Fuller, W. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Genton, M. G., He, L., and Liu, X. (2001). Moments of skew-normal random vectors and their quadratic forms. *Statistics & probability letters*, 51(4):319–325.

Genton, M. G. and Zhang, H. (2012). Identifiability problems in some non-gaussian spatial random fields. *Chilean Journal of Statistics*, 3(2):171–179.

Ghosal, S., Lember, J., Van Der Vaart, A., et al. (2008). Nonparametric bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.

Greene, W. H. (1990). A gamma-distributed stochastic frontier model. *Journal of econometrics*, 46(1-2):141–163.

Griffin, J. E. and Steel, M. F. (2004). Semiparametric bayesian inference for stochastic frontier models. *Journal of econometrics*, 123(1):121–152.

Gupta, M. and Ibrahim, J. G. (2009). An information matrix prior for bayesian analysis in generalized linear models with high dimensional data. *Statistica Sinica*, 19(4):1641–1663.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Chapman and Hall/CRC.

Hall, P. and Lahiri, S. N. (2008). Estimation of distributions, moments and quantiles in deconvolution problems. *The Annals of Statistics*, 36(5):2110–2134.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–2419.

Henze, N. (1986). A probabilistic representation of the'skew-normal'distribution. *Scandinavian journal of statistics*, 13(4):271–275.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30.

Hogan, S. L., Vupputuri, S., Guo, X., Cai, J., Colindres, R. E., Heiss, G., and Coresh, J. (2007).

Association of cigarette smoking with albuminuria in the united states: the third national health and nutrition examination survey. *Renal failure*, 29(2):133–142.

Hosmer, D. W. and Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10):1043–1069.

Hu, F. and Hu, J. (2000). A note on breakdown theory for bootstrap methods. *Statistics & probability letters*, 50(1):49–53.

Huber, P. and Ronchetti, E. (2009). *Robust Statistics*. John Wiley & Sons, New York.

Ibrahim, J., Chen, M., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer.

Jiménez-Gamero, M.-D., Alba-Fernández, V., Muñoz-García, J., and Chalco-Cano, Y. (2009). Goodness-of-fit tests based on empirical characteristic functions. *Computational Statistics and Data Analysis*, 53(12):3957–3971.

Jondrow, J., Lovell, C. K., Materov, I. S., and Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of econometrics*, 19(2-3):233–238.

Jones, G. L. (2004). On the markov chain central limit theorem. *Probability Surveys*, 1:299–320.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Kiefer, J. (1959). K-sample analogues of the kolmogorov-smirnov and cramer-v. mises tests. *The Annals of Mathematical Statistics*, 30(2):420–447.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione (on the empirical determination of a distribution law). *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91.

Kooperberg, C. and Stone, C. J. (1991). A study of logspline density estimation. *Computational Statistics & Data Analysis*, 12(3):327–347.

Kooperberg, C. and Stone, C. J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1(4):301–328.

Kopp, R. J. and Mullahy, J. (1990). Moment-based estimation and testing of stochastic frontier models. *Journal of Econometrics*, 46(1-2):165–183.

Kreiss, J.-P. and Lahiri, S. (2012). Bootstrap methods for time series. *Handbook of Statistics: Time Series Analysis: Methods and Applications*, 30(1).

Kuiper, N. H. (1960). Tests concerning random points on a circle. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A*, 63:38–47.

Kumbhakar, S. C. and Lovell, C. K. (2003). *Stochastic frontier analysis*. Cambridge university press.

Lahiri, S. N. (1994). On two-term edgeworth expansions and bootstrap approximations for studentized multivariate m-estimators. *Sankhyā: The Indian Journal of Statistics, Series A*, 56(2):201–226.

Lewis, S. M. and Raftery, A. E. (1997). Estimating bayes factors via posterior simulation with the laplaceâĂŤmetropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655.

Liang, F. (2002). Dynamically weighted importance sampling in Monte Carlo computation. *Journal of the American Statistical Association*, 97(459):807–821.

Liseo, B. and Loperfido, N. (2006). A note on reference priors for the scalar skew-normal distribution. *Journal of Statistical Planning and Inference*, 136(2):373–389.

Ma, Y. and Genton, M. G. (2004). Flexible class of skew-symmetric distributions. *Scandinavian Journal of Statistics*, 31(3):459–468.

Maag, U. R. and Stephens, M. A. (1968). The $\mathbf{V}_{NM}$ two-sample test. *The Annals of Mathematical Statistics*, 39(3):923–935.

Marín, J. M. (2000). A robust version of the dynamic linear model with an economic application. In *Robust Bayesian Analysis*, pages 373–383. Springer.

Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from cobb-douglas production functions with composed error. *International Economic Review*, 18(2):435–44.

Meng, X.-L. and Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586.

Mira, A. and Nicholls, G. (2004). Bridge estimation of the probability density at a point. *STATISTICA SINICA*, 14(2):603–612.

Ni, S., Sun, D., and Sun, X. (2007). Intrinsic bayesian estimation of vector autoregression impulse responses. *Journal of Business & Economic Statistics*, 25(2):163–176.

Otiniano, C., Rathie, P., and Ozelim, L. (2015). On the identifiability of finite mixture of skew-normal and skew-t distributions. *Statistics & Probability Letters*, 106:103–108.

Pettitt, A. N. (1976). A two-sample anderson-darling rank statistic. *Biometrika*, 63(1):161–168.

Primatesta, P., Falaschetti, E., Gupta, S., Marmot, M. G., and Poulter, N. R. (2001). Association between smoking and blood pressure: evidence from the health survey for england. *Hypertension*, 37(2):187–193.

Puddey, I. B. and Beilin, L. J. (2006). Alcohol is bad for blood pressure. *Clinical and Experimental Pharmacology and Physiology*, 33(9):847–852.

Robert, C. and Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer New York.

Rosenblatt, M. (1952). Limit theorems associated with variants of the von mises statistic. *The Annals of Mathematical Statistics*, 23(4):617–623.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39(3):577–91.

Salibian-Barrera, M. and Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statistics*, 30(2):556–582.

Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *Journal of Statistical Planning and Inference*, 136(12):4259–4275.

Schmidt, P. and Lin, T.-F. (1984). Simple tests of alternative specifications in stochastic frontier models. *Journal of Econometrics*, 24(3):349–361.

Scholz, F. W. and Stephens, M. A. (1987). K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924.

Shao, J. (1999). *Mathematical Statistics*. Springer-Verlag New York.

Sheather, S. J. (2004). Density estimation. *Statistical Science*, 19(4):588–597.

Simar, L., Van Keilegom, I., and Zelenyuk, V. (2017). Nonparametric least squares methods for stochastic frontier models. *Journal of Productivity Analysis*, 47(3):189–204.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Singh, K. (1998). Breakdown theory for bootstrap quantiles. *The annals of Statistics*, 26(5):1719–1732.

Smirnov, N. V. (1936). Sur la distribution de $\omega^2$. *Comptes Rendus de l'Académie des Sciences Paris*, 202:449–452.

Smirnov, N. V. (1937). Sur la distribution de $\omega^2$ (criterion of von mises). *Recueil Mathematique (Matematiceskii Sbornik), NS*, 2(44):973–993.

Smirnov, N. V. (1939a). Ob uklonenijah empiriceskoi krivoi raspredelenija. *Recueil Mathematique (Matematiceskii Sbornik), NS*, 6(48):3–26.

Smirnov, N. V. (1939b). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathèmatique de L'Universitè de Moscow*, 2(2):3–14.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvolving kernel density estimators. *Statistics*, 21(2):169–184.

Stephens, M. (1992). Introduction to kolmogorov (1933) on the empirical determination of a distribution. In *Breakthroughs in statistics*, pages 93–105. Springer.

Stevenson, R. E. (1980). Likelihood functions for generalized stochastic frontier estimation. *Journal of econometrics*, 13(1):57–66.

Stingo, F. C., Stanghellini, E., and Capobianco, R. (2011). On the estimation of a binary response model in a selected population. *Journal of statistical planning and inference*, 141(10):3293–3303.

Stock, J. H. and Watson, M. W. (2001). Vector autoregressions. *The Journal of Economic Perspectives*, 15(4):101–115.

van den Broeck, J., Koop, G., Osiewalski, J., and Steel, M. F. (1994). Stochastic frontier models: A bayesian perspective. *Journal of Econometrics*, 61(2):273–303.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With*

*Applications to Statistics*. Springer New York.

von Mises, R. (1931). *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*. Mary S. Rosenberg.

Wang, W. S., Amsler, C., and Schmidt, P. (2011). Goodness of fit tests in stochastic frontier models. *Journal of Productivity Analysis*, 35(2):95–118.

Weinberg, M. D. (2012). Computing the bayes factor from a markov chain monte carlo simulation of the posterior distribution. *Bayesian Analysis*, 7(3):737–770.

Willems, G. and Van Aelst, S. (2005). Fast and robust bootstrap for LTS. *Computational Statistics & Data Analysis*, 48(4):703–715.

Zhang, J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):281–294.

Zhang, J. (2006). Powerful two-sample tests based on the likelihood ratio. *Technometrics*, 48(1):95–103.

# APPENDIX A

# PROOFS OF THEOREMS IN CHAPTER 2

Let $c_{jw}(t) = \cos(t\overline{W}_j), d_{jw}(t) = \sin(t\overline{W}_j)$. Next define $e_{jw}(t) = M_x^{-1}\sum_{(l_1,l_2)\in\mathcal{S}_x}\cos\{(t/m_x)( W_{jl_1} - W_{jl_2})\}$. Denote the expectations by $c_{0w}(t) = E\{c_{jw}(t)\}, d_{0w}(t) = E\{d_{1w}(t)\}$, and $e_{0w}(t) = E\{e_{1w}(t)\}$. Then, $\boldsymbol{\Lambda}_{\boldsymbol{W}_j}(t) \equiv (c_{jw}(t) - c_{0w}(t), d_{jw}(t) - d_{0w}(t), e_{jw}(t) - e_{0w}(t))^T$ are iid mean zero random vectors. Similarly define $\boldsymbol{\Lambda}_{\boldsymbol{V}_j}(t)$ by replacing $\boldsymbol{W}_j$'s by $\boldsymbol{V}_j$'s in the definition of $\boldsymbol{\Lambda}_{\boldsymbol{W}_j}(t)$. Let $t_0 = \max\{|t_1|, |t_2|\}$, where recall that $\omega(t) = 0$ for all $t \notin [t_1, t_2]$. Define

$$\boldsymbol{Z}_n(t) = n_x^{-1/2}\begin{pmatrix} \sum_{j=1}^{n_x}\boldsymbol{\Lambda}_{\boldsymbol{W}_j}(t) \\ \sum_{j=1}^{n_y}\boldsymbol{\Lambda}_{\boldsymbol{V}_j}(t) \end{pmatrix}, \ |t| \le t_0.$$

Let $C, C(\cdot)$ denote generic constants with values in $(0, \infty)$ that may depend on their arguments (if any) but not on $n_x, n_y$. Also, let $\ell^\infty[-t_0, t_0]$ denote the set of all bounded measurable functions from $[-t_0, t_0]$ to the real line and let $\|x\|_\infty = \sup\{|x(t)| : t \in [-t_0, t_0]\}$, $x \in \ell^\infty[-t_0, t_0]$. Finally, let $A^{\mathrm{T}}$ denote the transpose of a matrix (vector) $A$.

Then we have the following result.

**Lemma 1.** *$\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{Z}$ as random elements of the space $(l^\infty[-t_0, t_0])^6$, where $\boldsymbol{Z}$ is a 6-dimensional zero-mean Gaussian process on $[-t_0, t_0]$ with the covariance function*

$$\Gamma(s,t) = \begin{bmatrix} \Gamma_w(s,t) & 0 \\ 0 & \rho^{-2}\Gamma_v(s,t) \end{bmatrix},$$

*with $\Gamma_v(s,t) = E\{\boldsymbol{\Lambda}_{\boldsymbol{W}_1}(s)\boldsymbol{\Lambda}_{\boldsymbol{W}_1}(t)\}$, $\Gamma_v(s,t) = E\{\boldsymbol{\Lambda}_{\boldsymbol{V}_1}(s)\boldsymbol{\Lambda}_{\boldsymbol{V}_1}(t)\}$, for $-t_0 \le s, t \le t_0$. Further, the paths of $\boldsymbol{Z}(\cdot)$ are continuous on $[-t_0, t_0]$ with probability one.*

*Proof.* Note that $i)$ $\boldsymbol{\Lambda}_{\boldsymbol{W}_j}(t)$ and $\boldsymbol{\Lambda}_{\boldsymbol{V}_j}(t)$ are bounded random vectors, $ii)$ the collection of functions $\{(\boldsymbol{\Lambda}_{\boldsymbol{w}}(t), \boldsymbol{\Lambda}_{\boldsymbol{v}}(t)); t \in [-t_0, t_0]\}$ is a VC-class, where $\boldsymbol{\Lambda}_{\boldsymbol{w}}(t) = [\cos(t\sum_{j=1}^{m_x} w_j/m_x), \sin$

$(t\sum_{j=1}^{m_x} w_j/m_x), M_x^{-1}\sum_{(l_1,l_2)\in S}\cos\{t(w_{l_1}-w_{l_2})/m_x\}]$, and $\boldsymbol{\Lambda_v}(t)$ is defined similarly. Hence, by using the Multivariate CLT (cf. Ch 11.1, Athreya and Lahiri, 2006), the finite dimensional distribution of the $\boldsymbol{Z}_n(\cdot)$-process converges in distribution to those of the $\boldsymbol{Z}(\cdot)$-process. Further, using the standard exponential inequalities (e.g., Hoeffding, 1963) and the chaining argument (van der Vaart and Wellner, 1996), it follows that $\boldsymbol{Z}_n \to \boldsymbol{Z}$ in distribution, where $\boldsymbol{Z}$ is a random element of $l^\infty([-t_0,t_0])]^6$ and it has continuous paths on $[-t_0,t_0]$ with probability one. □

Now I provide the proof of Theorems in Chapter 2.

*Proof of Theorem 1.* Recall the definitions of $\widehat{a}_x(t)$ and $\widehat{a}_{2x}(t)$ given in (2.4) and (2.5), respectively, and define $a_{2x}(t)=\phi_{u_x}^{m_x}(t/m_x)$ and let $Z_{kn}(t)$ be the $k$th component of $\boldsymbol{Z}_n(t)$ defined in Lemma 1. Then $a_x(t)=c_{0w}(t)/a_{2x}(t)$, and

$$\begin{aligned}
\sqrt{n_x}\{\widehat{a}_x(t)-a_x(t)\} &= \sqrt{n_x}\left\{\frac{n_x^{-1}\sum_{j=1}^{n_x}c_{jw}(t)}{\widehat{a}_{2x}(t)}-\frac{c_{0w}(t)}{a_{2x}(t)}\right\}\\
&= \sqrt{n_x}\left[\frac{n_x^{-1}\sum_{j=1}^{n_x}\{c_{jw}(t)-c_{0w}(t)+c_{0w}(t)\}}{\widehat{a}_{2x}(t)}-\frac{c_{0w}(t)}{a_{2x}(t)}\right]\\
&= \sqrt{n_x}\left[\frac{n_x^{-1}\sum_{j=1}^{n_x}\{c_{jw}(t)-c_{0w}(t)\}}{\widehat{a}_{2x}(t)}+\frac{c_{0w}(t)}{\widehat{a}_{2x}}-\frac{c_{0w}(t)}{a_{2x}(t)}\right]\\
&= \frac{Z_{1n}(t)}{\widehat{a}_{2x}(t)}-\frac{c_{0w}(t)\sqrt{n_x}\{\widehat{a}_{2x}(t)-a_{2x}(t)\}}{a_{2x}(t)\widehat{a}_{2x}(t)}. \qquad (A.1)
\end{aligned}$$

Now using the fact that $\widehat{a}_{2x}(t)=\{\phi_{u_x}^2(t/m_x)+Z_{3n}(t)/\sqrt{n_x}\}^{m_x/2}$, we get

$$\begin{aligned}
\sqrt{n_x}\{\widehat{a}_x(t)-a_x(t)\} &= \frac{Z_{1n}(t)}{a_{2x}(t)}-\frac{m_x c_{0w}(t)Z_{3n}(t)\phi_{u_x}^{m_x-2}(t/m_x)}{2a_{2x}^2(t)}+R_{nx}(t),\\
&\equiv A_{nx}(t)+R_{nx}(t), \qquad (A.2)
\end{aligned}$$

where

$$A_{nx}(t)=\frac{Z_{1n}(t)}{a_{2x}(t)}-\frac{m_x c_{0w}(t)Z_{3n}(t)\phi_{u_x}^{m_x-2}(t/m_x)}{2a_{2x}^2(t)}$$

and where, with a suitable constant $C(m_x) \in (0, \infty)$,

$$|R_{nx}(t)| \leq \frac{|Z_{1n}(t)||Z_{3n}(t)|}{\sqrt{n_x}} \times \frac{m_x\{1 + |Z_{3n}(t)/\sqrt{n_x}|^{m_x/2-1}\}}{2|a_{2x}(t)|\widehat{a}_{2x}(t)} \\ + \frac{|c_{0w}(t)|\{1 + |Z_{3n}(t)|^{m_x}\}}{|a_{2x}(t)|\sqrt{n_x}|\widehat{a}_{2x}(t)|} \times C(m_x) + \frac{m_x|c_{0w}||Z_{3n}(t)|^2\{1 + |Z_{3n}(t)/\sqrt{n_x}|^{m_x/2-1}\}}{2|a_{2x}(t)|^3\sqrt{n_x}|\widehat{a}_{2x}(t)|}.$$

Hence,

$$\begin{aligned} \int |R_{nx}(t)|^2 \omega(t) dt \quad \leq \quad & \frac{C(m_x)}{n_x}\bigg\{\int \frac{\omega(t)}{a_{2x}^2(t)} dt\bigg\}\bigg[||Z_{1n}||_\infty^2 ||Z_{3n}||_\infty^2 + \frac{\{1 + ||Z_{3n}||_\infty^{2m_x}\}}{\alpha_x^{4m_x}}\bigg] \\ & \times \frac{\{1 + (||Z_{3n}||_\infty/\sqrt{n_x})^{m_x/2-1}\}^2}{(\alpha_x^2 - ||Z_{3n}||_\infty/\sqrt{n_x})^{m_x}}, \end{aligned}$$

where $\alpha_x = \min\{|\phi_{u_x}(t/m_x)|; |t| \leq t_0\}$. Since $|| \cdot ||_\infty$ is continuous on $\ell^\infty[-t_0, t_0]$, it follows that $||Z_{kn}||_\infty \xrightarrow{d} ||Z_k||_\infty$ for $k = 1, \ldots, 6$. Hence

$$\int |R_{nx}(t)|^2 \omega(t) dt \to 0 \tag{A.3}$$

in probability. Next, we define $a_{2y}(t) = \phi_{u_y}^{m_y}(t/m_y)$ and write

$$\widehat{a}_{2y}(t) = \{\phi_{u_y}^2(t/m_y) + (\sqrt{n_x}/n_y) Z_{6n}(t)\}^{m_y/2}.$$

Then, using similar steps as above, we obtain

$$\sqrt{n_x}\{\widehat{a}_y(t) - a_y(t)\} \quad = \quad \frac{n_x}{n_y}\bigg\{A_{ny}(t) + R_{ny}(t)\bigg\},$$

where

$$A_{ny} = \frac{Z_{4n}(t)}{a_{2y}(t)} - \frac{m_y c_{0v}(t) Z_{6n}(t) \phi_{u_y}^{m_y-2}(t/m_y)}{2a_{2y}^2(t)},$$

and where, retracing arguments above, one can show that

$$\int |R_{ny}(t)|^2 \omega(t)dt \to 0 \tag{A.4}$$

in probability. Under $H_0 : \phi_x(t) = \phi_y(t)$, that means $a_x(t) = a_y(t)$ for all $t$. So, under $H_0$,

$$\begin{aligned} I_1 &= n_x \int \{\widehat{a}_x(t) - \widehat{a}_y(t)\}^2 \omega(t)dt \\ &= n_x \int [\{\widehat{a}_x(t) - a_x(t)\} - \{\widehat{a}_y(t) - a_y(t)\}]^2 \omega(t)dt \\ &= I_{11} + Q_n, \end{aligned} \tag{A.5}$$

where

$$I_{11} = \int \left\{ A_{nx}(t) - \frac{n_x}{n_y} A_{ny}(t) \right\}^2 \omega(t)dt,$$

and using the Cauchy-Schwartz inequality

$$|Q_n| \leq \int \{R_{nx}(t) + \frac{n_x}{n_y} R_{ny}(t)\}^2 \omega(t)dt + 2\left[ I_{11} \times \int \{R_{nx}(t) + \frac{n_x}{n_y} R_{ny}(t)\}^2 \omega(t)dt \right]^{1/2}. \tag{A.6}$$

By (A.3) and (A.4), $|Q_n| \to 0$ in probability. Next applying the continuous mapping theorem, we obtain

$$I_{11} \xrightarrow{d} I_{1\infty} \equiv \int \xi_1^2(t) \omega(t)dt, \tag{A.7}$$

where $\xi_1(t) = A_x(t) - \rho^2 A_y(t)$. Repeating the arguments above with $I_2 = n_x \int \{\widehat{b}_x(t) - \widehat{b}_y(t)\}^2 \omega(t)dt$ and using the joint weak convergence result of Lemma 1, one can show that

$$\begin{aligned} T_n &= I_1 + I_2 \\ &= I_{11} + \int \left[ \left\{ \frac{Z_{2n}(t)}{a_{2x}(t)} - \frac{m_x d_{0v}(t) Z_{3n}(t) \phi_{u_x}^{m_x - 2}(t/m_x)}{2a_{2x}^2(t)} \right\} \right. \end{aligned}$$

$$-\frac{n_x}{n_y}\Bigg\{\frac{Z_{5n}(t)}{a_{2y}(t)} - \frac{m_y d_{0v}(t) Z_{6n}(t)\phi_{u_y}^{m_y-2}(t/m_y)}{2a_{2y}^2(t)}\Bigg\}\Bigg]^2 \omega(t)dt + o_p(1)$$

$$\xrightarrow{d} \quad I_{1\infty} + \int\Bigg[\Bigg\{\frac{Z_2(t)}{a_{2x}(t)} - \frac{m_x d_{0v}(t) Z_3(t)\phi_{u_x}^{m_x-2}(t/m_x)}{2a_{2x}^2(t)}\Bigg\}$$

$$-\rho^2\Bigg\{\frac{Z_5(t)}{a_{2y}(t)} - \frac{m_y d_{0v}(t) Z_6(t)\phi_{u_y}^{m_y-2}(t/m_y)}{2a_{2y}^2(t)}\Bigg\}\Bigg]^2 \omega(t)dt$$

$$\equiv \int[\xi_1^2(t) + \xi_2^2(t)]\omega(t)dt.$$

This completes the proof of Theorem 1. □

*Proof of Theorem 2.* First suppose that $\int D_a^2(t)\omega(t)dt \neq 0$. Let $W_a(t) = \{\widehat{a}_x(t) - a_x(t)\} + \{a_y(t) - \widehat{a}_y(t)\}$, $|t| \leq t_0$. Then, it follows that

$$T_{1n_x} \equiv n_x \int \Big[\{\widehat{a}_x(t) - a_x(t)\} + \{a_x(t) - a_y(t)\} + \{a_y(t) - \widehat{a}_y(t)\}\Big]^2 \omega(t)dt \geq L_{1n_x},$$

where $L_{1n_x} = n_x \int\{a_x(t) - a_y(t)\}^2\omega(t)dt + 2n_x\int W_a(t)\{a_x(t) - a_y(t)\}\omega(t)dt$. Now, using the steps in the proof of Theorem 1 and the continuous mapping theorem, one can show that the second term of $L_{1n_x}$ is $O_p(\sqrt{n_x})$ while the first term diverges at the rate $n_x$. Thus, $L_{1n_x} = O_p(n_x)$. Hence, for

$$\text{pr}(T_{1n_x} \leq r) \leq \text{pr}(L_{1n_x} \leq r) \to 0 \text{ for any } r \in (0, \infty).$$

Next consider the case where $\int D_b^2(t)\omega(t)dt \neq 0$. Then, defining $T_{2n_x}$ by replacing $\widehat{a}_x$, $\widehat{a}_y$, $a_x, a_y$ in $T_{1n_x}$ by $\widehat{b}_x$, $\widehat{b}_y$, $b_x, b_y$ and using the arguments above, we have $\text{pr}(T_{2n_x} \leq r) \to 0$ for any $r \in (0, \infty)$. Thus, if $\int[D_a^2(t) + D_b^2(t)]\omega(t)dt \neq 0$, then for any $\alpha$,

$$\begin{aligned}\text{pr}(T_{n_x} > t_{n_x,\alpha}) &= 1 - \text{pr}(T_{1n_x} + T_{2n_x} \leq t_{n_x,\alpha}) \\ &\geq 1 - \min\big\{\text{pr}(T_{1n_x} \leq t_\alpha), \text{pr}(T_{2n_x} \leq t_\alpha)\big\} \to 1 \text{ as } n_x \to \infty,\end{aligned}$$

proving Theorem 2.

$\square$

*Proof of Theorem 3.* First we show that

$$\widehat{\phi}_x(t) \equiv \widehat{\phi}_{\overline{W}}(t)/\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x} = \widehat{\phi}_1(t)\phi_K(h_w t)/\{\widehat{\phi}_{u_x}(t/m_x)\}^{m_x}$$

converges to $\phi_x(t)$ uniformly over $|t| \leq t_0$, almost surely. Since $h_w \to 0$, it is enough to show that

$$\sup\{|\widehat{\phi}_1(t) - \phi_1(t)| : |t| \leq t_0\} \to 0 \text{ almost surely, and} \tag{A.8}$$

$$\sup\{|\widehat{\phi}_{u_x}(t) - \phi_{u_x}(t)| : |t| \leq t_0 m_x\} \to 0 \text{ almost surely.} \tag{A.9}$$

Since $\widehat{\phi}_1(t) = n_x^{-1} \sum_{j=1}^{n_x} \exp(it\overline{W}_j)$ is an average of i.i.d., bounded random variables, one can prove (A.8) using a discretization argument and Hoeffding's inequality (Hoeffding, 1963); see, e.g., Lahiri (1994). Next, for $h > 0$, write $e_{jw}(t,h) = M_x^{-1} \sum_{(l_1,l_2)\in\mathcal{S}_x} \cos\{(t/m_x)(W_{jl_1} - W_{jl_2})\}(1 - h^2t^2)^3 I(|ht| \leq 1)$ and $e_{0w}(t,h) \equiv E\{e_{jw}(t,h)\}$. Then, it is easy to check that $e_{0w}(t,h) = |\phi_{u_x}(t/m_x)|^2(1 - h^2t^2)^3 I(|ht| \leq 1)$, and hence, $\sup\{|e_{0w}(t,h_w) - \phi_{u_x}(t/m_x)| : |t| \leq t_0 m_x\} \to 0$, as $h_w \to 0$. Further, using arguments similar to those in the proof of (A.8), one can show that $\sup\{|\widehat{\phi}_{u_x}(t) - e_{0w}(t,h_w)| : |t| \leq t_0 m_x\} \to 0$, almost surely. Thus, (A.9) holds. Let $A$ be the event where (A.8) and (A.9) hold. Then pr$(A) = 1$. Next, let $B$ be the event where

$$\sup\{|\widehat{\phi}_2(t) - \phi_2(t)| : |t| \leq t_0\} \to 0, \text{ and}$$
$$\sup\{|\widehat{\phi}_{v_x}(t) - \phi_{v_x}(t)| : |t| \leq t_0 m_y\} \to 0,$$

as $n_x \to \infty$. Then, by similar arguments, pr$(B) = 1$, implying, pr$(A \cap B) = 1$.

We shall now show that $T^*_{n_x}$ converges in distribution to $T_\infty \equiv \int[[\xi_1^2(t) + \xi_2^2(t)]\omega(t)dt$, i.e., the Prohorov distance between the Bootstrap probability distribution of $T^*_{n_x}$ and the the probability distribution of $T_\infty$ goes to zero, on the set $A \cap B$. Let $\boldsymbol{Z}^*_n(t)$ be defined by replacing $(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{n_x})$ and $(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{n_y})$ in $\boldsymbol{Z}(t)$ by the corresponding Bootstrap variables $(\boldsymbol{W}_1^*, \ldots, \boldsymbol{W}_{n_x}^*)$ and $(\boldsymbol{V}_1^*, \ldots, \boldsymbol{V}_{n_y}^*)$, respectively. Also, let $\hat{\Gamma}(s,t)$ denote the covariance matrix function of $\boldsymbol{Z}_n^*(\cdot)$, i.e., $\hat{\Gamma}(s,t) = E_* \boldsymbol{Z}_n^*(t) \boldsymbol{Z}_n^*(t)^T, \;\; s,t \in [-t_0, t_0]$, where $E_*$ denotes expectation under $P_*$. Then, using Lemma 1, it is easy to check that on the set $A \cap B$,

$$\sup \Big\{ \|\hat{\Gamma}(s,t) - \Gamma(s,t)\| : s,t \in [-t_0, t_0] \Big\} \to 0 \;\text{ as } n_x \to \infty.$$

As a result, for any $\omega \in A \cap B$, the finite dimensional distributions of the $\boldsymbol{Z}_n^*$-process converges to those of the $\boldsymbol{Z}$-process, and further by Hoeffding's inequality, the tightness condition continues to hold. This implies that on the set $A \cap B$, $\boldsymbol{Z}_n^*$ converges in distribution to the same limiting process $\boldsymbol{Z}$ as in Lemma 1. Further, repeating the arguments in the proof of Theorem 1 and using uniform convergence of $\widehat{\phi}_1(t)$, $\widehat{\phi}_2(t)$, $\widehat{\phi}_{u_x}(t)$ and $\widehat{\phi}_{v_y}(t)$ o their respective limits on the set $A \cap B$, one can show that, for any $\omega \in A \cap B$,

$$T_{n_x}^* \to^d T_\infty.$$

Theorem 3 now follows from Theorem 1, Polya's Theorem, and the continuity of the limiting random variable $T_\infty$. □

# APPENDIX B

## PROOF OF PROPOSITION 1 IN CHAPTER 3

I assume that the true density $f_0$ for $y|\boldsymbol{x}$ is in the Hölder space $\mathcal{C}^\alpha[0,1]$ where a function $g \in \mathcal{C}^\alpha[0,1]$ satisfies $|g^{(m)}(x) - g^{(m)}(y)| \leq L|x-y|^\alpha$ for $\alpha \in (m, m+1]$ and a constant $L$. Let $F_0$ be the corresponding probability measure of $f_0$. With the choice of uniform prior on $\boldsymbol{\theta}_1$ in Section 3.3.1, the corresponding convergence rate is $a_{n,1} = n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$ (Example 4.2 in Ghosal et al., 2008). For the parametric family of models $\mathcal{F}_0$, the convergence rate is $a_{n,0} = n^{-1/2}$ so that $a_{n,1} > a_{n,2}$ for all $\alpha > 0$. We use Corollary 3.1 from Ghosal et al. (2008) to prove Proposition 1 and I state the conditions for Bayes factor to be consistent and check them. Recall that $\mathcal{F}_0 = \{f_{y|\boldsymbol{x}}(\cdot;\boldsymbol{\theta}_0) : \boldsymbol{\theta}_0 \in \Theta_0\}$ and $\mathcal{F}_1 = \{f_{y|\boldsymbol{x}}(\cdot;\boldsymbol{\theta}_1) : \boldsymbol{\theta}_1 \in \Theta_1\}$.

**1. When $f_0 \in \mathcal{F}_0$**

**N1** $\Pi_0(\boldsymbol{\theta}_0 \in \Theta_0 : \int \log(f_0/f_{y|\boldsymbol{x}}(\cdot;\boldsymbol{\theta}_0))dF_0 \leq a_{n,0}^2, \int \log(f_0/f_{y|\boldsymbol{x}}(\cdot;\boldsymbol{\theta}_0))^2 dF_0 \leq a_{n,0}^2) \geq e^{-na_{n,0}^2}$.

**N2** For a sufficiently large constant $M$,

$$\Pi_1(\boldsymbol{\theta}_1 \in \Theta_1 : d(f_{y|\boldsymbol{x}}(\cdot;\boldsymbol{\theta}_1), f_0) \leq M a_{n,1}) \leq o(e^{-3na_{n,0}^2}), \tag{B.1}$$

for some distance functions $d$ on set of densities.

**2. When $f_0 \notin \mathcal{F}_0$**

**A1** $\Pi_1(\boldsymbol{\theta}_1 \in \Theta_1 : \int \log(f_0/f_{y|\boldsymbol{x}}(\cdot;\boldsymbol{\theta}_1))dF_0 \leq a_{n,1}^2, \int \log(f_0/f_{y|\boldsymbol{x}}(\cdot;\boldsymbol{\theta}_1))^2 dF_0 \leq a_{n,1}^2) \geq e^{-na_{n,1}^2}$.

**A2** For every $n$ and some $I_n \to \infty$, $d(f_0, \mathcal{F}_0) \geq I_n a_{n,1}$.

*Proof of Proposition 1.* 1. The condition **N2** is satisfied due to Example 4.2 in Ghosal et al. (2008). Therefore, $B_{01} \to \infty$ when $f_0 \in \mathcal{F}_0$ by Corollary 3.1 (2) in Ghosal et al. (2008).

2. The condition **A1** is the assumption for the prior distribution under which the contraction rate $a_{n,1}$ is attained. In addition, since $f_0$ is not in $\mathcal{F}_0$ we expect $d(f_0, \mathcal{F}_0) > 0$ so that $d(f_0, \mathcal{F}_0) \geq I_n \epsilon_{n,1}$ for any $a_{n,1} \to 0$ and sufficiently slowly increasing $I_n$. Therefore, $B_{01} \to 0$ when $f_0 \notin \mathcal{F}_0$ by Corollary 3.1 (1) in Ghosal et al. (2008). □

# APPENDIX C

# COMPUTATIONAL COMPLEXITY OF THE TWO APPROACHES FOR THE LOGISTIC REGRESSION EXAMPLE IN CHAPTER 4

**Algorithm 1** Full Bootstrap method for the logistic regression model in Section 4.5.1

```
for $b = 1$ to $B$ do
    Draw a bootstrap sample
    Initialize $\alpha_0^{(b)}$ and $\beta_0^{(b)}$
    for $m = 1$ to $M + burn$ do
        Propose $\alpha^{cand}, \beta^{cand} \sim q(\alpha, \beta | \alpha_{m-1}^{(b)}, \beta_{m-1}^{(b)})$
        Calculate
        $r = \min\{1, \frac{\pi(\alpha^{cand}, \beta^{cand} | \boldsymbol{D}^{(b)}) q(\alpha_{m-1}^{(b)}, \beta_{m-1}^{(b)} | \alpha^{cand}, \beta^{cand})}{\pi(\alpha_{m-1}^{(b)}, \beta_{m-1}^{(b)} | \boldsymbol{D}^{(b)}) q(\alpha^{cand}, \beta^{cand} | \alpha_{m-1}^{(b)}, \beta_{m-1}^{(b)})}\}$
        Generate $u \sim U(0,1)$
        if $u < r$ then
            $\alpha_m^{(b)} = \alpha^{cand}$
            $\beta_m^{(b)} = \beta^{cand}$
        else
            $\alpha_m^{(b)} = \alpha_{m-1}^{(b)}$
            $\beta_m^{(b)} = \beta_{m-1}^{(b)}$
        end if
    end for
    Find averages $\widehat{\alpha}^{(b)}, \widehat{\beta}^{(b)}$ for the $b^{\text{th}}$ bootstrap sample:
    $\widehat{\alpha}^{(b)} = \sum_{j=1}^{M} \alpha_j^{(b)} / M$ and $\widehat{\beta}^{(b)} = \sum_{j=1}^{M} \beta_j^{(b)} / M$
end for
Evaluate standard deviations $sd_1(\alpha)$ and $sd_1(\beta)$ as in Section 4.5.1
```

**Algorithm 2** Proposed method for the logistic regression model in Section 4.5.1

```
for $b$ = 1 to $B$ do
    Draw a bootstrap sample, and obtain $(r_1^{(b)}, \ldots, r_n^{(b)})$
    for $m = 1$ to $M$ do
        evaluate $\omega^{(b)}(\alpha_m, \beta_m) = \prod_{i=1}^{n} f^{(r_i^{(b)}-1)}(X_i, Y_i|\alpha_m, \beta_m)$ $^\dagger$
    end for
    Find averages $\widehat{\alpha}^{(b)}$, $\widehat{\beta}^{(b)}$ for the $b^{\text{th}}$ bootstrap sample $^*$:
    $\widehat{\alpha}^{(b)} = \sum_{j=1}^{M} \alpha_j \omega^{(b)}(\alpha_m, \beta_m) / \sum_{j=1}^{M} \omega^{(b)}(\alpha_m, \beta_m)$ and
    $\widehat{\beta}^{(b)} = \sum_{j=1}^{M} \beta_j \omega^{(b)}(\alpha_m, \beta_m) / \sum_{j=1}^{M} \omega^{(b)}(\alpha_m, \beta_m)$
end for
Evaluate standard deviations $sd_2(\alpha)$ and $sd_2(\beta)$ as in Section 4.5.1
```

$^\dagger$ $(\alpha_1, \beta_1), \ldots, (\alpha_M, \beta_M)$ are from $\pi(\alpha, \beta|\boldsymbol{D})$.

$^*$ If we are interested in the $q^{\text{th}}$ quantile we will compute $\widehat{\alpha}_q^{(b)}$, $\widehat{\beta}_q^{(b)}$ based on equation (4.1) in Section 4.4.1 at this step.