NORTHWESTERN UNIVERSITY

Bayesian Analysis in Partially Identified Parametric and Non-Parametric Models

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Statistics

By

Yuan Liao

EVANSTON, ILLINOIS

August 2010

ABSTRACT

Bayesian Analysis in Partially Identified Parametric and Non-Parametric Models

Yuan Liao

This dissertation studies a general type of econometrc model characterized by moment conditions. Such a model, with different variations, has many important empirical applications in economics, biostatistics, and finance. The variations of the model have two dimensions: one is on the type of the moment conditions: either moment equality more inequality; the other is on the dimension of the structural parameter: either finite or infinite. As a result, the model contains most of the important econometric models. The key feature of the model that I am interested in is that the parameter is not completely identified. With limited knowledge of the underlying data distribution, it is only partially identified. I proceed with a Bayesian approach in this dissertation.

Chapter 1 This chapter introduces the model and corresponding Bayesian methods in the literature, followed by detailed examples of the models to be considered in this dissertation. I present in detail some closely related recent literature, from both frequentist and Bayesian perspectives.

Chapter 2 I study a type of moment condition that has been rapidly studied by econometricians in recent years: moment inequalities. Since the parameter of interest is allowed to be not point identified, the treatment is very flexible in dealing with incomplete data, such as missing data or censored data. I construct the posterior distribution of the structural parameter, and establish its large sample behaviors. Since in many applications, it is more straightforward to specify the moment inequalities than the distribution of the data generating process, hence instead of the true likelihood, the posterior density is derived based on the limited information likelihood, a moment condition based likelihood. It is shown that the posterior converges to zero exponentially fast outside any small neighborhood of the identified region. Inside the identified region, it is bounded below by a rate that is not exponentially small. The simulations provide evidence that the Bayesian approach has very attractive properties, in the sense that, with a proper choice of the prior, the posterior provides extra information about the true parameter inside the identified region.

Chapter 3 There exists a moment and model selection problem in the moment inequality model. Here only a subset of the moment inequalities are to be used and the true parameter vector is assumed to follow a submodel allowing only some selected components to be nonzero (which can be, e.g., the regression coefficients of some selected explanatory variables). The moment inequalities are called compatible if fixing the dimension of the parameter vector and the parameter space, the identified region defined by these moment inequalities is not empty. I derive the posterior distribution of the moment inequality/parameter subspace combination, and show that the incompatible combinations have exponentially small posteriors. While the posteriors of compatible combinations are positive, they are sensitive to the researchers' a priori information of the model, which is the choice of the priors.

Chapter 4 This chapter addresses the estimation of the semi-nonparametric conditional moment restricted model that involves a nonparametric structural function g_0. The posterior distribution of the parameter of interest is derived based on the limited information likelihood. I focus on the frequentist properties of the posterior distribution, allowing the nonparametric structural function to be partially identified. It is shown that the posterior converges to any small neighborhood of the identified region. I then apply the results to the single index model and the nonparametric instrumental regression model. In particular, the compactness assumption on the parameter space for nonparametric instrumental regression is relaxed, and a regularized prior is used to overcome the ill-posedness.

Chapter 5 I consider a Bayesian approach to making joint probabilistic inference on the action and the associated risk in data mining. The posterior probability is based on an empirical likelihood, which imposes a moment restriction relating the action to the resulting risk, but does not otherwise require a probability model for the underlying data generating process. The moment restriction partially identifies the parameters of interest, which include both the theoretical risk of interest and the parameters describing the associated actions. I illustrate with examples how this framework can be used to describe the posterior probability of actions to take in order to achieve a low risk, or conversely, to describe the posterior distribution of the resulting risk for a given action. The posterior distribution will cluster around the true risk-action relation with high probability for large data size, and that the actions can be generated from this posterior to reliably control the true resulting risk.

Acknowledgements

I would like to express my sincere appreciation to my advisor Professor Wenxin Jiang, for all the enthusiastic support and guidance he has given me. I have gained tremendously from his mentor not only professionally but also personally. It is his everlasting encouragement that determines my aim of pursuing an academic career goal for life.

I am deeply indebted to Professor Joel Horowitz, who has given me the key to the world of econometrics. It is his generous advice and above all care made all the difference. I am also grateful to Professor Thomas Severini and Professor Elie Tamer, because many things would not have been achieved without their kind help.

My gratitude also goes to Bruce Spencer for all his encouragement, as well as to all the professors in Northwestern statistics: Ajit Tamhane, Larry Hedges, Ji-ping Wang, Hong-mei Jiang, Sandy Zabell, Beth Andrews, Noelle Samia, and Martin Tanner. Many thanks are due to my fellow colleagues in statistics department, especially Zack Almquist, Beth Tipton, Eric Song, Huanhuan Wang, Chris Rhoads, Lili Yao, Xuan Mei and Eduardo Mendes. I also would like to thank Wei Zhao and Hang Zhou, who have made the life in Evanston enjoyable for me.

This dissertation is improved a lot from the discussions with Professor Han Hong and Ivan Canay, and from the seminar participants of my job interview at Chicago Both School of Business, especially Alan Bester.

Finally, I would like to especially thank Xiao Bai and my roommate Christian Kellner, for everything. This dissertation is also dedicated to them.

To My Parents Yasha and Dechang, and to Tibet, Where They Met

Preface

I have dedicated five years of my life, from September 2005 to June 2010, to pursue my Ph.D. degree, at the Department of Statistics, Northwestern University, Evanston, USA. This dissertation is based upon the studies conducted during this period.

This work systematically studies partially identified econometric models from a Bayesian asymptotic perspective. The motivation of the research originated to the Fall 2006, when I took the course ECON 481-1 with Professor Joel Horowitz, on nonparametric methods in econometrics. Ever since then, the door of econometrics has opened to me.

The second and third chapters have been combined and published on the Annals of Statistics in January 2010. Just upon this dissertation is written up, the fourth chapter is re-organized and submitted to the Annals of Statistics as well.

Charles Manski wrote in the preface of his book *Partial Identification of Probability Distributions* (Springer 2003):

> “He (Arthur Goldberger) especially lifted my spirits when, reacting to a new finding that I had excited shown him, he remarked ‘now you are flying’.”

I believe this dissertation is the starting point for me to fly eventually.

Yuan Liao

Evanston, IL, USA

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Many econometric and statistical models involve an unknown structural parameter of interest, which is either parametric or semi/non-parametric. Such a parameter often satisfies some moment conditions that are implied by the assumptions imposed on the distribution of the data generating process. Therefore, the variations of these models are in terms of two directions: one is the dimension of the structural parameter: either finite or infinite; the other is the type of the moment conditions: either moment equality or moment inequality. The moment equality is one of the most commonly seen moment conditions, defined as

$$Em(X,\theta)=0 \tag{1.1}$$

where the expectation is taken with respect to the observable random vector X, and $m(X,.)$ is a known function of the parameter of interest θ. Many statistical and econometric methods have been developed to estimate and make inference about θ from the frequentist perspective, such as estimating equations and generalized method of moments. At almost the same time, Bayesian statistical methods were also widely developed and applied to such a model. When the structural parameter is identifiable, a standard Bayesian procedure can be stated as following: first, specify the prior distribution of the parameter either from previous study experiences or using objective priors such as the Jeffery's prior if the Fisher's information matrix is nonsingular. Second, construct the likelihood function based on the distribution of the data generating process. Third, find the posterior distribution of the parameter by multiplying the prior with the likelihood function. Finally, obtain MCMC draws from the posterior by carrying out the Metropolis algorithm. From the frequentist point of view, under regularity conditions, the Bayesian procedure is consistent, meaning that asymptotically, the posterior distribution degenerates to a Dirac measure on the true value. As a result, the posterior mean estimated by averaging the MCMC draws should consistently estimate the unknown parameter.

In recent years, the statistical models where the structural parameters are not identifiable have been brought into attention by both statisticians and econometricians. The problem with loss of identifiability occurs in diverse areas such as industrial organization, reliability theory and survival analysis (Prakasa Rao 1992 and Tamer 2009). One of the most important types of econometric models with loss of identifiability is the "moment inequality", as given by

$$Em(X,\theta) \geq 0 \tag{1.2}$$

The structural parameter θ may be unidentifiable in this model since there can be more than one θ satisfying model(1.2), regardless of the dimension of m.

One of the biggest problems of the loss of identifiability is that the regular estimators, developed in mathematical statistics and econometrics, are not consistent anymore. The reason is that, generally speaking, in classical statistical models with identifiability, the structural parameter is usually identified as the minimizer of some nonnegative function over the parameter space. That is to say, let θ_0 denote the true parameter of interest, which lies in some parameter space Θ. $\theta_0 = \arg\min_{\theta\in\Theta} Q(\theta)$, where $Q(\theta)$ is a nonnegative function on Θ, and depends on the distribution of the data generating process. Let $Q_n(\theta)$ denote the sample analog of $Q(\theta)$, under some regularity conditions, the estimator defined by $\hat{\theta}_n = \arg\min_{\theta\in\Theta} Q_n(\theta)$ then converges in probability to θ_0. On the other hand, the loss of identifiability often occurs when the set of minimizers of $Q(\theta)$ is not a singleton, say $\arg\min_{\theta\in\Theta} Q(\theta) \neq \{\theta_0\}$, hence the minimizer of $Q_n(\theta)$ is usually not a singleton either, and even if it were, it would not converge to the true parameter of interest. Therefore, the consistency of point estimation cannot be achieved. From the Bayesian perspective, the classical posterior consistency is established based on the fact that, up to the leading order, the likelihood function is proportional to $\exp(-nQ_n(\theta))$, where $Q_n(\theta)$ converges to some unknown function $Q(\theta)$ uniformly over θ, and $Q(\theta)$ is uniquely minimized at θ_0. In addition, the prior distribution has support on a neighborhood of θ_0 (Chernozhukov and Hong 2003). However, when θ_0 is not identifiable, the posterior distribution will not only asymptotically supported on a neighborhood of θ_0, but also on all the minimizers of $Q(\theta)$ (Liao and Jiang 2010).

A third type of important econometric model is the "conditional moment restricted model", as defined by

$$E[\rho(Z,\theta)|W] = 0 \tag{1.3}$$

where the conditional expectation is taken with respect to the conditional distribution of Z given W, with both Z and W observable. Here $\rho(Z,.)$ is a known function of the structural parameter θ, usually called the "residual function". Model (1.3) is usually satisfied by the regression models, with the dimension of θ being either finite or infinite. For example, when θ is finite dimensional and $\rho(Z,\theta)$ linearly depends on θ, model (1.3) is then either the simple linear regression model or the linear regression model with instrumental variable W. When θ is infinite dimensional, it is the semi-parametric or nonparametric regression model. Depending on different cases, the sufficient conditions for the identifiability of θ can be different.

In organizing this introductory chapter, I will start by illustrating some empirical examples in economics, survival analysis, and treatment effects, where the structural parameter is not identifiable (or point identified) by the statistical model assumptions. This section will then be followed by detailed reviews of the theories and methodologies developed in dealing with the problems of loss of identifiability in the literature, from both the frequentist and Bayesian perspectives. The literature review comes from both the statistics and econometrics literatures, where in the latter case, the term "loss of identifiability" is usually referred to "partial identification". Followed the partial identification, I will then introduce the literature of the conditional moment restricted model (1.3), with both the point identification and partial identification cases.

1.1. Examples of Loss of Identifiability

This section illustrates some important examples of loss of identifiability in statistics and econometrics. These examples are: the empirical English auctions, the treatment selections, the censored data in survival analysis, and the binary choice model.

Example 1.1.1 (Empirical English Auctions)**.** Haile and Tamer (2003) carried out an empirical analysis of the English auctions that relies on an incomplete model consisting of two simple assumptions.

Assumption 1: Bidders do not bid more than they are willing to pay.

Assumption 2: Bidders do not allow an opponent to win at a price they are willing to beat.

At the auction, an initial price of the object is designated, and monotonically increasing bids are then accepted from the participating bidders, subject to a minimum bid increment $\Delta \geq 0$. For each bidder i, let V_i denote her valuation on the object, i.e., the maximal value that she is willing to pay. The valuation V_i is not observable to econometricians. Often one is directly interested in how valuations are affected by auction-specific observables such as the bidder's demographic characteristics, the terms of a government contract, or the seller's reputation. Let X_i denote a vector of the observales for bidder i. The parameter of interest θ is defined as an element that lies in some parameter space Θ, satisfying

$$E(V_i|X = x_i) = l(x_i, \theta)$$

for all i, where $l(.,.)$ is known and may take any form; it may be either linear or a polynomial, for example. In addition, instead of V_i, econometricians can observe Y_i, bidder i's final bid,$Y_{\max}$, the maximal bid of all the bidding participants for the object, and Δ, the minimum increment. By Assumptions 1 and 2, $Y_i \leq V_i \leq Y_{\max} + \Delta$. It then follows that

$$E(Y_i|X_i = x_i) \leq l(x_i, \theta) \leq E(Y_{\max} + \Delta|X_i = x_i). \tag{1.4}$$

The parameter of interestθ may not be identifiable if the set of θ that satisfies inequalities (1.4) is not a singleton, i.e., that are more than one elements in Θ that are consistent with data. If it is true, we cannot consistently estimate the true parameter θ even if we had infinitely many $(Y_i, X_i, Y_{\max})$. Therefore it results to the loss of identifiability.

Example 1.1.2 (Missing Data Problem)**.** Manski (2003) discussed and worked out a type of problem in which the loss of identifiability of the parameter of interest arises from the missing data. Let us start with the simplest case where only the outcome is missing, and then extend to the case where both the covariate and outcome are missing, previously studied by Horowitz and Manski (2000).

Let Y be an outcome dummy random variable that indicates whether a treatment is successful $(Y = 1)$ or unsuccessful $(Y = 0)$. Z is an indicator of missing data. Y is observed if $Z = 1$ and unobserved if $Z = 0$. The parameter of interest here is $P(Y = 1)$, the probability that the treatment is successful. Note that

$$P(Y = 1) = P(Y = 1|Z = 1)P(Z = 1) + P(Y = 1|Z = 0)P(Z = 0)$$

The data generating process identifies $P(Y = 1|Z = 1)$, $P(Z = 1)$ and $P(Z = 0)$, but not $P(Y = 1|Z = 0)$, therefore $\theta = P(Y = 1)$ is not identifiable or point identified. Note that $P(Y = 1|Z = 0)$ can take any possible value in $[0, 1]$, hence we can derive the lower and upper bounds for θ, which would be

$$P(Y = 1|Z = 1)P(Z = 1) \leq \theta \leq P(Y = 1|Z = 1)P(Z = 1) + P(Z = 0) \tag{1.5}$$

Horowitz and Manski (2000) considered nonparametric missing data problem with both outcome and covariate missing. Let Y be a binary outcome variable same as before, and X be a covariate. Let Z_y and Z_x denote the indicators of missing data. Y is observed if $Z_y = 1$ and unobserved if $Z_y = 0$. X is observed if $Z_x = 1$ and unobserved if $Z_x = 0$. The covariate X is assumed to be "missing completely at random", meaning that

$$P(Z_x = i|Y = j, X = x, Z_y = k) = P(Z_x = i)$$

for all x in th support of X and all $i, j, k \in \{0, 1\}$. The parameter of interest is $g(x) = P(Y = 1|X = x)$, which is assumed to be nonparametric. Under the Missing Completely at Random assumption, it can be shown that

$$\begin{aligned} g(x) &= P(Y = 1|X = x, Z_x = 1, Z_y = 1)P(Z_y = 1|X = x, Z_x = 1) \\ &\quad +P(Y = 1|X = x, Z_x = 1, Z_y = 0)P(Z_y = 0|X = x, Z_x = 1) \end{aligned} \tag{1.6}$$

All quantities on the right side of (1.6) are identifiable by the data generating process except $P(Y = 1|X = x, Z_x = 1, Z_y = 0)$, which can take any value in $[0, 1]$. Therefore $g(x)$ is not point identified. The upper and lower bounds similar to (1.5) can be derived. Let $h(x) = P(Z_y = 1|X = x, Z_x = 1)$, then

$$\begin{aligned} g(x) &\geq P(Y = 1|X = x, Z_x = 1, Z_y = 1)h(x) \\ g(x) &\leq P(Y = 1|X = x, Z_x = 1, Z_y = 0)h(x) + 1 - h(x) \end{aligned} \tag{1.7}$$

Example 1.1.3 (Censored Data)**.** Consider the parameter θ in the model

$$y_i = x_i'\theta + \epsilon_i$$

where x_i is a vector of covariates, θ is the unknown parameter of interest, and ϵ denotes the unobserved error term with Median$(\epsilon_i|x_i) = 0$. In survival analysis, y_i is usually censored. In particular, we observe the random vector (x_i, v_i, d_i) such that

$$v_i = \min\{y_i, c_i\}$$

$$d_i = I_{(y_i < c_i)}$$

where d_i is a binary variable that indicates whether an observation v_i is censored or not. The random variable c_i is only observed for censored observations. If the censoring value c_i is assumed to be independent of (x_i, ϵ_i), then under non-singularity conditions for some matrices, θ is point identified (Honore, Khan and Powell 2002). In fact, let $F_i(y_i;\theta|x_i)$ be the conditional distribution function of y_i, with density function $f_i(y_i;\theta|x_i)$. The likelihood function can be written as

$$L(\theta) = \prod_{d_i=1} f_i(y_i, \theta|x_i) \prod_{d_i=0} [1 - F_i(c_i, \theta|x_i)]$$

The structural parameter θ can then be consistently estimated by parametric methods (maximum likelihood), or nonparametric methods (first estimate the survival function using the Kaplan-Meier estimator). However, in many applications, the independence assumption is suspected. For example, suppose y_i and c_i are the survival time and censoring time for a patient respectively. When patients whose conditions worsen significantly are less likely to continue in the study, the result would be a positive correlation between censoring time and the survival time. In other situations, the censoring can also be affected by unobservables that also affect outcomes. Without assuming the independence between the censoring and surviving time, the point identification often does not hold. In the literature of survival analysis, there is difficulty of estimating survival functions when survival and censoring are not independent, see Tsiatis (1975).

Under the assumption Median$(\epsilon_i|x_i) = 0$ alone, Khan and Tamer (2009) showed that θ satisfies *conditional moment inequalities*:

$$\begin{aligned} E[I_{(v_i \geq x_i'\theta)}|x_i] &\leq \frac{1}{2} \\ E[d_i I_{(v_i \leq x_i'\theta)}|x_i] &\leq \frac{1}{2} \end{aligned} \tag{1.8}$$

The proof is straightforward:

$$\begin{aligned} E[I_{(v_i \geq x_i'\theta)}|x_i] &= P(v_i \geq x_i'\theta|x_i) = P(c_i \geq x_i'\theta, y_i \geq x_i'\theta|x_i) \\ &= P(c_i \geq x_i'\theta, \epsilon_i \geq 0|x_i) \\ &\leq P(\epsilon_i \geq 0|x_i) = \frac{1}{2} \end{aligned}$$

$$\begin{aligned} E[d_i I_{(v_i \leq x_i'\theta)}|x_i] &= E[I_{(v_i \leq x_i'\theta, y_i \leq c_i)}|x_i] = P(v_i \leq x_i'\theta, y_i \leq c_i|x_i) \\ &= P(y_i \leq c_i, y_i \leq x_i'\theta|x_i) = P(\epsilon_i \leq c_i - x_i'\theta, \epsilon_i \leq 0|x_i) \\ &\leq P(\epsilon_i \leq 0|x_i) = \frac{1}{2} \end{aligned}$$

Without further assumptions on the distribution of (x_i, ϵ_i, c_i), conditional moment inequalities (1.8) generally do not guarantee the point identification of θ.

Example 1.1.4 (Probit Binary Choice Model)**.** Consider the binary choice model

$$y_i = I_{(x_i'\theta + \epsilon_i < 0)}$$

The error term ϵ is assumed to be $N(0, \sigma^2)$ and independent of x_i, where σ^2 is the unknown variance. We observe $(x_i, y_i)_{i=1}^n$. Note that

$$P(y_i = 1|x_i) = P(x_i'\theta + \epsilon_i < 0|x_i) = \Phi\left(-\frac{x_i'\theta}{\sigma}\right)$$

where $\Phi(.)$ denotes the cumulative distribution function of standard normal distribution. Here $\frac{\theta}{\sigma}$ is point identified, but not θ (Imai and van Dyk 2004, and McCulloch et al. 2000). In fact, for any $c \neq 0$, let $\tilde{\theta} = c\theta$, and $\tilde{\sigma} = c\sigma$, then

$$\Phi\left(-\frac{x_i'\tilde{\theta}}{\tilde{\sigma}}\right) = P(y_i = 1|x_i)$$

1.2. Literature Review on Partially Identified Models

The inference on the structural parameter that is only partially identified has been largely initiated and popularized in recent years. The recent literature on partially identified models in econometrics starts from the study of interval identified models. Horowitz and Manski (2000) consistently estimated the bounds of the conditional probability of a successful treatment which is partially identified on an interval, with missing data. They also derived confidence intervals that asymptotically cover the entire identification region with fixed probability. Manski and Tamer (2002) considered the linear models with interval data, where the structural parameter is partially identified on a set of minimizers of a particular objective function $Q(\theta)$, i.e. the true parameter $\theta_0 \in \arg\min_{\theta\in\Theta} Q(\theta)$, where Θ is the parameter space. Different from the classical M-estimator, $\arg\min_{\theta\in\Theta} Q(\theta)$ is no longer a singleton. Manski and Tamer (2002) constructed a set of parameters A_n that consistently estimates the identified region $\arg\min_{\theta\in\Theta} Q(\theta)$ in Hausdorff distance. Define

$$d_H(A, B) = \max\{\sup_{a\in A} d(a, B), \sup_{b\in B} d(b, A)\} \tag{1.9}$$

where $d(b, A) = \inf_{a\in A} ||b - a||$. Let $\Omega = \arg\min_{\theta\in\Theta} Q(\theta)$, they showed that $d_H(A_n, \Omega) \rightarrow^p 0$ in probability. In addition, Imbens and Manski (2004) considered inference on the partially identified structural parameter. They constructed a uniform confidence interval C_n for θ_0 with coverage probability α, say,

$$\lim_{n\to\infty}\inf_{\theta\in\Omega} P(\theta \in C_n) \geq \alpha$$

A more general form of partially identified models is the moment inequality model

$$Em(X,\theta) \geq 0$$

where $m(.,.) : Supp(X) \times \Theta \to \mathbb{R}^d$ is a known function of observable random variable X and θ, where $supp(X)$ denotes the support of X.

Example 1.2.1 (Empirical English Auction Continued)**.** Consider example 1.1.1. The parameter of interest satisfies (1.4). Then for any positive function $h(x)$, such that $E|h(X)| < \infty$,

$$E[h(X_i)Y_i] \leq E[h(X_i)l(X_i,\theta)] \leq E[h(X)(Y_{\max} + \Delta)]$$

Therefore we have moment inequalities, with

$$m(X,Y,\theta) = \begin{pmatrix} h(X)l(X,\theta) - h(X)Y \\ h(X)(Y_{\max} + \Delta) - h(X)l(X,\theta) \end{pmatrix}$$

Example 1.2.2 (Missing Data Problem Continued)**.** Consider example 1.1.2, where only the outcome is missing. The parameter of interest satisfies (1.5). Write $P(Y = 1|Z = 1)P(Z = 1) = E(I_{(Y=1,Z=1)})$, and $P(Z = 0) = E(I_{(Z=0)})$, then we have moment inequality model with

$$m(Z,Y,\theta) = \begin{pmatrix} \theta - I_{(Y=1,Z=1)} \\ I_{(Y=1,Z=1)} + I_{(Z=0)} - \theta \end{pmatrix}$$

Example 1.2.3 (Censored Data Continued)**.** Consider example 1.1.3. The parameter of interest satisfies 1.8. Then for any positive function $h(x)$ such that $E|h(X)| < \infty$, we have

$$E[h(X)I_{(V\geq X'\theta)}] \leq \frac{1}{2}E[h(X)]$$
$$E[h(X)DI_{(V\leq X'\theta)}] \leq \frac{1}{2}E[h(X)]$$

Then the moment inequalities hold with $m(.,.)$ given by:

$$m(X,V,\theta) = \begin{pmatrix} \frac{1}{2}h(X) - h(X)I_{(V\geq X'\theta)} \\ \frac{1}{2}h(X) - h(X)DI_{(V\leq X'\theta)} \end{pmatrix}$$

1.2.1. Consistent Set Estimation and Confidence Regions

Suppose the true parameter of interest θ_0 belongs to some parameter space Θ. The set of parameter values that satisfy the moment inequality models (1.2) is given by

$$\Omega = \{\theta \in \Theta : Em(X,\theta) \geq 0\}.$$

The true parameter θ_0 is not point identified because normally Ω is not a singleton: there are more than one elements in Θ that satisfy the moment inequalities. The set Ω that captures all the information about θ_0 is called the *identified region* for θ_0. Under the general moment inequality setting, the identified region is neither necessarily an interval nor the Cartesian products of intervals.

Chernozhukov, Hong and Tamer (2007) are among the first who construct the consistent set estimator in Haursdorff distance of Ω and its confidence region, given the general moment inequality assumptions. Their inference was based on the observation that the identified region is the set of minimizers of a criterion function $Q(\theta)$. Let

$$m(X,\theta) = \begin{pmatrix} m_1(X,\theta) \\ \vdots \\ m_p(X,\theta) \end{pmatrix}$$

and $w_1(\theta), ..., w_p(\theta)$ be strictly positive functions on Θ. Define

$$Q(\theta) = \sum_{i=1}^{p} w_i(\theta)[Em_i(X,\theta)]^2 I_{(Em_i(X,\theta)<0)} \tag{1.10}$$

Function $Q(\theta)$ is minimized to zero if and only if $\theta \in \Omega$. In other words, $\Omega = \arg\min_{\theta\in\Theta} Q(\theta)$. Therefore, the inference on Ω may be based on the empirical analog of Q: define

$$Q_n(\theta) = \sum_{i=1}^{p} w_{ni}(\theta)\bar{m}_i(\theta)^2 I_{(\bar{m}_i(\theta))<0}$$

where $\bar{m}_i(\theta) = \frac{1}{n}\sum_{j=1}^{n} m_i(X_j,\theta)$, and $w_{ni}(\theta)$ is an estimate that converges to $w_i(\theta)$ uniformly in $\theta \in \Theta$. For a sequence $a_n \to \infty$, define $C_n(c) = \{\theta \in \Theta : a_n Q_n(\theta) \leq c\}$. Chernozhukov, Hong and Tamer (2007) showed that if $c_n \geq \sup_{\theta\in\Omega} a_n Q_n(\theta)$ with probability approaching 1, and $c_n/a_n \to^p 0$, then $d_H(C_n(c_n), \Omega) = o_p(1)$. They further derived the rate of convergence, which is close to $1/\sqrt{n}$.

For the confidence region of Ω, note that $\Theta \subset C_n(c)$ is equivalent to $\sup_{\theta\in\Omega} a_n Q_n(\theta) \leq c$. Hence if c_α is such that

$$P\left(\sup_{\theta\in\Omega} a_n Q_n(\theta) \leq c_\alpha\right) = 1-\alpha$$

then $P(\Theta \subset C_n(c_\alpha)) = 1-\alpha$. Chernozhukov Hong and Tamer (2007) constructed the confidence region based on the approximation to the quantiles of $\sup_{\theta\in\Omega} a_n Q_n(\theta)$.

In addition, Romano and Shaikh (2008) provided a similar subsampling procedure for the confidence region. Bugni (2010) proposed a Bootstrap procedure to approximate the quantiles of $\sup_{\theta\in\Omega} a_n Q_n(\theta)$, and showed that the convergence rate of the Bootstrap procedure is faster than subsampling. Other related papers in the literature regarding the confidence regions for moment inequality models can be found in Rosen (2008), Pakes, Porter, Ishiii and Ho (2006), Andrews and Jia (2008), among others.

1.2.2. Optimal Inference

There currently exist a variety of inferential methods for the inference of partially identified parameters in models with moment inequalities. As illustrated earlier, the moment inequality models are usually represented via an objective criterion function $Q(\theta)$ so that $\Omega = \arg\min_{\theta\in\Theta} Q(\theta)$. Function (1.10) is one of the examples applied by Chernozhukov, Hong and Tamer (2007). Rosen (2006) provided an alternative formulation for the criterion function, which is defined by

$$Q(\theta) = \min_{\lambda\geq 0}[Em(X,\theta) - \lambda]' V(\theta)^{-1}[Em(X,\theta) - \lambda]$$

where $V(\theta)$ is the variance of $m(X,\theta)$. In fact, any function $Q(\theta)$ such that:

(1) $Q(\theta) \geq 0$ for all $\theta \in \Theta$, and

(2) $Q(\theta) = 0$ if and only if $\theta \in \Omega$

can be adopted as a criterion function. See Andrews and Soares (2010) for a more formal definition of criterion functions.

Since there are a variety of criterion functions $Q(\theta)$ that have Ω as the set of minimizers, with each criterion function resulting to a different test statistic, and so a different confidence region, a natural question arises immediately: is there a criterion function that is better than the others? Canay (2010) defined the optimality criterion in terms of the asymptotic power of the test statistic for testing

$$H_0 : Em(X, \theta) \geq 0.$$

He showed that the criterion function constructed based on the empirical likelihood function (Owen 1990) achieves the optimal asymptotic power than others under certain rate restriction. Suppose we observe i.i.d. data $X^n = (X_1, ..., X_n)$ from the population X, the empirical log-likelihood for θ of moment inequality models is given by

$$\begin{aligned} l_{el}(\theta) &= \max_{p_1,\ldots,p_n} \sum_{i=1}^{n} \log p_i \\ \text{subject to:} \quad & p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i m(X_i, \theta) \geq 0 \end{aligned} \tag{1.11}$$

where p_i denotes the probability mass placed at X_i by a discrete distribution with support $\{X^n\}$. The unrestricted empirical log-likelihood $l_u(\theta)$ is defined similarly except that the moment restriction $\sum_{i=1}^{n} p_i m(X_i, \theta) \geq 0$ is not imposed. Based upon both restricted and unrestricted empirical log-likelihood $l_{el}(\theta)$ and $l_u(\theta)$, the empirical likelihood ratio statistic is given by

$$ELR_n(\theta) = 2[l_{el}(\theta) - l_u(\theta)] = \max_{\lambda \leq 0} 2 \sum_{i=1}^{n} \log(1 + \lambda' m(X_i, \theta))$$

For each given θ, large values of $ELR_n(\theta)$ suggest that the moment inequalities $Em(X, \theta) \geq 0$ are not supported by the data.

Canay (2010) showed that, for any $\theta \in \Theta$, and any $c > 0$, under H_0,

$$P(ELR_n(\theta) \geq 2cn) \leq e^{-cn}.$$

In addition, Any test for H_0 can be defined as $r(\theta) = 1_{(X^n \in R(\theta))}$, where $R(\theta)$ is the rejection region of the sample space, i.e., test $r(\theta)$ rejects H_0 if and only if $X^n \in R(\theta)$. For some $\delta > 0$, denote the neighborhood of $R(\theta)$ by $R^\delta(\theta) = \cup_{x \in R(\theta)} B(x, \delta)$. It is then showed that if a test $r(\theta)$ is such that for any $c > 0$, under H_0, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} P(r(\theta) = 1) \le \limsup_{n\to\infty} P(X^n \in R^\delta(\theta)) \le e^{-cn} \tag{1.12}$$

then

$$\liminf_{n\to\infty} \frac{P(r(\theta) = 0)}{P(ELR_n(\theta) < 2cn)} \ge 1$$

if H_0 is false. This result says that, among all the tests for $H_0 : Em(X, \theta) \ge 0$ whose type I error rate is restricted by (1.12), the empirical log-likelihood ratio test has the best asymptotic power.

1.2.3. Bayesian approach to the partially identified models

The Bayesian methods have been extensively applied to models with loss of identifiability. Gelfand and Sahu (1999) have studied issues surrounding non-identifiability and improper priors in the context of generalized linear models. Neath and Samaniego (1997) considered Bayesian updating for a non-identified two parameter binomial model. Gustafson (2005) studied Bayesian inference in non-identified scenarios involving misclassification and measurement errors, which was discussed by a number of prominent researchers.

With a Bayesian approach McCulloch, Polson and Rossi (2000) have studied the multinomial probit model (MNP) with the case of non-identifiability. The MNP is a multivariate case of the binary choice model, defined as:

$$W = X\theta + \epsilon$$

$$Y = \begin{cases} 0, & \max(W) < 0 \\ i, & \max(W) = W_i > 0 \end{cases} \tag{1.13}$$

where ϵ is $N(0, \Sigma)$. X is a $p \times k$ matrix, and $\max(W)$ means the maximal element of $W' = (W_1, ..., W_p)'$. In practice, econometricians observe a set of observations of (Y, X). In the setting of (1.13), the parameters (θ, Σ) are not point identified. One of the commonly adopted frequentist methods to achieve the point identification is to set the first diagonal element $\sigma_{11}$of Σ equal to 1. However, it is not straightforward to adopt this approach in Bayesian analysis because of the difficulty in imposing a prior on the parameter space such that σ_{11} is one. McCulloch, Polson and Rossi (2000) proposed two ways of prior specification: one imposes priors on θ and Σ directly, and the other imposed priors on θ and $\Sigma|\sigma_{11} = 1$.

A more related work to this dissertation is by Moon and Schorfheide (2009a), who were the first to study the posterior properties of partially identified models, and compare the Bayesian approach to the frequentist's. In Moon and Schorfheide (2009a), the model can involve three types of parameters: the structural parameters of interest θ, a reduced form parameter vector that is point identified by data ϕ, and also a vector of auxiliary parameters α, which links the structural and reduced form parameters via a known function $\theta = \theta(\phi, \alpha)$. For a particular value of ϕ, the auxiliary parameter takes its value in some set A_ϕ. The identified set for θ can be written as

$$\Theta(\phi) = \{\theta(\phi, \alpha) : \alpha \in A_\phi\}$$

Assuming the log-likelihood function $l(\phi)$ for ϕ is known, we can consistently estimate ϕ by maximum likelihood estimator $\hat{\phi}$. It is then straightforward to obtain a consistent set estimator of $\Theta(\phi)$, which is $\Theta(\hat{\phi})$.

For any set A, let $p(\theta \in A|Data)$ denote the posterior probability that $\theta \in A$. Moon and Schorfheide (2009a) showed that for any $\tau \in (0,1)$, there exists a set $T_n(\tau) \subset \Theta(\hat{\phi})$ such that in probability,

$$P(\theta \in T_n(\tau)|Data) \to^p 1 - \tau$$

Hence $T_n(\tau)$ is the asymptotic $1-\tau$ credible set for θ. In addition, let $G(\theta,\alpha) = \phi$ be the link function deduced by $\theta = \theta(\phi,\alpha)$. Let $\hat{\alpha}(\theta) = \arg\max_\alpha l_n(G(\theta,\alpha))$. Define $Q_n(\theta) = 2[l(G(\theta,\hat{\alpha}(\theta)) - l(\hat{\phi})]$. Consider a confidence region for θ that is of the form $CS(c) = \{\theta : Q_n(\theta) \geq -c\}$ for some positive constant c. It is a confidence set of level $1-\tau$ that is uniformly valid asymptotically if

$$\lim_{n\to\infty} \inf_{\theta\in\Theta} P(\theta \in CS(c_\tau)) \geq 1 - \tau$$

Since $Q_n(\theta) = 0$ if $\theta \in \Theta(\hat{\phi})$, it follows immediately that $\Theta(\hat{\phi}) \subset CS(c)$ for any $c > 0$. Therefore with probability approaching one,

$$T_n(\tau) \subset \Theta(\hat{\phi}) \subset CS(c_\tau) \tag{1.14}$$

Result 1.14 indicates that the Bayesian credible set for θ locates inside the identified region, whereas the frequentist confidence region extends beyond the identified region. Hence the Bayesian credible set is asymptotically smaller. This result is different from the point identified case.

Example 1.2.4 (Interval censored data)**.** Consider the simple location model $Y_i = \phi + \epsilon_i$, where $\phi = EY_i$. Suppose the structural parameter $\theta = EX_i$ is the expectation of an unobservable X_i, which is censored almost surely in the interval $[Y_i, Y_i + \lambda]$ for a known constant $\lambda > 0$. Therefore $\phi \leq \theta \leq \phi + \lambda$. Since econometricians can observe Y_i, hence ϕ is point identified, and is linked with θ by an auxiliary parameter α through

$$\phi = \theta - \alpha, \alpha \in [0, \lambda]$$

The identified region for θ is then $\Theta(\phi) = \{\phi + \alpha : \alpha \in [0, \lambda]\}$. Suppose the prior $p(\theta|\phi)$ is uniform on $[\phi, \phi + \lambda]$, and $p(\phi)$ is flat. In addition, suppose $p(Y|\theta, \phi) = p(Y|\phi) \sim N(\phi, \sigma^2)$ for some known σ^2, then under some week regularity conditions, we can show that

$$\begin{aligned} p(\theta|Data) &\propto \int p(Data|\phi)p(\theta|\phi)p(\phi)d\phi \\ &\propto \int_{\theta-\lambda}^{\theta} \exp\left(-\frac{\sum_{i=1}^{n}(Y_i - \phi)^2}{2\sigma^2}\right) d\phi \\ &\propto P(\theta - \lambda \leq Z \leq \theta) \\ &\sim Uniform[\bar{Y}, \bar{Y} + \lambda] \end{aligned}$$

where Z in the third line is $N(\bar{Y}, \sigma^2/n)$. Therefore, the interval $T_n(\tau) = [\bar{Y} + \tau\lambda/2, \bar{Y} + \lambda - \tau\lambda/2]$ is an asymptotic $1 - \tau$ credible interval for θ. Apparently $T_n(\tau) \subset \Theta(\bar{Y})$. One can additionally verify that

$$Q_n(\theta) = \begin{cases} 0 & \bar{Y} < \theta < \bar{Y} + \lambda \\ -\frac{n}{\sigma^2}(\bar{Y} - \theta)^2, & \theta \leq \bar{Y} \\ -\frac{n}{\sigma^2}(\bar{Y} - \theta + \lambda)^2, & \theta \geq \bar{Y} + \lambda \end{cases}$$

Therefore the confidence set CS that satisfies $\lim_{n\to\infty}\inf_{\phi}\inf_{\theta\in\Theta(\phi)} P(\theta \in CS|\phi) = 1-\alpha$ is given by $CS = [\bar{Y} - z_{1-\tau/2}\sigma/\sqrt{n}, \bar{Y} + \lambda + z_{1-\tau/2}\sigma/\sqrt{n}]$, where $z_{1-\tau/2}$ is the $1-\tau/2$ quantile of the standard normal distribution. We can see $T_n(\tau) \subset \Theta(\bar{Y}) \subset CS$.

Recently, Bollinger and Hasselt (2009) applied Moon and Schorfheide (2009a)'s setting to binary misclassification problem. Let $Z \in \{0,1\}$ be a binary random variable with $P(Z = 1) = \pi$, where π is the parameter of interest. Instead of observing Z, we observe $X \in \{0,1\}$, with measurement error. Suppose the misclassification probability is $p = P(X \neq Z|Z)$. On can show that $\mu = E(X) = \pi(1-2p)+p$. Here μ is point identified, p is the reduced parameter, and π is the parameter of interest, which is partially identified as we do not directly observe Z.

Suppose the data $X^n = \{X_1, ..., X_n\}$ are i.i.d., let $n_1 = \sum_i X_i$. Because the data are from Bernoulli distribution, the likelihood function is given by

$$f(X^n|\mu) = \mu^{n_1}(1-\mu)^{n-n_1}$$

One can start with putting a joint prior $f_1(\pi, p)$, and obtain the joint prior $f_1(\pi, \mu)$. Finally, the posterior of π is obtained:

$$p(\pi|X^n) \propto \int f_1(\pi,\mu) f(X^n|\mu) d\mu.$$

1.3. Literature Review on Conditional Moment Restricted Models

The conditional moment restricted model is given by

$$E(\rho(Z, g_0)|W) = 0 \tag{1.15}$$

where (Z^T, W^T) is a vector of observable random variables, and W might or might not be included in Z. Here ρ is a residual function known up to g_0. The conditional expectation is taken with respect to the conditional distribution of Z given W, assumed unknown. The parameter of interest is g_0, which is infinite dimensional. Model (1.15) is a very general setting, which encompasses many important classes of nonparametric and semiparametric models.

Example 1.3.1 (Regular nonparametric regression)**.** Consider model

$$Y = g_0(W) + \epsilon$$

assuming $E(\epsilon|W) = 0$. Let $Z = (Y, W)$, then it can be written as the conditional moment restricted model with $\rho(Z, g_0) = Y - g_0(W)$.

Example 1.3.2 (Nonparametric IV regression)**.** Consider nonparametric model

$$Y = g_0(X) + \epsilon$$

where X is an endogenous regressor, meaning that $E(\epsilon|X)$ does not vanish. However, suppose we have observed an instrumental variable W for which $E(\epsilon|W) = 0$, then it becomes a nonparametric regression model with instrumental variables, studied by Newey and Powell (2003) and Hall and Horowitz (2005). Define $\rho(Z, g_0) = Y - g_0(X)$, with $Z = (Y, X)$. Then we have conditional moment restriction

$$E(\rho(Z, g_0)|W) = 0$$

Example 1.3.3 (Single index model)**.** Consider single index model

$$Y = h_0(W^T\theta_0) + \epsilon$$

where $E(\epsilon|W) = 0$. The parameter of interest is (h_0, θ_0), with h_0 being nonparametric. This type of model is studied by Ichimura (1993) and Antoniadis et al (2004). By defining $Z = (Y, W)$ and $g_0 = (h_0, \theta_0)$, we can write $E(\rho(Z, g_0)|W) = 0$.

Note that g_0 is infinite dimensional, hence model (1.15) is either nonparametric or semiparametric. Therefore it is different from the conditional restriction considered by Kitamura (2005) and Smith (20007), who assume g_0 lies in a finite dimensional compact parameter space. In this section, I will review the work done by Ai and Chen (2003), where the model was studied in a general setting, and the parameter space was assumed to be compact. One of the important applications of the model is the nonparametric instrumental regression. In this particular model, the estimation is very difficult when the compactness assumption on the parameter space is relaxed, and there is a huge literature on solving the so-called "ill-posed" problem. I will review the literature of nonparametric IV regression in a separate section.

Let Θ be a compact parameter space that contains g_0. Define $m(w, g) = E[\rho(Z, g)|W = w]$, then under the assumption that g_0 is point identified by model (1.15),

$$g_0 = \arg \inf_{g \in \Theta} E[m(W, g)' \Sigma(W)^{-1} m(W, g)] \tag{1.16}$$

where $\Sigma(W)$ is a positive definite matrix for any given W. To consistently estimate g_0, Ai and Chen (2003) replaced Θ by a finite dimensional sieve space Θ_q^n that becomes dense in Θ as n increases, where $q = \dim(\Theta_q)$. In addition, suppose $\hat{m}(w, g)$ is a nonparametric estimator of $m(w, g)$, then Ai and Chen (2003) proposed a sieve minimum distance (SMD) parameter that minimizes the sample analog of (1.16)

$$\hat{g} = \arg\min_{g \in \Theta_q^n} \frac{1}{n} \sum_{i=1}^{n} \hat{m}(W_i, g)' \hat{\Sigma}(W_i)^{-1} \hat{m}(W_i, g) \tag{1.17}$$

where $\hat{\Sigma}(W)$ is a consistent estimator of $\Sigma(W)$.

To consistently estimate $m(W, g)$, consider a sieve estimator. Let $p^k(W) = (p_1(W), ..., p_k(W))'$, where $\{p_j(W), j = 1, 2, ...\}$ is a sequence of known basis functions whose linear combination can approximate any square integrable real-valued function of W well. Then for each $g \in \Theta$, $m(W, g)$ can be approximated by $p^k(W)'\beta$ for some vector of coefficients β as $k \to \infty$. The linear sieve estimator for $m(W, g)$ is $\hat{m}(W, g) = p^k(W)'\hat{\beta}$, where $\hat{\beta}$ is the ordinary least squares estimate obtained by regressing $\rho(Z_i, g)$ on $p^k(W_i)'$. Note that with $\hat{\Sigma}(W) = I$, the SMD estimator has a GMM interpretation: define $P = (p^k(W_1), ..., p^k(W_n))'$, then the SMD estimator defined by (1.17) with $\hat{m}(W, g) = p^k(W)'\hat{\beta}$ is the GMM estimator based on the following unconditional moment restrictions:

$$E[\rho(Z, g) p_j(W)] = 0, j = 1, 2, ..., k \tag{1.18}$$

with weighting matrix $P'P$. Note that, this GMM interpretation implies an overidentifying restriction: $k \geq q$.

Under some regularity conditions imposed on $\rho, k,$ and $p^k(W)$, it can be shown that the SMD estimator is consistent. With further assumptions imposed on the metric in Θ and on the approximation rate of Θ_q^n, the SMD estimator can achieve a convergence rate $||\hat{g} - g_0|| = o_p(n^{-1/4})$. When g_0 consists of a parametric part θ_0, the $o_p(n^{-1/4})$ convergence rate is useful for constructing the asymptotic normality of $\hat{\theta}$. Ai and Chen (2003) in addition showed that the SMD estimator $\hat{\theta}$ is semiparametric efficient.

Besides their work, Newey and Powell (2003) also considered the estimation of conditional moment restricted model in a general setting. Their approach is similar to Ai and Chen (2003) while they were more focusing on the identification and consistency. In particular, they were among the first to consider the identification conditions of g_0 in nonparametric IV regression models (Example 1.3.2). Recently, Chen and Pouzo (2009a, 2009b) relaxed the compactness assumption on the parameter space. To achieve the consistency, they imposed a penalty term.

1.4. Nonparametric Instrumental Variable Regression

Although the nonparametric instrumental variable regression model is a special case of the nonparametric conditional moment restricted models, it is particularly of interest to us, because of its importance in many empirical applications and as a natural extension from the conventional instrumental variable regression model. Consequently, ever since Newey an Powell (2003), nonparametric IV has been receiving tremendous attentions in both statistics and econometrics literature.

The nonparametric IV regression is formally defined as:

$$Y = g_0(X) + \epsilon \tag{1.19}$$

where Y is the response variable, and X is the explanatory variable, which can be either scalar or mutivariate. Here ϵ denotes unobservable disturbances. The function g_0 is nonparametric. It satisfies regularity conditions but does not belong to a known, finite-dimensional parametric family. In addition, ϵ is assumed to be correlated with the explanatory variable X, and hence

$E(\epsilon|X) \neq 0$. Therefore classical nonparametric methods can't be applied here to estimate g_0. Instead, suppose we have available another observable W, for which

$$E(\epsilon|W) = 0 \tag{1.20}$$

In the literature of econometrics, W is known as ”instrumental variable”. We then have an opportunity to estimate g_0 based upon observed simple random samples of the triple (Y, X, W). The model (1.19) and (1.20) together is known as nonparametric regression model with presence of instrumental variables.

1.4.1. Identification

In a nonparametric setting, the restriction $E(\epsilon|W) = 0$ is important to the identification of structural function g_0. Newey and Powell (2003) characterized the identification in terms of the completeness of the conditional distribution of X given W as follows.

Consider the model

$$Y = g_0(X) + \epsilon, \qquad E(\epsilon|W) = 0,$$

Taking conditional expectation of Y yields

$$E(Y|W) = E[g_0(X)|W] \tag{1.21}$$

Since $E(Y|W)$ depends upon the conditional distribution of observable random variable $Y|W$, it is identified; hence the identification of g_0 depends upon the existence and uniqueness of integral equation (1.21), say, if $E[g(X)|W] = E[\tilde{g}(X)|W] = E(Y|W)$ almost surely implies

$g(X) = \tilde{g}(X)$. This is equivalent to the completeness of the conditional distribution of X given W. Therefore Newey and Powell (2003) obtained the following proposition.

Theorem 1.4.1 (Newey and Powell 2003)**.** *If* $E(Y|W) = E(g_0(X)|W)$ *is satisfied almost surely, then* g_0 *is point identified if and only if for all* $\delta(X)$ *with finite expectation,* $E[\delta(X)|W] = 0$ *implies* $\delta(X) = 0$ *a.s..*

Example 1.4.1 (Severini and Tripathi 2006, Example 3.2)**.** Let $Y = g_0(X) + \epsilon$, where $g_0 \in L_2(X)$. The regressor X is endogenous and we have an instrument W satisfying $E(\epsilon|W) = 0$ a.s. Suppose that $X = W + U$, where W and U are independent and identically distributed as Uniform $[-\frac{1}{2}, \frac{1}{2}]$. Define $\mathcal{M} = \{\delta \in L_2(X) : E(\delta(X)|W) = 0 a.s.\}$. Note that $E(\delta(X)|W = w) = \int_{w-1/2}^{w+1/2} \delta(x)dx$, it can be shown that $\int_{w-1/2}^{w+1/2} \delta(x)dx = 0$ for almost all $w \in [-\frac{1}{2}, \frac{1}{2}]$ if and only if $\delta(x) = \delta(1 + x)$ for almost all $x \in [-1, 0]$ and $\int_{-1}^{0} \delta(x)dx = 0$. Therefore, we have $\mathcal{M} = \{\delta \in L^2(X) : \delta(x) = \delta(1 + x) a.a. x \in [-1.0], \text{ and } \int_{-1}^{0} \delta(x)dx = 0\}$, which is clearly not a singleton $\{0\}$. Therefore, g_0 is not point identified.

In addition to the completeness of the conditional distribution, Severini and Tripathi (2006) framed the identification problem in a general Hilbert space setting. Consider Y, X and W as elements of a separable Hilbert space with inner product $(., .)$. Assume that $\mathcal{W}$ is a known linear subspace that contains the support of W, and call vector a orthogonal to $\mathcal{W}$ if $(a, w) = 0$ for all $w \in \mathcal{W}$, which we write as $a \perp \mathcal{W}$. Let $\mathcal{M}$ denote a linear subspace, and assume that, corresponding to Y, there exists an element $\mu_y \in \mathcal{M}$. Here μ_y is a summarization of the distribution of Y and may be viewed as the parameter of interest. In addition, assume that $Y - \mu_y \perp \mathcal{W}$. We call $\mathcal{M}$ the"model space" and $\mathcal{W}$ the"instrumental space". It is easy to see that if μ_y is not identified, say, suppose corresponding to Y, there exist two different μ_y, $\mu'_y \in \mathcal{M}$

such that $Y - \mu_y \perp \mathcal{W}$ and $Y - \mu_y' \perp \mathcal{W}$, then we have a nonzero vector $\mu_y - \mu_y' \in \mathcal{M}$ such that $\mu_y - \mu_y' \perp \mathcal{W}$, which follows that $\mathcal{M} \perp \mathcal{W}$. Hence we have

Theorem 1.4.2 (Severini and Tripathi 2006). *μ_y is identified if and only if $\mathcal{M}$ is not orthogonal to $\mathcal{W}$.*

In nonparametric regression problem, $\mathcal{M} = L_2(X)$ and $\mathcal{W} = L_2(W)$ are infinite dimensional linear subspaces consisting of square integrable functions. The restriction $E(\epsilon|W = 0)$ implies $Y - g_0(X) \perp L_2(W)$. Therefore it follows from Theorem 1.4.2 that a sufficient and necessary condition for g_0 to be point identified, is that if a function $f \in L_2(X)$ satisfies $E[f(X)h(W)] = 0$ for all $h \in L_2(W)$, then $f = 0$ a.s, which is equivalent to saying if $E[f(X)|W] = 0$ a.s. for $f \in L_2(W)$, then $f = 0$ a.s. This is the completeness of conditional distribution of $X|W$.

Another equivalent statement of point identification of g_0 was provided by Hall and Horowitz (2005). Define

$$\tilde{T} : L_2(X) \rightarrow L_2(W), \text{ such that } \tilde{T}(g)(w) = E[g(X)|W = w], \forall g \in L_2(X)$$
$$\mu(w) = E[Y|W = w]$$

It then follows from $E[\epsilon|W] = 0$ that $\tilde{T}(g_0) = \mu$. Therefore, g_0 is point identified if and only if $\tilde{T}$ is nonsingular.

1.4.2. Ill-posed inverse problem

As previously discussed, when g_0 is point identified, it is the unique solution to an integral equation $\tilde{T}(g_0) = \mu$, where $\tilde{T}$ is nonsingular, and is defined as

$$\tilde{T}(g)(w) = \int g(x) f_{X|W}(x|w) dx$$

where $f_{X|W}(x|w)$ is the conditional density function of $X|W$. if $f_{X|W}(x|w)$ is continuous in x for all w, then the integral operator $\tilde{T} : L_2(X) \rightarrow L_2(W)$ has a continuous kernel, and hence is compact. Therefore, although $\tilde{T}$ is continuous and nonsingular, it does not have a bounded inverse. As a result, g_0 cannot be estimated consistently by $\hat{T}^{-1}\hat{\mu}$, simply replacing $\tilde{T}^{-1}$ and μ by their consistent estimators, because even small error in the data will cause arbitrarily large error in the estimation. This problem is known as "ill-pose" problem, and was discussed by Kress (1999). The ill-posed problem is the main task to overcome in the estimation of nonparametric IV models, and there has been a huge literature concerned about it, see for example, Newey and Powell (2003), Chen and Pouzo (2009a,b), Florens and Simoni (2009a), and Hall and Horowitz (2005), among others.

Let us assume $g_0 \in L^2(X)$. Even if g_0 is not identified, in the sense that $\Theta_I = \{g \in L^2(X) : E(Y - g(X)|W)\} \neq \{g_0\}$, the problem of recovering Θ_I from the data is still ill-posed. To illustrate the problem, we need to introduce some additional notation. For any $g_1 \in L^2(X)$, let

$$[g_1] = \{g : E(g(X)|W) = E(g_1(X)|W)\}$$

which is an equivalence family of g_1 under $\tilde{T}$. Let $\mathcal{N}(\tilde{T}) = \{g : E(g(X)|W) = 0\}$, the null space of $\tilde{T}$. It deduces a quotient space $L^2(X)/\mathcal{N}(\tilde{T})$. Define

$$A : L^2(X)/\mathcal{N}(\tilde{T}) \rightarrow L^2(W)$$

$$A[g] = \tilde{T}g$$

It can be shown that $A^{-1} : L^2(W) \to L^2(X)/\mathcal{N}(\tilde{T})$ is not continuous. Therefore to recover $[g_0] = \Theta_I$ from $A^{-1}\mu$ requires some regularization techniques.

1.4.3. Estimation with point identification

In this subsection, we assume that g_0 is point identified, and focus on the consistent estimation of g_0. To overcome the inverse problem, one way is to restrict the parameter space for g_0 to a compact subspace of $L_2(X)$. This approach is based on a fact that, if $\tilde{T} : \Theta \to \mathcal{W}$, where $\tilde{T}$ is a compact nonsingular linear operator and Θ is a compact function space, then $\tilde{T}^{-1}$ is continuous on $\mathcal{W}$. Therefore the ill-posed problem is removed. Both Newey and Powell (2003) and Ai and Chen (2003) followed this approach. Alternatively, Hall and Horowitz (2005) and Chen and Pouzo (2009a,b) relaxed the compactness assumption, but put a regularization parameter on the operator. This so-called Tikhonov regularization approach requires the convergence rate assumptions on the eigenvalues of the linear operator. We present it in detail here.

In order for the regularization approach to be valid, it is essential to require the operator that identifies g_0 have positive eigenvalues. Therefore, we need to do some transformation first. Denote by f_X, f_W and f_{XW} the marginal densities of X and W, and the joint density of X and W, respectively. It may be proved from (1.1) and (1.2) that

$$E_W\{E(Y|W)f_{XW}(z,W)\} = \int\int g(x)f_{XW}(x,w)f_{XW}(z,w)dxdw \tag{1.22}$$

Define $\phi(z) = E_W\{E(Y|W)f_{XW}(z,W)\}$, and $t(x,z) = \int f_{XW}(x,w)f_{XW}(z,w)dw$. In addition define the linear operator T on $L_2(z)$ by $(T\psi)(z) = \int t(x,z)\psi(x)dx$. Therefore (1.22) can be written into the form of integral equation

$$Tg_0 = \phi$$

Assuming T is nonsingular, all its eigenvalues are positive. However, the ill-posed problem is still there because if we order the eigenvalues to be $\lambda_1 \geq \lambda_2 \geq ... > 0$, we have $\lambda_k \rightarrow 0$ as k increases. Let $\hat{T}$ be a nonparametric consistent estimator of T, instead of inverting $\hat{T}$, Hall and Horowitz (2005) imposed a regularization parameter a_n, which converges to zero as the sample size increases. Let $\hat{\phi}^{(-i)}$ be the kernel estimator of ϕ with the ith observation missing. Their proposed estimator is given by

$$\hat{g} = \frac{1}{n}\sum_{i=1}^{n}(\hat{T} + a_n)^{-1}\hat{\phi}^{(-i)}$$

For the consistency of $\hat{g}$, appropriate assumptions on the convergence rates of the eigenvalues o T and a_0 are necessary. In particular, a_0 should not decease too fast, otherwise the ill-posed problem comes back again. In addition, although $\hat{g}$ will still be consistent even if the restrictions on the eigenvalues of T do not hold, the rate of convergence of $\hat{g}$ will be very slow if the eigenvalues decrease too slowly. See Hall and Horowitz (2005) for detailed regularity conditions.

1.4.4. Inference with partial identification

It is also interesting to make inference about nonparametric IV models without assuming the point identification of g_0, for the following two reasons: First, as discussed earlier, the identification of g_0 depends on the completeness of the conditional distribution $X|W$. When the conditional distribution does not belong to the exponential family, the completeness assumptions is hard to check, and is in fact hard to be satisfied (see, for example, Example 1.4.1).

Another reason is that, sometimes instead of g_0 itself, we are only interested in one of its particular characteristics, say its linear functional $h(g_0)$. For example, in the nonparametric IV regression, if $g_0(x)$ is the inverse demand function, then its consumer surplus at some level x^* can be written as a functional $h(g_0) = \int_0^{x^*} g_0(x)dx - g_0(x^*)x^*$. In this case, the identification of g_0 might not be necessary. Severini and Tripathi (2006) have shown that without assuming g_0 to be identified, it is still possible to point identify its functional $h(g_0)$.

Let Θ be the parameter space for g_0, which is a collection of all the possible functions of g_0, defined to be a compact space $\Theta = \{g : \|\theta\|_s \leq B\}$ for some known B and norm $\|.\|_s$. Define

$$\Theta_I = \{g \in \Theta : E(Y|W) = E[g(X|W)]a.s.\}$$

When g_0 is not point identified, Θ_I is not equal to $\{g_0\}$. Santos (2007) developed methods for hypothesis testing in a nonparametric IV setting within a partial identification framework. The kind of hypothesis tests he allowed for are of the form

$$H_0 : \Theta_I \cap R \neq \emptyset \qquad H_1 : \Theta_I \cap R = \emptyset$$

where R is a set of functions that satisfy a property we wish to test for. Under the assumption that R is compact, the null hypothesis is equivalent to $H_0 : \inf_{g \in \Theta \cap R} E[(E(Y - g(X)|W))^2 f_W^2(W)] = 0$. The advantage of this transformed null hypothesis over $H_0 : \Theta_I \cap R \neq \emptyset$ is that no estimation of Θ_I is required.

Define $Q_n(g) = \frac{1}{n}\sum_{i=1}^n h_n(g, W_i)^2 \hat{f}_W^2(W_i)$, where $h_n(g, W_i)$ is the Nadaraya-Watson kernel estimator for $E[Y - g(X)|W = W_i]$, and $\hat{f}_W$ is the kernel estimator for f_W. Let $T_n(g)$

be a properly centered version of $Q_n(g)$. The test statistic is then $\inf_{\Theta\cap R} T_n(g)$. Santos (2007) showed with some regularity conditions that, under the null hypothesis, $\inf_{\Theta\cap R} T_n(g)$ converges in distribution to a Gaussian process.

For most hypotheses, the computation of the test statistic requires solving a minimization problem over a nonparametric set of functions $\Theta \cap R$. Santos (2007) addressed this challenge by approximating $\Theta \cap R$ with a sieve space $\Theta_q \cap R$, and the minimizing problem can be solved over $\Theta_q \cap R$ without losing the asymptotic results. Finally, the critical values can be found by either subsampling or Bootstrap procedure (Santos 2008b).

Recently, Kovchegov and Yildiz (2010) considered consistent estimation of the identified region Θ_I. Recall that in Section 1.4.2, we established that $\Theta_I = [g_0]$, which is an equivalence family in the quotient space $L^2(X)/\mathcal{N}(\tilde{T})$, hence it is straightforward to estimate $[g_0]$ using the same regularization techniques as what have been used for the point identification case. Kovchegov and Yildiz (2010) also considered constructing the confidence region for Θ_I.

1.4.5. Bayesian approach

Recently, Florens and Simoni (2009a) proposed a quasi-Bayesian nonparametric approach to estimating the structural function g_0.

Let $L^2(X)$ denote the square integrable Hilbert space for g_0. It is assumed that the prior of g_0 is a Gaussian measure on $L^2(X)$ that defines a mean element g^* and a covariance operator $\sigma^2\Omega_0$, and that the error term is $N(0,\sigma^2)$. Under some conditions, the conditional posterior of g_0 given σ^2 is Gaussian with mean $n^{-1/2}A(Y_1,...,Y_n)^T + b$, and covariance operator $\sigma^2(\Omega_0 - AK\Omega_0)$, where

$$A = \Omega_0 K^*(K\Omega_0 K^*)^{-1}, \qquad b = (I - AK)g^*$$

$K(g) = E(g(X)|W)$ for each $g \in L^2(X)$, and $K^*(h) = E(h(W)|X)$ for each $h \in L^2(W)$. Note that $K\Omega_0K^*$ is a compact operator which is not continuously invertible. Therefore $(K\Omega_0K^*)^{-1}$ is a non-continuous operator that amplifies the measurement error in $(Y_1, ..., Y_n)$ and thus the posterior is not consistent in the frequentist sense.

In order to solve the lack of continuity of $(K\Omega_0K^*)^{-1}$, Florens and Simoni (2009a) replaced the standard posterior distribution with a regularized posterior distribution, by applying a Tikhonov regularization scheme to the inverse of $K\Omega_0K^*$, so that to get $(K\Omega_0K^* + a_nI)^{-1}$, where a_n is a regularization parameter that plays the same regularization rule in Hall and Horowitz (2005). By using this regularization scheme, it can be shown that the posterior mean is consistent.

Note that, if we write $U = g_0(X) - E(g_0(X)|W) + \epsilon$, and $K(g_0) = E(g_0(X)|W)$, then $Y = K(g_0) + U$. The regularization scheme therefore used here to achieve the posterior consistency can be applied to general statistical linear inverse problems of the form $Y = K(g_0) + U$, where the operator K is compact so that its inverse is not continuous on the whole space of reference. Florens and Simoni (2009b) proposed an extended version of Zellner's g-prior (Zellner 1986) to correct the ill-posedness.

1.5. Bayesian GMM and Bayesian Empirical Likelihood

As we have seen so far, many econometric and statistical models can be characterized by moment conditions, either conditional $Em(X, \theta) = 0$ or unconditional: $E[m(X, \theta)|W] = 0$, where the parameter θ can be either finite dimensional or infinite dimensional. (Note that for moment inequality model $Em(X, \theta) \geq 0$, we can always impose a bias parameter $\lambda \geq 0$ and rewrite it as $E[m(X, \theta) - \lambda] = 0$.) To conduct appropriate Bayesian analysis of θ, one needs to specify the likelihood function. However, in many applications, what is known or assumed, is just the moment conditions instead of the distribution of the data generating process. On the other hand, even if the data generating process can be directly assumed, for instance, in linear regression model, $E[Y - x'\theta|x] = 0$, the error term $Y - x'\theta_0$ can be assumed to be normal, however, the assumption inevitably suffers from the model mis-specification problem. When the likelihood function is mis-specified, the result will be inconsistent. Therefore, it is more robust to construct the likelihood function of θ directly from the moment conditions, without assuming the true underlying likelihood function.

One way of constructing such a moment-based likelihood is to use the Bayesian GMM (Yin 2009). Write $\bar{m}(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta)$, and let $V = Var(m(X, \theta_0))$, for any $\theta \in \Theta$ where θ_0 denotes the true structural parameter. Define the Bayesian GMM likelihood:

$$L_{GMM}(\theta) = e^{-n\bar{m}(\theta)'V(\theta)\bar{m}(\theta)}$$

Suppose $p(\theta)$ is the prior of θ, the GMM posterior is then given by $p(\theta|Data) \propto p(\theta)L_{GMM}(\theta)$. When V is not known, one can replace it by a consistent estimator $\hat{V}$. Let

$$\Theta_I = \arg\min_{\theta \in \Theta} Em(X, \theta)'VEm(X, \theta).$$

If θ_0 is point identified, $\Theta_I = \{\theta_0\}$, Chernozhukov and Hong (2003) showed that the GMM posterior is consistent, meaning that $p(\theta|Data)$ converges in probability to any small neighborhood of θ_0. Alternatively, one can use the empirical likelihood (Owen 1990):

$$L_{EL}(\theta) = \max_{p_i : i=1,\ldots,n} \{\prod_{i=1}^{n} p_i | \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i m(X_i, \theta) \geq 0, p_i \geq 0\}$$

Again, the Empirical likelihood posterior can be constructed via $p(\theta|Data) \propto p(\theta)L_{EL}(\theta)$. This Bayesian procedure was considered by Lazar (2003), who heuristically showed that L_{EL} can be used as a valid likelihood function for every absolutely continuous prior in terms of the posterior coverage probability. As for the formal posterior consistency, Chernozhukov and Hong (2003) also showed that the posterior converges in probability to any small neighborhood of θ_0 if θ_0 is identified by the moment conditions. However, to my best knowledge, so far there is no result on the posterior consistency of partial identification case (when Θ_I is not minimized uniquely) yet.

One of the biggest concerns of using either Bayesian GMM or Bayesian empirical likelihood, is the Bayesian interpretation of the moment-based likelihood function. Ideally, by the Bayes' law, the likelihood function should be interpreted as the conditional probability of the data given the parameter. Since the true underlying likelihood is not assumed, one needs to figure out a similar interpretation of the "artificial" likelihood function to be used. Fortunately, both Bayesian GMM and Bayesian empirical likelihood have such a good interpretation. Kim (2002) gave a Bayesian interpretation of L_{GMM} with the "limited information likelihood" idea. He showed that, although L_{GMM} is not the true likelihood function, it is the best approximation subject to the moment restrictions. It is also known that the empirical likelihood L_{EL} is the solution to the following constrained optimization problem:

$$\min_{L} K(L||L_n), \text{ subject to } \int m(x,\theta)dL, \forall\theta \in \Theta$$

where $K(L||L_n)$ is the Kulback-Leibler divergence, and L_n is the empirical distribution based on $X_1, ..., X_n$. Therefore, L_{EL} is the best approximation to the empirical distribution of the data, and the latter uniformly converges to the underlying true likelihood according to Gliverko-Cantelli theorem.

1.6. Structure of this Dissertation

The remainder of this dissertation is organized as follows: Chapter 2 studies the moment inequality models (1.2) with a Bayesian approach. I will first derive the moment-based likelihood function, which is the Bayesian GMM, and then study the large sample properties of the posterior distribution of the partially identified parameter. In particular, I will consider two cases subsequently: when the identified region has nonempty interior and when the identified region has empty interior.

Chapter 3 addresses the moment and model selection problem in moment inequality models. I will derive the posterior of each moment and model combination, and show that the posterior is exponentially small if the selected combination defines an empty identified region.

Chapter 4 studies the Bayesian estimation of the semi-nonparametric conditional moment restricted model (1.3). Because the structural nonparametric function may not be point identified, as discussed in Section 1.4, I will put the discussion into the partial identification setup. The nonparametric function is approximated by a sieve series, with the number of sieve terms diverging to infinity as n increases. The posterior distribution will then be constructed on the sieve approximation instead of the true nonparametric function directly. I will show that the posterior is consistent. Finally, an empirical example on estimating the Engel curve will be conducted.

Chapter 5 is concerned with the Bayesian classification problem, where we measure both the classification rule and classification risk simultaneously using the posterior distribution.

Suppose the loss function is given by $\rho(W, \theta)$, where the classification rule depends on certain parameter θ, and the risk is denoted by r, it then follows by definition that

$$E\rho(W, \theta) = r$$

I will use the Bayesian empirical likelihood function approach to construct the posterior distribution based on this moment condition.

The technical proofs are collected in the Appendices.

CHAPTER 2

Bayesian Analysis in Moment Inequality Models

2.1. Introduction

In this chapter, I study a type of moment conditions that have been rapidly studied by econometricians in recent years: moment inequalities. In moment inequality models, since the parameter of interest is allowed to be not point identified, the treatment is very flexible in dealing with incomplete data, such as missing data or censored data. Consequently, the moment inequality model has many important applications in economics and biostatistics. See Tang (2008) and Haile and Tamer (2003) for the application in empirical auctions, Ciliberto and Tamer (2009) in game theory, Khan and Tamer (2009) in survival analysis, Horowitz and Manski (2000), Manski (2003), and Molinari (2010) in missing treatments, and references therein.

Let $(\bar{\Omega}, \mathcal{A}, P)$ denote a probability space. Suppose we are interested in some structural parameter $\theta_0 \in \mathbb{R}^d$ that satisfies a set of *moment inequality* conditions:

$$Em_j(X, \theta_0) \geq 0, j = 1, ..., p \tag{2.1}$$

where $m_j(., \theta), i = 1, ..., p$ are known real-valued moment functions. X is an observable random vector defined on $(\bar{\Omega}, \mathcal{A}, P)$, and assume we observe independent and identically distributed or stationary realizations $X^n = \{X_1, ..., X_n\}$ of X. A model that is characterized by moment inequalities (2.1) is usually called a *moment inequality model*.

As discussed in Chapter 1, a key feature of moment inequality models is that θ_0 is not necessarily point identified: there exists more than one solution to the inequalities in (1.1) if $Em_j(X, \theta_0)$ is viewed as a function of θ_0. In other words, let Θ be the parameter space that contains θ_0, and define

$$\Omega = \{\theta \in \Theta, Em_j(X, \theta) \geq 0, j = 1, ..., p\} \tag{2.2}$$

then Ω can be a non-singleton set. In this case, we say that θ_0 is *partially identified* on Ω, and Ω is called the *identified region*.

Many partially identified models are characterized by such moment inequalities, where the parameter of interest is only partially identified and therefore cannot possibly to be consistently estimated. Under this framework, since the identified region captures all the information about the parameter, it becomes one of the most interesting subjects of study in moment inequality models. See Section 1.2 in Chapter 1 for the examples of moment inequality models and the corresponding literature in both frequentist and Bayesian approach.

In this chapter, I study a Bayesian approach to the moment inequality models. The Bayesian procedure provides distributional information for the partially identified parameter both inside and outside the identified region, through its posterior distribution. The advantages of using posterior distributions to characterize the parameters are many. First of all, as pointed out by Poirier (1998), a Bayesian analysis of partial identification models is always possible if a proper prior for the parameters is specified. If we have some a priori information on θ_0, then by using a properly chosen prior distribution, the resulting posterior density may not be flat within the identified region; this provides evidence that the parameter is more likely to lie in some particular area. Secondly, even with a flat prior distribution, when θ_0 is multi-dimensional, the posterior density of some components of θ_0 may no longer be flat, due to the shape of the identified region. Hence if we are interested in these components of θ_0, the posterior density can still provide extra information on their locations within the identified region. As a third advantage, it can be shown asymptotically that the posterior density has support only on the identified region. Containing more information, a posterior density can always be used to estimate the identified region, but not vice versa. Finally, the MCMC method is a very powerful method to draw samples from the posterior, which can be used for approximations of the calculation of the posterior statistics. In addition, those posterior samples can also be used in frequentist methods to estimate the identified region, by, e.g., minimizing an econometric criterion function in Chernozhukov Hong and Tamer (2007).

To my best knowledge, so far all the methods proposed in the partial identification literature other than Liao and Jiang (2010) use traditional posteriors based on the likelihood function instead of the moment inequalities. Our Bayesian approach proceeds with a more general framework. In contrast to the previous work, we do not need to have a full probability model for the observed data. Starting from moment inequalities $Em(X,\theta_0) \geq 0$, where $m(X,.)$ is a known function of θ_0, we put some bias parameter $\lambda_0 \geq 0$ so that $Em(X,\theta_0) = \lambda_0$, and place prior distributions on (θ_0, λ_0). Then the posterior density of θ_0 can be derived based on a limited information likelihood function, which is generated by the conditional asymptotic distribution of $\frac{1}{n}\sum_{i=1}^{n} m(X_i,\theta_0) - \lambda_0$ given (θ_0, λ_0), integrating out λ_0. I study in detail the frequentist behaviors of the posterior density function of θ_0. I derive the bounds of convergence rates of the posterior density both inside and outside of the identified region. It will be shown that there is a big "gap" between them. Once the posterior density and its frequentist properties are obtained, it is easy to derive consistent estimators for the identified region. However, I point out that a posterior density provides more information than a region estimation, since it can also incorporate prior information and describe how likely the true parameter is distributed both inside and outside of the identified region.

2.2. Moment Inequality Models

2.2.1. Limited Information Likelihood

Suppose for $\theta \in \mathbb{R}^d$, we have moment inequality conditions:

$$Em_j(X_i, \theta) \geq 0, j = 1, ..., p$$

Let $m(X, \theta) = (m_1(X, \theta), m_2(X, \theta), ..., m_p(X, \theta))^T$. The moment inequalities can then be rewritten as

$$Em(X, \theta) = \lambda, \text{ for some } \lambda \in [0, \infty)^p \tag{2.3}$$

Here θ is the structural parameter of interest, e.g., $\theta = EY$, the mean of the unobserved random variable Y in Example 1.1 and 1.2, and λ is the bias parameter of $Em(X, \theta)$, e.g., $\lambda = (EY_2 - \theta, \theta - EY_1)^T$, in Example 1.1. Let θ_0 be the true parameter value of θ, and λ_0 be the true bias parameter when $\theta = \theta_0$. Suppose the prior of θ_0 is supported on a large enough compact set that contains the identified region. We are interested in constructing the marginal posterior for θ_0.

In addition, let $\bar{m}(\theta) = \frac{1}{n}\sum_{i=1}^n m(X_i, \theta)$, and $G(\theta, \lambda) = \bar{m}(\theta) - \lambda$, then after the bias parameter λ is introduced, G can be considered as the "de-biased" sample moment. In other words, G is an estimating function with $EG(\theta, \lambda) = 0$. It is over-parameterized, meaning that the dimension of (θ, λ) is bigger than the dimension of G, and hence we can not consistently estimate θ_0 by solving $G(\theta, \lambda) = 0$ directly.

Under some regularity conditions, by the central limit theorem,

$$\sqrt{n}G(\theta, \lambda)|_{\theta=\theta_0,\lambda=\lambda_0} \rightarrow^{d} N_p(0, V_0) \tag{2.4}$$

where $V_0 = Var(m(X, \theta_0))$. We can therefore formally construct a "likelihood" function:

$$p(X^n|\theta, \lambda) = \frac{1}{\sqrt{\det(\frac{2\pi}{n}V_0)}} e^{-\frac{n}{2}G(\theta,\lambda)^T V_0^{-1} G(\theta,\lambda)} \tag{2.5}$$

Note that for $\theta \neq \theta_0$, (2.4) is not true in general. In fact, we can't find a $\lambda \in [0, \infty)^p$ such that $Em(X, \theta) = \lambda$ for $\theta \notin \Omega$. Hence (2.5) is not the large sample conditional pdf of G for general (θ, λ). The asymptotic result (2.4) alone would not allow us to derive a likelihood function over the entire $\Theta \times [0, \infty)^p$. To solve this problem, Kim (2002) introduces the concept of Limited Information Likelihood. For each parameter $\theta \in \Theta$, although (2.5) may not be the true probability density of X^n, it is shown to be proportional to the density that is closest to the true density in the Kullback-Leibler distance, among a family of densities satisfying the moment condition $EG(\theta, \lambda) = 0$. The "likelihood" in (2.5) is therefore the limited information likelihood of θ and λ, which is the best approximation to the true density that satisfies the moment restrictions. The concept of the Kullback-Leibler information distance and applications of it can be found in a number of works such as Cover and Thomas (1991) and Zellner (1994).

Let $p(\lambda)$ be the marginal prior of λ. Assume λ and θ are independent, i.e., the conditional prior of λ given θ is equal to the marginal prior of λ. Since we are only interested in θ, we thus integrate out λ to obtain the Limited Information Likelihood function for θ:

$$\begin{aligned} L(\theta) &= p(X^n|\theta) \\ &= \int_{[0,\infty)^p} p(X^n|\theta,\lambda)p(\lambda|\theta)d\lambda \\ &= \int_{[0,\infty)^p} p(X^n|\theta,\lambda)p(\lambda)d\lambda \end{aligned} \tag{2.6}$$

The fact that λ is a location parameter of (2.5) makes the problems solvable. This will be described in detail in Section 3.1.

In practice, the asymptotic variance V_0 in (2.5) is not known, but it can be shown to have very little influence on the inference about θ, in the current situation of partially identified moment inequality models. In future expositions, we will replace V_0 by a pre-specified nonsingular matrix V, and show that $L(\theta)$ has good and very similar frequentist properties for inference on θ, whatever V is chosen. (A more delicate treatment would be to approximate V_0 by a sample analog and replace the true parameter θ_0 in V_0 by the unknown argument θ. This will be left for future work. We expect that similar techniques will lead to similar results in this treatment, but the technical details can be much more complicated.)

2.2.2. A General Result on the Posterior Set Estimation

I first define some notation that will be used subsequently. Throughout this chapter, let A^c and $int(A)$ denote the complement and interior of a set A respectively. In addition, following CHT's notation, $\forall\delta > 0$, let $(\Omega^c)^{-\delta}$ be the $\delta-$contraction of Ω^c,

$$(\Omega^c)^{-\delta} = \{\theta \in \Theta : d(\theta, \Omega) \geq \delta\}$$

Let $B(\omega, r)$ denote an open ball around ω: $B(\omega, r) = \{\theta : d(\omega, \theta) < \delta\}$, where $d(\omega, \theta)$ denotes the Euclidean distance between ω, θ. Let $d_H(A, B)$ denote the Hausdorff distance between set A and B.

$$d_H(A, B) = \max\{\sup_{a \in A} d(a, B), \sup_{b \in B} d(A, b)\}$$

where $d(a, B) = \inf_{b \in B} d(a, b)$. We say a set estimator A_n consistently estimates Ω, if

$$d_H(A_n, \Omega) \to 0 \text{ in probability.}$$

Moreover, for two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n \succ b_n$ if $\frac{a_n}{b_n} \to \infty$. Finally we write w.p.a.1 to represent "with probability approaching one in the probability distribution of X^n as $n \to \infty$".

Let $p(\theta)$ be the prior of θ, then by Bayes' rule the posterior of θ satisfies

$$p(\theta | X^n) \propto p(\theta) L(\theta) \tag{2.7}$$

It is desirable for the posterior to possess some "good" frequentist properties. Roughly speaking, we want to see that the posterior density of θ concentrates near Ω and drops dramatically to zero outside Ω, with a high probability as n increases. The significant difference of such an asymptotic behavior between inside and outside the identified region implies that the resulting posterior has the capability to produce consistent set estimation for Ω. Such a relation between a "good" posterior and its capability to estimate Ω is demonstrated below for a scalar function of Ω. (A more general estimation of Ω itself will also be discussed later in Section 2.3)

The posterior probability that θ belongs to a set A is

$$P(\theta \in A|X^n) = \int_A p(\theta|X^n)d\theta$$

Definition 2.2.1 (dense)**.** *A subset $A \subset \Omega$ is said to be dense in Ω if $\forall \omega \in \Omega \backslash A$, and any neighborhood U_w of ω, $U_w \cap A \neq \phi$.*

An equivalent definition of dense subsets in real analysis is that the closure of A is Ω, i.e. cl$(A) = \Omega$. I will consider the large sample behavior of the posterior distribution on a dense subset of Ω.

Suppose instead of θ we are interested in the functions of θ: $g(\theta)$, where $g : \Theta \to \mathbb{R}$ is some known continuous mapping. For instance, if we are interested in the ith component of θ, then $g(\theta) = \theta_i$. Let $g(\Omega) = \{g(\theta) : \theta \in \Omega\}$, the image of g. We are interested in estimating $g(\Omega)$ directly. Let us impose the following assumptions:

Assumption 2.2.1. *Θ is compact.*

Assumption 2.2.2. *Ω is compact and connected.*

In moment inequality models, the compactness of Ω follows from assuming $Em_j(X, .) : \Theta \to \mathbb{R}$ to be continuous for each j. Here Ω is assumed to be connected so that the intermediate value theorem on a topological space holds.

Assumption 2.2.3. *$g : \Theta \to \mathbb{R}$ is continuous on Θ.*

The estimation of $g(\Omega)$ to be constructed is based on the inverted posterior cdf of $g(\theta)$. Let $F_g(x) = P(g(\theta) \leq x|X^n)$, the posterior cdf of $g(\theta)$. Denote

$$F_g^{-1}(y) = \inf\{x : F_g(x) \geq y\}.$$

Then $x \geq F_g^{-1}(y)$ if and only if $F_g(x) \geq y$. The following theorem provides a general consistency result of a set estimator of $g(\Omega)$ based on the posterior cdf. Notice that since it can be shown $g(\Omega) = [\inf_{\theta\in\Omega} g(\theta), \sup_{\theta\in\Omega} g(\theta)]$, one might think that a more natural set estimator can be constructed by finding estimators for the end points of the interval $g(\Omega)$. This idea works, for example, when $g(\Omega) = [EY_1, EY_2]$ where $EY_1 < EY_2$ and both Y_1 and Y_2 are observable. In this case, Ω can be estimated by $[\bar{Y_1}, \bar{Y_2}]$. However, in a more general setting, estimating the end points $\inf_{\theta\in\Omega} g(\theta)$ and $\sup_{\theta\in\Omega} g(\theta)$ would require the estimation of Ω first. The estimator proposed in the following theorem provides a way of estimating the interval directly.

Theorem 2.2.1. *Under Assumptions 2.2.1-2.2.3, assume there exists* $\{\pi_n\}_{n=1}^{\infty}$, $\pi_n \to 0$ *such that*

(1) $\forall\delta > 0$, $P(\theta \in (\Omega^c)^{-\delta}|X^n) = o_p(\pi_n)$

(2) *There exists a dense subset* $A \subset \Omega$*, such that* $\forall\omega \in A$*, and* $\forall\rho > 0$*,*

$P(\theta \in B(\omega, \rho)|X^n) \succ \pi_n$ *w.p.a.1*

Let $\hat{g} = [F_g^{-1}(\pi_n), F_g^{-1}(1 - \pi_n)]$*, then*

$$d_H(\hat{g}, g(\Omega)) \to 0 \quad \textit{in probability.}$$

There are some remarks regarding this theorem:

(1) The consistent set estimator depends on the choice of π_n. However, we do not pursue an operational way of constructing the estimator based on the posterior distribution in this chapter, because there are many frequentist methods to achieve this purpose, for instance, CHT, Beresteanu and Molinari (2008), etc. this chapter is more focused on the posterior distribution itself. The purpose of this theorem is to demonstrate that the posterior can be used to consistently estimate the identified region, if needed. The posterior distribution can actually provide more information than the identified region, when taking into account the prior.

(2) We can also provide an exact credible region (based on, say, setting $\pi_n = 0.025$ for instance) for the true parameter, conditional on the observed data. This is parallel to the provision of the confidence intervals with required coverage probabilities in the frequentist approaches of Imbens and Manski (2004), Rosen (2008), etc.

(3) It is possible to get an optimal rate of π_n for optimal convergence rate in Hausdorff distance. We leave it as a future work.

We will see in the next section that under some regularity conditions, the posterior distribution of θ satisfies conditions 1 and 2 in this theorem, which describe the frequentist properties of the posterior. In addition, we will also propose a consistent estimator for Ω directly based on the log-posterior density.

2.3. Posterior Properties: When the Identified Region Has Nonempty Interior

In this section it is assumed that the identified region contains a non-empty interior $int(\Omega)$. I assume it is dense in Ω, then it is of interest to study the asymptotic properties of the posterior distribution inside $int(\Omega)$.

2.3.1. The Posterior Density

Following the discussions in Section (2.2), let us define the limited information likelihood for θ:

$$L(\theta) = \int_{[0,\infty)^p} \frac{1}{\sqrt{\det(\frac{2\pi V}{n})}} e^{-\frac{n}{2}(\bar{m}(\theta)-\lambda)^T V^{-1}(\bar{m}(\theta)-\lambda)} p(\lambda) d\lambda \tag{2.8}$$

where V is some pre-selected positive definite matrix that doesn't depend on θ. We will use a multivariate exponential distribution as the prior on λ throughout this chapter.

$$p(\lambda) = (\prod_{i=1}^{p} \psi_i) e^{-\psi^T \lambda}, \quad \psi = (\psi_1, ..., \psi_p)^T \in [0,\infty)^p, \lambda \in [0,\infty)^p$$

where ψ is pre-specified. I use the exponential prior for ease of integration over λ. More general choices of $p(\lambda)$ may not allow the integration to be carried out analytically, but the large sample behavior of the posterior should remain unchanged.

Let Z_θ be a p- dimensional multivariate normal random vector, with mean $(\bar{m}(\theta) - \frac{V\psi}{n})$, and variance covariance matrix $\frac{V}{n}$. Then a straightforward calculation of (3.1) leads to

$$L(\theta) = P(Z_\theta \geq 0) e^{-\psi^T \bar{m}(\theta) + \frac{1}{2n}\psi^T V \psi} (\prod_{i=1}^{p} \psi_i) \tag{2.9}$$

and we have $p(\theta|X^n) \propto p(\theta)L(\theta)$.

For large values of n, by uniform WLLN, $\bar{m}(\theta)$ is bounded on Θ w.p.a.1. Thus for fixed ψ and V, $e^{-\psi^T \bar{m}(\theta) + \frac{1}{2n}\psi^T V \psi}(\prod_{i=1}^{p} \psi_i)$ is bounded away from zero and infinity. Therefore the only term that characterizes the large sample properties of the posterior should be $P(Z_\theta \geq 0)$. Moreover, the variance covariance matrix of Z has order $O_p(n^{-1})$, so we would expect that $\lim_{n\to 0} P(Z_\theta \geq 0) = 1$ in probability if and only if $\bar{m}(\theta) - \frac{V\psi}{n} \geq 0$ w.p.a.1. This depends on whether θ belongs to Ω. For large n, the posterior density is positive inside Ω, and drops to zero exponentially fast as θ gets away from Ω. I will formally examine these asymptotic properties and derive the convergence rate of the posterior probabilities.

2.3.2. Large Sample Analysis

I now conduct a large sample analysis to the posterior distribution of the parameter θ.

Assumption 2.3.1. *$int(\Omega)$ is non-empty and is dense in Ω.*

The assumption that $int(\Omega)$ is dense in Ω can be restated as follows: for any ω on the boundary of Ω, and any neighborhood U_w of ω, U_w contains points in $int(\Omega)$. Most of the identified regions characterized by moment inequalities possess such property. The case when $int(\Omega)$ will be empty is considered in the next section.

Assumption 2.3.2. *$Em_j(X,.) : \Theta \rightarrow \mathbb{R}$ is continuous, for each $j = 1, ..., p$.*

This assumption guarantees that $Em(X,\theta)$ is bounded in any compact set, and that the uniform law of large number holds. The next assumption puts a regularity condition on the prior of θ.

Assumption 2.3.3. *(i) $p(\theta)$ is continuous, and bounded away from zero and infinity on Ω. (ii) $P(\min_j Em_j(X,\theta) = 0) \equiv \int_{\{\theta:\min_j Em_j(X,\theta=0\}} p(\theta)d\theta = 0$.*

Let v_{jj} be the jth diagonal element of V. We can write

$$\Omega^c = \{\theta : \min_j Em_j(X,\theta) < 0\} = \left\{\theta : \min_j \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} < 0\right\}$$

For any $\delta > 0$, let

$$A_\delta = \left\{ \theta : \min_j \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} < -\delta \right\}$$

Apparently, $A_\delta \subset \Omega^c$.

Lemma 2.3.1. *Under Assumptions 2.2.1, 2.2.2 and 2.3.2, if* $\exists$ *some* $a_n \to 0$ *such that* $\forall \delta > 0$, $P(\theta \in A_\delta | X^n) = o_p(a_n)$, *then* $\forall \epsilon > 0$, $P(\theta \in (\Omega^c)^{-\epsilon} | X^n) = o_p(a_n)$.

Theorem 2.3.1. *Under Assumptions 2.2.1, 2.2.2, and 2.3.1-2.3.3,*

(1) $\forall \delta > 0$, *for some* $\alpha > 0$,

$$P(\theta \in (\Omega^c)^{-\delta} | X^n) = o_p(e^{-\alpha n})$$

(2) $\forall$ *nonempty open set* $\Xi \subset \Omega$, *in probability*

$$\liminf_{n \to \infty} P(\theta \in \Xi | X^n) > 0$$

Hence we are able to distinguish the asymptotic behavior of the posterior: for large value of n, the posterior density is only supported on a neighborhood of the identified region, and the posterior distribution drops to zero exponentially fast on any subset that is separated from Ω. Based on these findings, we can construct consistent estimators for both Ω and its continuous mappings. For the latter task we can now apply Theorem 2.2.1. Suppose $g(.)$ is a continuous real-valued function on Θ, let $F_g^{-1}(y)$ be the $y-$quantile of the posterior cdf of $g(\theta)$.

Theorem 2.3.2. *Under Assumption 2.2.1-2.2.3 and 2.3.1-2.3.3, for any sequence* $\pi_n = o_p(1)$ *satisfying* $\forall a > 0$, $e^{-an}/\pi_n \to 0$,

$$d_H([F_g^{-1}(\pi_n), F_g^{-1}(1-\pi_n)], g(\Omega)) \to 0 \textit{ in probability.}$$

It is also possible to consistently estimate Ω directly using the posterior density function. The consistency is based on the fact that the posterior density attains its peak inside Ω and is asymptotically supported on the entire identified region. In addition, it drops to zero outside Ω at an exponential rate. Therefore, by properly choosing a cut off value ϵ_n, the region where the log-posterior density function exceeds its peak subtracting ϵ_n should eventually converge to Ω.

Theorem 2.3.3. *Under Assumptions 2.2.1-2.2.3, 2.3.1-2.3.3, let $n \succ \epsilon_n \succ 1$. Define*

$$A_n = \{\theta : \max_{\omega \in \Theta} \ln p(\omega|X^n) - \ln p(\theta|X^n) \leq \epsilon_n\}$$

then

$$d_H(A_n, \Omega) \to 0 \textit{ in probability}$$

Remark 2.3.1. The estimation established in Theorem 2.3.3 is easy to implement, because:

(1) Note that

$$\max_{\omega \in \Theta} \ln p(\omega|X^n) - \ln p(\theta|X^n)$$

$$\begin{aligned} =\; & \max_{\omega \in \Theta} \left(\ln p(\omega)L(\omega) - \ln \int_{\Theta} p(\theta)L(\theta)d\theta \right) - \Bigg(\ln p(\theta)L(\theta) \\ & - \ln \int_{\Theta} p(\theta)L(\theta)d\theta \Bigg) \\ =\; & \max_{\omega \in \Theta} \ln p(\omega)L(\omega) - \ln p(\theta)L(\theta) \end{aligned}$$

Thus it's no need to normalize $p(\theta)L(\theta)$, avoiding numerically integrating $p(\theta)L(\theta)$.

(2) Maximizing $\ln p(\theta)L(\theta)$ is computationally workable, since the maxima is attained only inside Ω, where $p(\theta)L(\theta)$ is quite smooth, hence Newton-Raphson's algorithm can carry out the maximization.

(3) Set $a_n = \max_{\omega \in \Theta} \ln p(\omega | X^n) - \epsilon_n$, then $A_n = \{\theta : \ln p(\theta) \geq a_n\}$. The boundary $\{\theta : \ln p(\theta) - a_n = 0\}$ is a closed curve with dimension $d - 1$.

2.4. Posterior Properties: When the Identified Region Has Empty Interior

When Ω has no interior, moment inequality models may contain exact moment conditions.

$$\begin{aligned} &Em_{1j}(X, \theta_0) \geq 0, j = 1, ..., r \\ &Em_{2j}(X, \theta_0) = 0, j = 1, ..., p \end{aligned} \tag{2.10}$$

Moon and Schorfheide (2009b) have considered the estimation problem assuming θ_0 is point identified by the exact moment conditions. Let

$$m_1(X, \theta) = (m_{11}(X, \theta), ..., m_{1r}(X, \theta))^T, m_2(X, \theta) = (m_{21}(X, \theta), ..., m_{2p}(X, \theta))^T$$

If $p \geq \dim(\theta_0)$, and there doesn't exist a pair of moment functions (m_{2i}, m_{2j}) such that $\{\theta \in \Theta : Em_{2i}(X, \theta) = 0\} = \{\theta : Em_{2j}(X, \theta) = 0\}$, then θ_0 is point identified by $Em_2(X, \theta_0) = 0$. Moon and Schorfheide (2009b) showed that by using the overidentifying information provided by $Em_1(X, \theta_0) \geq 0$, the empirical likelihood estimators reduce the asymptotic mean squared errors. In this section, I will relax this point identifying restriction, and allow θ_0 to be partially identified by model (2.10).

The identified region is defined by

$$\Omega = \{\theta : Em_1(X, \theta) \geq 0, Em_2(X, \theta) = 0\}$$

In our setting, Ω shrinks to a lower dimension sub-manifold of $\{\theta : Em_2(X,\theta) = 0\}$ with boundaries defined by linear or nonlinear hyperplanes $\{\theta : Em_1(X,\theta) = 0\}$. One of the problem one needs to take into account when considering the asymptotic behaviors of the posterior distribution is that Ω has zero Lebesgue measure, due to the loss of dimensionality. Thus integrating over Ω is always zero. The limit of posterior density is known as Dirac function:

$$\lim_{n\to\infty} p(\theta|X^n) = \begin{cases} +\infty & \theta \in \Omega \\ 0, & \theta \notin \Omega \end{cases} a.s.$$

Thus $\lim_{n\to\infty} p(\theta|X^n)$ is not a real valued function of θ.

However, it is still possible to study the large sample properties of the posterior distributions completely on Θ. Like in $int(\Omega) \neq \phi$ case, a dense subset in Ω plays an important role in characterizing such behaviors. Define

$$\Xi = \{\theta \in \Omega : Em_1(X,\theta) > 0\} \tag{2.11}$$

It is assumed that Ξ is dense in Ω.

2.4.1. Derivation for Limited Information Likelihood

Suppose $X^n = \{X_1, ..., X_n\}$ is a stationary realization of X. Define $\bar{m}_j(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_j(X_i,\theta)$, for $j = 1, 2$. Like before, we introduce auxiliary parameter λ to moment inequalities and define

$$G(\theta,\lambda) = \begin{pmatrix} \bar{m}_1(\theta) - \lambda \\ \bar{m}_2(\theta) \end{pmatrix}, \theta \in \Theta, \lambda \in [0,\infty)^r$$

For any positive definite $r \times r$ matrix V not depending on θ, define limited information likelihood:

$$L(\theta) = \int_{[0,\infty)^r} \frac{1}{\sqrt{\det(\frac{2\pi V}{n})}} e^{-\frac{n}{2}G(\theta,\lambda)^T V^{-1} G(\theta,\lambda)} p(\lambda) d\lambda$$

Write V^{-1} into subblocks

$$V^{-1} = \begin{pmatrix} \Sigma_1 & \Sigma_3 \\ \Sigma_3^T & \Sigma_2 \end{pmatrix}, \Sigma_1 : r \times r, \Sigma_2 : p \times p$$

Let us still place an exponential prior: $p(\lambda) = (\prod_{i=1}^r \psi_i) e^{-\psi^T \lambda}, \psi, \lambda \in [0,\infty)^r$, then we have

$$\begin{aligned} L(\theta) &= \int_{[0,\infty)^r} \frac{1}{\sqrt{\det(\frac{2\pi V}{n})}} \exp\left(-\frac{n}{2}(\bar{m}_1(\theta) - \lambda, \bar{m}_2(\theta)) \begin{pmatrix} \Sigma_1 & \Sigma_3 \\ \Sigma_3^T & \Sigma_2 \end{pmatrix} \begin{pmatrix} \bar{m}_1(\theta) - \lambda \\ \bar{m}_2(\theta) \end{pmatrix} \right) p(\lambda) d\lambda \\ &= \frac{\prod_{i=1}^r \psi_i}{\sqrt{\det(V_2)}} P(Z \geq 0) e^{\tau} \end{aligned}$$

where:

- Z follows multivariate normal distribution with mean μ, variance covariance matrix $\frac{\Sigma_1^{-1}}{n}$, $\mu = \bar{m}_1(\theta) + \Sigma_1^{-1}\Sigma_3^T \bar{m}_2(\theta) - \frac{1}{n}\Sigma_1^{-1}\psi$.
- $V_2 = (\Sigma_2 - \Sigma_3^T \Sigma_1^{-1} \Sigma_3)^{-1}$. If $V = Var(m_1, m_2)$, then by the matrix inversion formula, $V_2 = Var(m_2)$.
- $\tau = -\frac{n}{2}\bar{m}_2(\theta)^T V_2^{-1} \bar{m}_2(\theta) - \psi^T(\Sigma_1^{-1}\Sigma_3^T \bar{m}_2(\theta) + \bar{m}_1(\theta)) + \frac{1}{2n}\psi^T \Sigma_1^{-1}\psi$

Roughly speaking, when $\theta \notin \Omega$, either $Em_2(X,\theta) \neq 0$ or $\exists Em_{1j}(X,\theta) < 0$. When $Em_2(X,\theta) \neq 0$, since V_2^{-1} is also positive definite, $e^{\tau} \to 0$; when $Em_2(X,\theta) = 0$ but $Em_{1j}(X,\theta) < 0$ for some j, then for large n, the jth component of $\mu < 0$. Since the covariance matrix of Z has order $O(n^{-1})$, $P(Z \geq 0) \to 0$. Therefore, $L(\theta) \to 0$ outside Ω. When

$\theta \in \Omega$, by central limit theorem, $\bar{m}_2(\theta) = O_p(n^{-1/2})$, hence $e^{\tau} = O_p(1)$. In addition, for large n, $P(Z \geq 0) \approx 1$. Thus $L(\theta) = O_p(1)$.

2.4.2. Posterior Distribution

Let $p(\theta)$ denote the prior on θ, then $p(\theta|X^n) \propto p(\theta)L(\theta)$.

Assumption 2.4.1. Ξ *defined in (2.11) is dense in* Ω.

This assumption states that if θ_0 satisfies $Em_2(X, \theta_0) = 0$ and $Em_{1j}(X, \theta_0) = 0$ for some $j = 1, ..., r$, then in any neighborhood of θ_0 we can find θ_1 such that $Em_1(X, \theta_1) > 0$ and $Em_2(X, \theta_1) = 0$.

Suppose all the other components of $Em_1(X, \theta_0)$ except for j are positive. By continuity of $Em_1(X, .)$, they remain to be positive in a small neighborhood of θ_0. Suppose Assumption 2.4.1 does not hold, then within some neighborhood U of θ_0, $\forall \theta \in U \cap \Omega$, $Em_{1j}(X, \theta) = 0$, and $Em_{1i}(X, \theta) > 0$, for $i \neq j$. Since Ω is connected, we argue that $Em_{1j}(X, \theta) \equiv 0$ on $U \cap \Omega$. Hence intuitively, Assumption 2.4.1 says that for each i, hyperplane $\{\theta : Em_{1i}(X, \theta) = 0\}$ has no part that overlaps with $\{\theta : Em_2(X, \theta) = 0\}$.

Example 2.4.1. This example shows that Assumption 2.4.1 is satisfied by the interval regression model. Suppose we have moment inequalities $E(Z_1 Y_1) \leq E(Z_1 X^T)\theta \leq E(Z_1 Y_2)$ and exact moment condition $EZ_2(Y_3 - X^T\theta) = 0$, where Z_i, $i = 1, 2$ are r_1 and r_2 dimensional vectors of instrumental variables respectively, with each instrument being positive almost surely, and not sharing same components. Y_i is scalar $i = 1, 2, 3$, and $\theta \in \mathbb{R}^d$. $Y_2 > Y_3 > Y_1$ a.s. Let

$W = (Z_1, Z_2, X, Y_1, Y_2, Y_3)$, then

$$m_1(W,\theta) = \begin{pmatrix} Z_1(Y_2 - X^T\theta) \\ Z_1(X^T\theta - Y_1) \end{pmatrix}, m_2(W,\theta) = Z_2(Y_3 - X^T\theta)$$

Assume $r_2 < d$ so that θ can not be point identified by $Em_2(W,\theta) = 0$. Let us also assume there exists a unit vector δ such that $EZ_2X^T\delta = 0$ but $EZ_{11}X^T\delta < 0$, where Z_{11} denotes the first component of Z_1. In this interval instrumental variable regression model,

$$\Xi = \{\theta : E(Z_1Y_1) < E(Z_1X^T)\theta < E(Z_1Y_2); EZ_2Y_3 = EZ_2X^T\theta\}$$

We now show Ξ is dense.

For any $\theta \in \Omega\backslash\Xi$, we have $EZ_2(Y_3 - X^T\theta) = 0$. For simplicity, let us assume the first component of m_1: $EZ_{11}(Y_2 - X^T\theta) = 0$, and for the jth component of m_1: $Em_{1j}(W,\theta) > 0$, for all $j > 1$. Then in a small neighborhood of θ, $Em_{1j}(W,.) > 0$ for all $j > 1$. For small enough $\epsilon > 0$, let $\theta_1 = \theta + \epsilon\delta$, then

$$Em_2(W,\theta_1) = EZ_2(Y_3 - X^T\theta) - \epsilon EZ_2X^T\delta = 0$$

$$Em_{11}(W,\theta_1) = EZ_{11}(Y_2 - X^T\theta_0) - \epsilon EZ_{11}X^T\delta = -\epsilon EZ_{11}X^T\delta > 0$$

Therefore $\theta_1 \in B(\theta, 2\epsilon) \cap \Xi$.

Assumption 2.4.2. *(i) $Em_{1j}(X,\theta)$ is continuous on Θ for each j.*

(ii) $Em_{2j}(X,\theta)$ is Lipschitz continuous on Θ for each j.

Assumption 2.4.3. *w.p.a.1, for any $\beta_n \to \infty$,*

$$\sup_{\theta \in \Theta} ||\bar{m}_2(\theta) - Em_2(X, \theta)||^2 \leq \frac{\ln \beta_n}{n}$$

Assumption 2.4.4. *$p(\theta)$ is continuous, and bounded away from zero and infinity on Ω.*

Theorem 2.4.1. *Under Assumptions 2.2.1,2.2.2, and 2.4.2-2.4.4,*

(1) $\forall \delta > 0$, *for some $\alpha > 0$,*

$$P(\theta \in (\Omega^c)^{-\delta} | X^n) = o_p(e^{-\alpha n})$$

(2) $\forall \omega \in \Xi$, $\forall \delta > 0$, *for all $\beta_n \to \infty$, we have in probability*

$$P(\theta \in B(\omega, \delta) | X^n) \succ \frac{1}{\beta_n} n^{-d/2}$$

where $d = \dim(\theta_0)$

Like the case when $int(\Omega) \neq \phi$, let $g(.)$ be a continuous real-valued function on Θ, let $F_g^{-1}(y)$ be the $y-$quantile of the posterior cdf of $g(\theta)$.

Theorem 2.4.2. *Under Assumptions of Theorem 2.4.1, if $\{\pi_n\}_{n=1}^{\infty}$ is such that $e^{-\alpha n} \prec \pi_n \prec n^{-\beta}$, for any $\alpha > 0$ and some $\beta > \frac{d}{2}$, then*

$$d_H([F_g^{-1}(\pi_n), F_g^{-1}(1 - \pi_n)], g(\Omega)) \to 0 \text{ in probability.}$$

2.5. Monte Carlo Experiments

This section presents some Monte Carlo simulation results. I first provide evidence on the finite sample behaviors of the consistent estimators described in the previous sections as well as the posterior distribution. The models described in Example 1 and Example 2 in Chernozhukov Hong and Tamer (2007) are simulated.

Example 2.5.1 (Interval data)**.** Consider the interval censored data problem, where the parameter of interest $\theta = E(Y)$ satisfies moment inequalities:

$$E(Y_2 - \theta) \geq 0, E(\theta - Y_1) \geq 0$$

Set $Y_1 \sim N(0, 0.1)$ and $Y_2 \sim N(5, 0.1)$, then $\Omega = [0, 5]$. Y_1 and Y_2 are generated independently, and observations with $Y_1 > Y_2$ are discarded. I also set $\psi_1 = 0.1$, $\psi_2 = 0.5$, $V = I$, the identity matrix in the likelihood function. In addition, let us place flat prior on θ. The estimated identified interval of θ described both in Theorem 2.3.2 with $g(\theta) = \theta$ and in Theorem 2.3.3, for sample size $N = 500, 1000, 5000$, and various choices of ϵ_n, π_n are reported.

Table (2.1) reports the estimation of Ω given by Theorem 2.3.3. To compare the results corresponding to the choices of ϵ_n, for each interval $[a, b]$, we calculate $\gamma = (a-0)^2 + (b-5)^2$. We find $\epsilon = \ln \ln n$ performs better than the other two choices, for it has a lower γ value.

To construct the estimator based on the posterior distribution function, I carried out the Metropolis algorithm to draw $B = 5000$ samples from the posterior distribution, and calculated the π_n- quantile of the empirical cdf with various choices of π_n. For the Metropolis algorithm, the initial value was set to $\theta_0 = 1$ and a jump distribution $\tilde{\theta} \sim N(\theta_j, 0.5)$. Table(2.2) reports the findings with $\pi_n = e^{-\sqrt{n}}, n^{-1}$, and $1/\ln n$. As can be seen, $\pi_n = \frac{1}{n}$ appears to be a better choice compared with other two. It is also noticed that $\pi_n = 1/\ln n$ tends to zero too slow to fully estimate the entire identified interval: the estimated interval shrinks too much inside Ω.

Table 2.1. Estimation based on posterior density

ϵ_n	$\sqrt{n}$	$\ln n$	$\ln\ln n$
$n = 500$	[-0.2841, 5.2634]	[-0.123, 5.113]	[-0.0389, 4.702]
$n = 1000$	[-0.2362, 5.2267]	[-0.1135, 5.0977]	[-0.0342, 4.9110]
$n = 5000$	[-0.1158, 5.1233]	[-0.0477, 5.0476]	[-0.0202, 4.9779]

Table 2.2. Estimation based on empirical cdf

π_n	$e^{-\sqrt{n}}$	$\frac{1}{n}$	$\frac{1}{\ln n}$
$n = 500$	[-0.0716, 5.0418]	[-0.0498, 5.0069]	[0.4048, 3.3447]
$n = 1000$	[-0.0422, 4.9983]	[-0.0383, 5.0164]	[0.3304, 3.2542]
$n = 5000$	[-0.0155, 5.0098]	[-0.0063, 4.9927]	[0.2717, 3.8012]

In addition, Figure 2.1 plots the posterior density function of θ with two choices of priors: flat prior and $N(0, 0.25)$ prior. Theoretically one needs to truncate the normal distribution so that the priors are supported on a compact set. However, since the tail of normal density function is very thin and we can choose a very large parameter space, we believe a normal prior is workable here. We see that when a flat prior is used, the posterior density function is high on the entire identified interval $[0, 5]$, but when the prior is set to be $N(0, 0.25)$, most posterior mass falls in $[0, 2]$, which tends to underestimating the true identified interval. However, with this more informative prior, the posterior provides more information about the location of θ.

Example 2.5.2 (Interval outcomes in regression models)**.** I simulated the instrumental inequality model

$$E(ZY_1) \leq E(ZX^T)\theta \leq E(ZY_2)$$

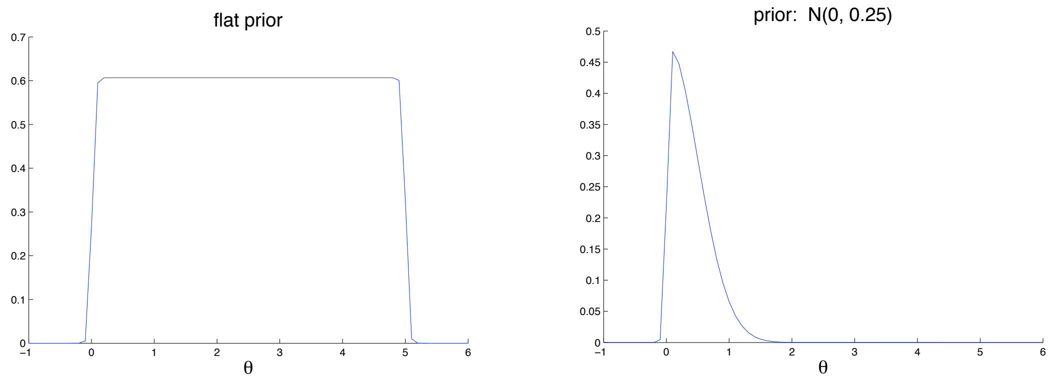


Figure 2.1. The posterior density function of θ

where $\theta = (\theta_1, \theta_2)^T, X = (X_1, X_2)^T, Y = (Y_1, Y_2)^T \in \mathbb{R}^2$. Generate $X \sim N_2((1,1)^T, I_2)$. Let $Z_1 = X_1 + X_2$ and $Z_2 = X_1 + 2X_2$. Generate $Y_1 \sim N(3, 0.1)$, $Y_2 \sim N(6, 0.1)$ independently. We discard a stack of generated data if either Z_1 or Z_2 is negative. The identified region is $\Omega = \{\theta : 2 \leq \theta_1 + \theta_2 \leq 4, 9 \leq 4\theta_1 + 5\theta_2 \leq 18\}$, a two dimensional region with parallelogram boundary. To estimate this model, set $\psi = (0.1, 0.1, 0.5, 0.5)^T$, $V = I$. Fixing sample size $n = 500$, we conduct the Metropolis algorithm to draw $B = 5000$ samples from the posterior distribution.

Let us first put a flat prior on θ. Figure 2.2 (left) displays the parallelogram boundary of Ω as well as 5000 draws from the posterior distribution. Most of the draws fall uniformly inside the identified set except for those close to the two opposite angels of the parallelogram. We can see there is small "bias" at boundaries.

In order to show that when a more informative prior is applied, the posterior distribution indeed provides more information about the location of the true parameter inside the identified region, I repeated the same MCMC procedure but with prior distribution

$$\theta_1 \sim N(10, 12^2),\ \theta_2 \sim N(-6, 12^2) \tag{2.12}$$

where θ_1 and θ_2 are a priorily independent. This prior can be used when, for instance, a previous study estimates that $E\theta_1 \approx 10$ and $E\theta_2 \approx -6$, with the same standard deviation 12. Figure 2.2 (right) displays 5000 MCMC draws from the posterior derived from prior (2.12). We see that the draws mostly concentrate on the right bottom corner inside the identified region, which is close to $(10, -6)$, showing that our Bayesian approach indeed provides more information on θ in this case than the frequentist method, which would only estimate the identified region and provide confidence set, but not tell how θ is distributed inside it.

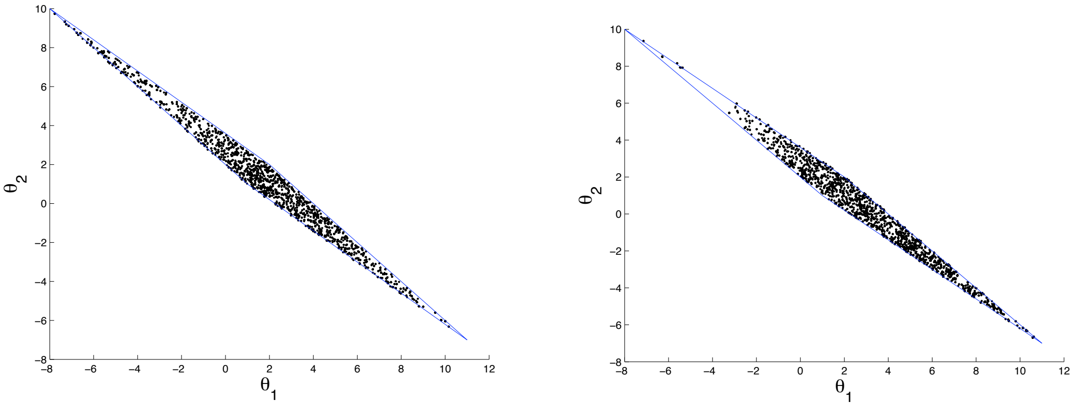


Figure 2.2. The identified set and MCMC draws
Left: flat prior; Right: prior(5.1)

Example 2.5.3 (When $int(\Omega)$ is empty)**.** In this example, I simulated an interval instrumental regression model with exact moment conditions. Consider,

$$E(Z_1Y_1) \leq E(Z_1X^T)\theta \leq E(Z_1Y_2), \qquad E(Z_2X^T)\theta = E(Z_2Y_3)$$

Generate $(X_1, X_2) \sim N_2((1,1)^T, I_2)$, and $Z_1 = X_1 + X_2$, $Z_2 = -2X_1 + 2X_2$. $(Y_1, Y_2)^T \sim N_2((3,6)^T, 0.1I_2)$, independent of X. Let $Y_3 = Z_1 + 3$. The identified region is then given by

$$\Omega = \{(\theta_1, \theta_2) : \theta_1 = \theta_2, 2\theta_1 + \theta_2 \leq 4\}$$

To estimate Ω, let us choose a positive definite weight matrix

$$V^{-1} = \begin{pmatrix} I_2 & \Sigma_3 \\ \Sigma_3^T & 6 \end{pmatrix}$$

where $\Sigma_3 = (1,2)^T$.

Figure (2.3) displays the identified region as well as 10,000 draws using Metropolis algorithm, with two choices of $\psi^1 = (0.5, 0.5)^T$, and $\psi^2 = (0.01, 0.01)^T$ respectively.

The identified interval of θ_1 was also estimated, which is [1, 2] theoretically. Table (2.3) reports $[F_e^{-1}(\pi_n), F_e^{-1}(1-\pi_n)]$ based on the empirical cumulative distribution function F_e of 5000 draws from the posterior distribution.

Table 2.3. Estimation of $\Omega_1 = [1, 2]$ based on the empirical cdf

π_n	$e^{-\sqrt{n}}$	$\frac{1}{\sqrt{n}}$	$\frac{1}{\ln n}$
$n = 500$	[1.1384, 2.0295]	[1.0068, 1.9331]	[1.0904, 1.6207]
$n = 1000$	[1.0809, 1.9425]	[0.9620, 1.8844]	[1.1183, 1.8874]
$n = 5000$	[1.1045, 1.8551]	[0.9944, 1.9575]	[1.1878, 1.9729]

2.6. An Empirical Missing Data Example with Fictitious Data

In this section I apply the proposed Bayesian approach to a simple missing data problem, with fictitious data. Suppose we conduct a survey to estimate the employment rate in a certain population. Let $Y_i = I(i$ is employment$)$ indicate whether person i is employed. Because there

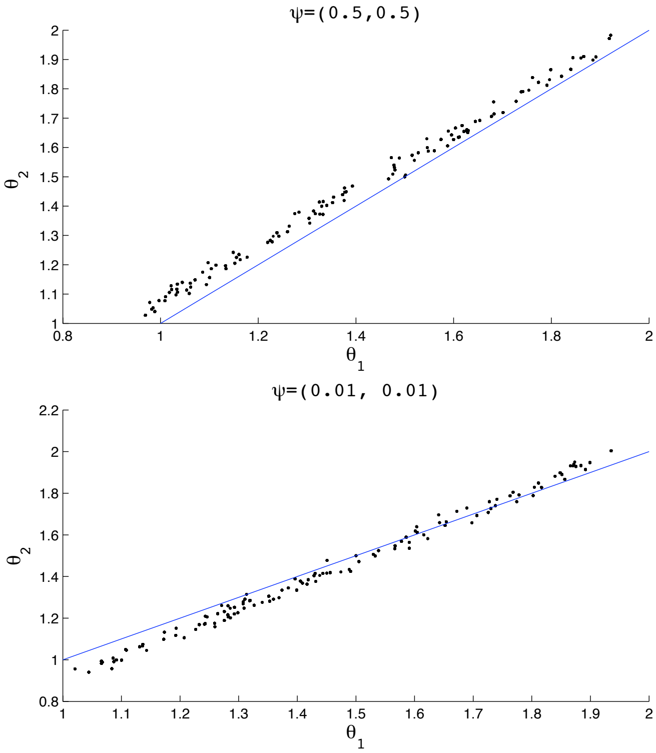


Figure 2.3. The identified set and MCMC draws

are people who refused to answer the questionnaires, hence the nonresponse is where the missing data problem comes from. Let $Z_i = I(i \text{ is not missing})$ indicates whether Y_i is not missing. Suppose we draw a simple random sample from a population (Y, Z). We are interested in the overall employment rate in this population $\theta = P(Y = 1)$, which is the parameter of interest. Because in practice people who respond the questionnaires are more likely to be employed than those who didn't response, therefore we assign the following fictitious population parameter values:

$$P(Y = 1|Z = 1) = 0.3, P(Y = 1|Z = 0) = 0.7, P(Z = 0) = 0.33$$

The missing data assumption does not hold in this example. The true parameter is $\theta_0 = 0.5$, which is the true employment rate in this specific population. By the discussion in Example 1.1.2, θ_0 is not point identified, but satisfies the moment inequalities (1.5).

I simulated 5,000 data, of which 1,622 were missing. The proportion of employment among people who responded is

$$\hat{P}(Y = 1|Z = 1) = \frac{2,355}{5,000 - 1,622} = 0.69.$$

Hence the missing-at-random assumption, although guarantees the point identification, tends to over-estimate the employment rate. The confidence interval calculated based on missing-at-random is $[0.68, 0.71]$.

Suppose we make another survey study on another 25 people. With such a small sample size, we are able to put additional effort to track all the surveyed 25 people such that none of them are missing. It is then observed that the proportion of employment of this new sample is 0.45, with standard error $\sqrt{0.45(1 - 0.45)/25} = 0.1$. We can then incorporate this as an informative prior $p(\theta) \sim N(0.45, 0.1^2)$, and then obtain the posterior based on $p(\theta)$. This procedure is equivalent to combining both the original and the new data.

Figure (2.4) displays the posterior density curve of θ_0 using either uniform prior on $[0, 1]$ or the informative prior $p(\theta)$ respectively. When the informative prior is used, the posterior density is not flat within the identified region, which also achieves the peak around the true parameter value. If the additional sampling is representative of the entire population, this informative prior and therefore the corresponding posterior should be reliable.

Finally, we compare the Bayesian credible interval with the frequentist confidence interval obtained by Imbens and Manski (2004). The Bayesian 95% credible interval $[a, b]$ is defined

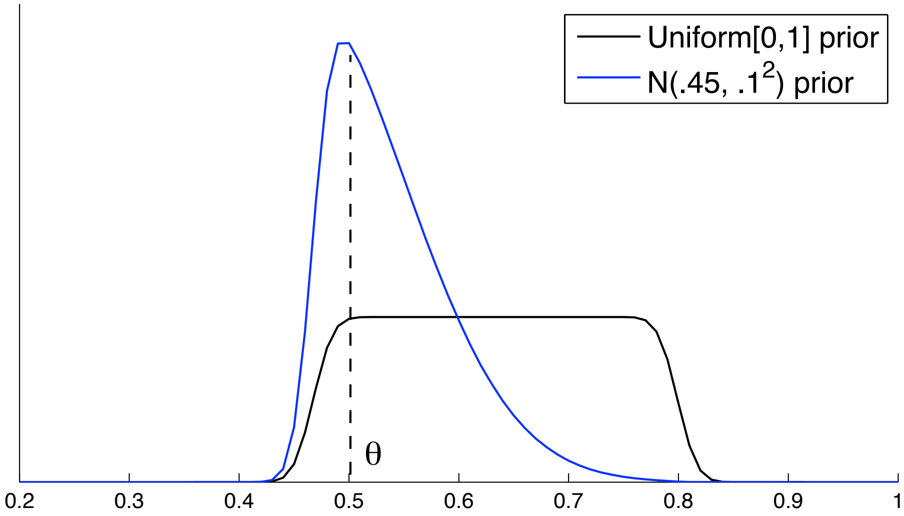


Figure 2.4. The identified set and MCMC draws

such that $P(\theta \leq a|Data) = 0.025$, and $P(\theta \leq b|Data) = 0.975$; the frequentist 95% confidence interval $[c, d]$ converges uniformly such that $\lim_n \inf_{\theta \in \Theta} P(\theta \in [c, d]) \leq 95\%$.

Table 2.4. 95% Confidence v.s. 95% Credible Interval

Method	C.I.	Length
Freq. Confidence Imbens & Manski (2004)	$[0.465, 0.801]$	0.336
Bayes. Credible Uniform$[0, 1]$ prior	$[0.477, 0.792]$	0.315
Bayes. Credible $N(.45, 0.1^2)$ prior	$[0.465, 0.679]$	0.214

Moon and Schorfheide (2009a) showed that in partially identified models, the Bayesian interval is always smaller than the frequentist confidence interval. Our computed results are consistent with their conclusion. Especially when the informative prior is used, the Bayesian credible region is much smaller than that was obtained by Imbens and Manski (2004).

2.7. Conclusions

In partially identified models, there are two different objects to make inferences: one is the identified region and the other is the true parameter. The simulation results demonstrate that when dealing with the first goal, a flat prior is appropriate; to achieve the second goal, an informative prior is more preferable. Hence in this case one should include as much information on the prior as possible. The Bayesian approach is specially attractive in dealing with the second goal, since the posterior distribution can provide more information about the inside of the identified region, because of the prior distribution.

Based on the posterior distribution, we can in principle construct a credible set for the true parameter conditional on the data with a required coverage probability (This is out of the scope of this chapter, but it is straightforward by using the posterior density function). Moon and Schorfheide (2009a) derived a Bayesian credible set for the true parameter and compared it with the frequentist confidence interval and concluded that while frequentist condence intervals usually extend beyond the boundaries of the identified set, the Bayesian credible sets are located in the interior of the identified set. In the framework of this chapter, it is also possible to derive a Bayesian credible set for the identified region if one can express the identified region explicitly in terms of θ and λ, which can be an interesting topic for future work.

CHAPTER 3

Model and Moment Selection in Moment Inequality Models

3.1. Introduction

Similar to the selecting the moment conditions in GMM, there is a moment and model selection problem in moment inequality models. Suppose there are p candidate moment inequalities

$$Em_j(X,\theta) \geq 0, j = 1, ..., p$$

with a k-dimensional parameter vector $\theta = (\theta_1, ..., \theta_k)^T$ that belongs to the parameter space $\Theta_1 \times \cdots \times \Theta_k$. The moment selection problem refers to selecting the best subset of the moment inequalities among all the possible candidates, while the model selection procedure addresses the problem of selecting the best model that is characterized by setting some components of the parameter to be zero. Such a candidate model can be a parameter subspace like $\{0\} \times \Theta_2 \times \cdots \times \Theta_k$. Therefore, the moment/model selection procedure produces a combination of moment inequalities and a parameter subspace. Consider the following example:

Example 3.1.1 (Interval censored regression)**.** (See, e.g., Example 1 of CHT 2007.) Let Y be a real valued random variable which lies in $[Y_1, Y_2]$ almost surely; Y_1 and Y_2 are observed random variables, but Y is not observed. (Sometimes one may assume that $Y_2 = Y_1 + 1$ as in the case when Y_1 is the recorded integer part of Y.) . Assume that

$$Y = X^T\theta + \epsilon$$

where X is a regressor vector. In addition, there exists a observable random vector Z such that $E(\epsilon|Z) = 0$. Here Z can be part of the regressors in X or a set of instrumental variables when X is endogenous. It the follows that $E(ZY) = E(ZX^T)\theta$. Due to $Y_1 \leq Y \leq Y_2$, we then have moment inequalities

$$EZ(Y_2 - X^T\theta) \geq 0, EZ(X^T\theta - Y_1) \geq 0 \tag{3.1}$$

In this example, the moment selection problem can correspond to selecting the instrumental variables (components of Z), while the model selection problem is related to selecting the useful explanatory variables (components of X) that have nonzero regression coefficients.

□

A similar selection problem in point identified case was previously considered by Andrews and Lu (2001), where they applied their approach to dynamic panel data models. We provide a similar example as follows.

Example 3.1.2 (Dynamic Panel Data)**.** Consider a dynamic panel data model

$$y_{it} = z_{it}'\theta_t + \eta_i + v_{it}$$

Here y_{it} is the dependent variable, censored between $y_{it}^L \leq y_{it} \leq y_{it}^U$. Hence instead of y_{it}, econometricians can only observe y_{it}^L, y_{it}^U. Also, v_{it} is an unobserved error, η_i is an unobserved individual effect, and θ_t are unknown parameters of interest. The distributions of η_i and v_{it} are not specified, and hence the limited information likelihood based on some moment conditions given below may be preferable. All of the random variables are assumed to be independent across individuals i.

The regressor $z_{it} = (X_{it}, w_{it})$ is an observed vector, where $X_{it} = (x_{it-1}, ..., x_{it-L})$ includes L lags of some covariates that may be exogenous, predetermined, or endogenous, where $L \geq 0$. The true lag length L_0 may be unknown. In addition, z_{it} also includes exogenous variables w_{it}, which are contained in an observed vector Z_{it}. The vector Z_{it} may also contain variables that do not enter the regression function. Such variables can be employed as instrumental variables. The following assumptions are imposed to derive the moment conditions.

$E\eta_i = Ev_{it} = 0, \forall t = 1, ..., T.$

$Ev_{it}Z_{it+1} = ... = Ev_{it}Z_{iT} = 0, \forall t = 1, ..., T.$

We may further partition Z_{it} into variables that are either uncorrelated with η_i or not, and achieve additional moment conditions. We do not do so here for simplicity. The moment conditions implied are

$$\begin{aligned}
&EZ_{it}(y_{it} - y_{it-1}) = EZ_{it}(z'_{it}\theta_t - z'_{it-1}\theta_{t-1}), \forall t = 1, ..., T \\
&Ey_{it} = Ez'_{it}\theta_t, \forall t = 1, ..., T.
\end{aligned}$$

Assume $\{Z_{it} : i = 1, ..., n,\}$ is generated from a distribution with support $supp(Z_t)$, which is compact for each t. Since one can always transform Z_{it} into $Z_{it} - \inf supp(Z_t)$, hence without loss of generality, we assume $Z_{it} \geq 0$. As y_{it} is censored in $[y^L_{it}, y^U_{it}]$, we have moment inequalities

$$\begin{aligned}
&EZ_{it}(y^L_{it} - y^U_{it-1}) \leq EZ_{it}(z'_{it}\theta_t - z'_{it-1}\theta_{t-1}) \leq EZ_{it}(y^U_{it} - y^L_{it-1}), \forall t = 1, ..., T \\
&Ey^L_{it} \leq Ez'_{it}\theta_t \leq Ey^U_{it}, \forall t = 1, ..., T.
\end{aligned}$$

Setting different lag coefficients to zero yields models with different number of lags in X_{it}. Therefore, the model selection refers to selecting the lagged variables of X_{it}. In addition, θ_t

also contains covariates of other variables w_{it}. If the effects of some particular components of w_{it} are of special interest, we should always keep the corresponding coefficients in the model.

□

As illustrated by the previous example, by allowing the dimension of the parameter space to change, we consider the case where the parameter vector may incorporate several models. By setting different elements of θ equal to zero, one obtains different models. Andrews and Lu (2001) also gave another example where a model may have structural breaks in the parameters. For example, when $t \leq t_0$, $\theta_t = \theta_0$, which is the pre-break value; when $t > t_0$, $\theta_t = \theta_0 + d\theta_1$, which adds a post-break deviation to the pre-break value. Here t_0 may be unknown. There may be multiple time breaks, and one can stack the break values into a single vector of parameters $\theta = (\theta_0, d\theta_1, ...)$. Different sets of post-break deviations can denote changes at different times. If the post-break deviations are set equal to zero, then one obtains the model with no structural breaks at that time.

3.2. Posterior Setup

Suppose we have p candidate moment inequalities

$$Em_j(X, \theta) \geq 0, j = 1, ..., p$$

with a k-dimensional parameter vector $\theta = (\theta_1, ..., \theta_k)^T \in \Theta_1 \times \cdots \times \Theta_k$. Here the possible moment inequalities and corresponding subsets of the parameter space are known. What is not known is which ones are the best.

Instead of selecting the moment inequalities and the parameter subspace as two separate procedures, I select them as a combination simultaneously. The selection procedure is based on the posterior probabilities. I assign prior probabilities to each candidate moment/model, and then derive the posterior probabilities based on the limited information likelihood described previously in Chapter 2, by integrating out the structural and nuisance parameters (θ, λ).

Let us define a combination $C_s = (M_{s_1}, \Theta_{s_2})$, with a vector index $s = (s_1, s_2)$, $s_1 \in \{1, 2, ..., 2^p - 1\}$, and $s_2 \in \{1, ..., 2^k\}$. Here M_{s_1} denotes a subset of moments, for instance, $M_{s_1} = \{m_1\}$, or $M_{s_1} = \{m_1, m_2\}$, etc. Then there are $2^p - 1$ number of such possible subsets. In addition, we denote by Θ_{s_2} as the parameter subspace corresponding to the selected model. By definition, Θ_{s_2} is the subset of vectors with one or more components fixed to be zero. There are 2^k possible Θ_{s_2}'s. (Notice that we can select none of the parameters, in which case the model is a reduced model, for example, in Cox proportional hazard model, if all the parameters are set to be zero, we get the baseline model.) The combination C_s combines both the candidate moment functions and the parameter subspace together. When selecting a subset of moment inequalities, we also specify a subspace of the structural parameter.

Example 3.2.1 (Example 3.1.1 continued)**.** Let $\Theta_1 \times \Theta_2$ be the parameter space for (θ_1, θ_2), chosen large enough so that $\{(\theta_1, \theta_2) : 0.2 \leq \frac{1}{3}\theta_1 + \theta_2 \leq 0.4, -0.1 \leq \theta_2 \leq 0.1\} \subset \Theta_1 \times \Theta_2$. A scope of candidate combinations can be any of the following:

$$\begin{array}{ll}
\{E(Z_1X^T\theta - Z_1Y_1)\}, & \Theta_1 \times \Theta_2 \\
\{E(Z_1X^T\theta - Z_1Y_1), E(Z_1Y_2 - Z_1X^T\theta)\}, & \Theta_1 \times \Theta_2 \\
\{E(Z_2X^T\theta - Z_2Y_1)\}, & \{0\} \times \Theta_2 \\
\{E(Z_1X^T\theta - Z_1Y_1), E(Z_1Y_2 - Z_1X^T\theta), E(Z_2Y_2 - Z_2X^T\theta)\}, & \Theta_1 \times \{0\} \\
 & \vdots
\end{array}$$

$\{E(Z_2Y_2 - Z_2X^T\theta)\}$,	$\Theta_1 \times \Theta_2$

Definition 3.2.1. *A combination $C_s = (M_{s_1}, \Theta_{s_2})$ is compatible if and only if*

$$\inf_{\theta \in \Theta_{s_2}, \lambda \in [0,\infty)^m} ||EM_{s_1}(X,\theta) - \lambda||^2 = 0$$

where m denotes the number of candidate moment functions in M_{s_1}.

Assumption 3.2.1. *(i) $\Theta = \Theta_1 \times ... \times \Theta_k$ is compact.*

(ii) $\forall j = 1, ..., p$, $Em_j(X,.) : \Theta \to \mathbb{R}$ is continuous.

Lemma 3.2.1. *Under Assumption 3.2.1, the following statements are equivalent.*

(i) C_s is compatible.

(ii) $\Omega_s = \{\theta \in \Theta_{s_2} : EM_{s_1}(X,\theta) \geq 0\}$ is not empty.

(iii) For all positive definite V_0,

$$\inf_{\theta \in \Theta_{s_2}, \lambda \in [0,\infty)^m} (EM_{s_1}(X,\theta) - \lambda)^T V_0 (EM_{s_1}(X,\theta) - \lambda) = 0.$$

Let us partition the parameters θ and λ into "restricted" and "unrestricted" parts according to the biases of the selected and unselected moment functions. Formally, let

$$\lambda = EM(X,\theta)$$

where $M(X,\theta) = (m_1(X,\theta), ..., m_p(X,\theta))^T$, the vector of all the candidate moments, and $\theta = (\theta_1, ..., \theta_k)^T$, the vector of full parameters supported on $\Theta_1 \times ... \times \Theta_k$. Suppose a combination $C_s = (M_{s_1}, \Theta_{s_2})$ selects m moment conditions M_{s_1}, and leaves the rest of the moments (denoted by $M_{s_1}^c$ unused). It also selects a submodel parameterized by $\theta_s \in \Theta_{s_2}$, setting all the other components of θ, (which is denoted by θ_s^c) to be zero.

One can view model selection as placing a restriction on θ, while moment selection can be reviewed as placing a restriction on λ. Let λ_s be the subvector of λ corresponding to the selected moments. Let λ_s^c be the remaining components of λ corresponding to $M_{s_1}^c$. Then we have

$$EM_{s_1}(X, \theta_s) = \lambda_s, \lambda_s \geq 0$$

$$EM_{s_1}^c(X, \theta_s) = \lambda_s^c, \lambda_s^c \in \mathbb{R}^{p-m}$$

The bias λ_s for the selected moments is restricted to be nonnegative, while the bias λ_s^c for the unselected moments is left unrestricted. We thus have partitioned the moment functions into $M(X, \theta_s) = (M_{s_1}(X, \theta_s)^T, M_{s_1}^c(X, \theta_s)^T)^T$, and λ into $\lambda = (\lambda_s, \lambda_s^c)$. We put prior:

$$\begin{aligned} p(\lambda_s^c | C_s) &\sim N_{p-m}(0, \Sigma) \\ p(\lambda_s | C_s) &\sim Exp(\psi) \end{aligned} \tag{3.2}$$

where N_{p-m} denotes the $p - m$ dimensional multivariate normal distribution, assumed to be a priorly independent so that $\Sigma = diag\{\sigma_1^2, ...\sigma_{p-m}^2\}$. $Exp(\psi)$ is the exponential distribution with parameter ψ, as in Chapter 2.

We include both selected M_s and unselected M_s^c to construct the limited information likelihood, which depends only on the unrestricted θ_s since $\theta_s^c = 0$.

$$L(X^n | \theta_s, \lambda, C_s) = \frac{1}{\sqrt{\det(\frac{2\pi}{n} V)}} e^{-\frac{n}{2}(\bar{M}(\theta_s) - \lambda)^T V^{-1} (\bar{M}(\theta_s) - \lambda)} \tag{3.3}$$

where $\bar{M}(\theta_s) = \frac{1}{n}\sum_{i=1}^{n} M(X_i, \theta_s)$. The prior of C_s is imposed. Then the posterior of C_s can be obtained by integrating out θ_s and $\lambda = (\lambda_s^T, \lambda_s^{cT})^T$, which is proportional to the "integrated likelihood":

$$\begin{aligned} p(C_s|X^n) \quad \propto \quad & \iint_{\Theta_{s_2}\times[0,\infty)^m\times\mathbb{R}^{p-m}} L(X^n|\theta_s, \lambda, C_s)p(\theta_s|C_s)p(\lambda_s|C_s)p(\lambda_s^c|C_s) \\ & \cdot p(C_s)d\theta_s d\lambda_s d\lambda_s^c \end{aligned} \tag{3.4}$$

3.3. Posterior Consistency of Liao and Jiang (2010)

This chapter is actually written based on the materials of my published paper Liao and Jiang (2010, hereafter LJ), which considered the moment/model selection problem in interval censored regression problem. Due to the limited space, in this section I only briefly go over the main results, and details can be found in the published paper. The selection problems considered in LJ consist of two parts: selecting the compatible combinations of moment /model, and among the compatible combinations, selecting the "optimal" one. The optimal compatible combination is defined as the one with maximal $\dim(M_{s_1}) - \dim(\Theta_{s_2})$. This is because it is desirable that the optimal combination should contain as many moment inequalities as possible, since intuitively the more moment inequalities, the smaller the identified region, and hence the more information we have about the parameter. Meanwhile, it is required that the model should be as simple as possible, since simpler models are easier to interpret.

The selection procedure was known as the maximal posterior criterion (MPC), by maximizing the posterior of the combinations. In order for the MPC procedure to asymptotically select the optimal combination, the variance covariance matrix Σ in prior (3.2) should depend on the sample size:

$$\begin{aligned} &\sigma_i^2 = \sigma_n^2, \text{ where } \sigma_n^2 \to \infty \text{ but not exponentially fast} \\ &p(\theta_s | C_s) \sim N_t(0, n\sigma_n^2 I_t), \text{ where } t = \dim(\theta). \end{aligned} \tag{3.5}$$

Under further regularity assumptions (see LJ Assumptions 4.2-4.6), it can be shown that

Theorem 3.3.1 (LJ Theorem 4.3, MPC Consistency)**.** *Let*

$$C^* = \arg\max_{C_s} p(C_s | X^n)$$

where $p(C_s) > 0$ *does not depend on* n *for each* C_s*, and priors on specified by (3.2) and (3.5) for interval censored regression model, with probability approaching one,* $C^* = (M_{s_1}, \Theta_{s_2})$ *then is compatible and has the largest* $\dim(M_{s_1}) - \dim(\Theta_{s_2})$.

We can impose the following assumption which is similar to Assumption IDbc in Andrews and Lu (2001):

Assumption 3.3.1. *The true model and moment combination is the unique combination of* (M, Θ)*, such that it has the maximal* $\dim(M) - \dim(\Theta)$.

If this assumption holds, the previous theorem implies that by maximizing the combination posterior, we can asymptotically select the actual true combination of model and moments. In particular, when the dimension of the true parameter is fixed, and we are only selecting the corresponding moment inequalities that are satisfied by the true parameter of interest, Assumption 3.3.1 becomes: There exists a unique set of maximal number of moment inequalities that are satisfied by the true parameter. In this case, MPC consistently selects all the true moment inequalities.

3.4. More Reliable Setting

Note that Assumption 3.3.1 plays the central role of identifying the "true" moment inequalities and the parameter space. When the parameter space is fixed, meaning that the true parameter is assumed to exist, the selection problem then becomes to select the "true" moment inequalities that are satisfied by the true parameter. In this case, MPC procedure in LJ asymptotically selects the true moment inequalities.

However, in practice Assumption 3.3.1 is not satisfied naturally, and when it is not, MPC may select a set of incorrect moment inequalities with probability approaching one. The problem is that, the moment inequalities that are not satisfied by the true parameters can still be compatible.

Example 3.4.1. Suppose the true parameter $\theta_0 = 1.7$, with parameter space $\Theta = [0, 5]$. Consider the following moment inequalities

$$\theta \geq EY_1 (= 1.5) \tag{3.6}$$

$$\theta \leq EY_2 (= 2) \tag{3.7}$$

$$\theta \geq EY_3 (= 3) \tag{3.8}$$

$$\theta \geq EY_4 (= 3.5) \tag{3.9}$$

Apparently, only (3.6) and (3.7) are satisfied by θ_0, which correspond to interval $[1.5, 2]$. However, the MPC procedure will select all the other three inequalities (3.6), (3.8) and (3.9), because their combination has the maximal (3) number of inequalities, and the corresponding interval in Θ is $[3.5, 5]$. In this example, the compatible interval defined by the maximal number of inequalities does not contain the true parameter.

□

The following theorem shows that, when Assumption 3.3.1 is relaxed and priors (3.2) and $p(\theta_s)$ are data-independent, the posterior probability of incompatible combinations (where the corresponding identified region is empty) is still exponentially small, as opposed to compatible combinations, whose posterior is proportional to a positive constant multiplied by the combination prior.

Assumption 3.4.1. *For any compatible $C_s(M_{s_1}, \Theta_{s_2})$, $p(\theta|C_s)$ is uniformly bounded on Θ_{s_2}.*

Theorem 3.4.1. *Under Assumption 3.2.1, 3.4.1, the parameter priors are given in (3.2), and $V > 0, \Sigma > 0$ are fixed,*

(1) *If C_s is compatible and $p(C_s) > 0$, in probability*

$$\liminf_{n\to\infty} p(C_s|X^n) > 0$$

(2) *If C_s is not compatible, then for some $\alpha > 0$,*

$$p(C_s|X^n) = o_p(e^{-\alpha n})p(C_s)$$

The MPC procedure consistently selects the maximal $\dim(M) - \dim(\Theta)$, because the data-size-dependent priors (3.5) for unrestricted parameters (θ_s, λ_s^c), corresponding to unselected moments and selected parameters, have very thick tails asymptotically, which force the posterior of combinations with many unrestricted parameters to be very small. As illustrated in the previous example, however, the true parameter (if any) may satisfy only a few inequalities. Therefore a more reliable setting is to use data-independent prior for unrestricted parameters. If the prior of (θ_s, λ) is jointly specified as $p_{\theta,\lambda}(\theta_s, \lambda|C_s)$, which does not depend on the sample size n, the selection among compatible combinations using posterior probabilities is no longer consistent in terms of selecting the maximal $\dim(M_{s_1}) - \dim(\Theta_{s_2})$, because when Assumption 3.3.1 is relaxed, we fail to identify the true set of inequalities. Suppose $C_s(M_{s_1}, \Theta_{s_2})$ is a compatible combination. Write λ as the parameter that satisfies the moment condition $\lambda = EM(X, \theta)$, and λ is ordered and partitioned as (λ_s, λ_s^c), then λ takes its value in $\Lambda = [0, \infty)^m \times \mathbb{R}^{p-m}$, where m denotes the dimension of M_{s_1}, i.e., the number of selected moments. We impose the following regularity conditions on the parameter prior. Note that Condition (i) can be achieved if $p_{\theta,\lambda}(\theta, \lambda|C_s)$ is uniformly bounded by a constant $k > 0$ on $\Theta_{s_2} \times \Lambda$, given that Θ_{s_2} is bounded.

Assumption 3.4.2. *(i) For any C_s, and $\theta \in \Theta_{s_2}$, there exists $g(\theta) > 0$ satisfying $\int_{\Theta_{s_2}} g(\theta)d\theta < \infty$, such that $p_{\theta,\lambda}(\theta, \lambda|C_s) \leq g(\theta)$ for all $\lambda \in \Lambda$.*
(ii) For any fixed θ, $p_{\theta,\lambda}(\theta, \lambda|C_s)$ is continuous with respect to λ on Λ.

The following theorem shows that, in this case, the posterior heavily depends on the prior of combinations $p(C_s)$, which may be obtained by, if any, a priori information about some specific moment inequalities/ submodels.

Let $\Omega(\Theta, \Lambda) = \{\theta \in \Theta : EM(X, \theta) \in \Lambda\}$.

Theorem 3.4.2. *Under Assumptions 3.2.1 and 3.4.2, with fixed $V > 0$, in probability,*

$$(3.10) \qquad plim_{n\to\infty} p(C_s|X^n) = \frac{p(C_s)\int_{\Omega(\Theta,\Lambda)} p_{\theta,\lambda}(\theta_s, EM(X,\theta)|C_S)d\theta}{\sum_{C_s} p(C_s)\int_{\Omega(\Theta,\Lambda)} p_{\theta,\lambda}(\theta_s, EM(X,\theta)|C_S)d\theta}$$

We can see from this theorem, that asymptotically the posterior depends on $p(C_s|X^n)$, the combination's prior, and on $p_{\theta,\lambda}$, which is the prior distribution of (θ, λ) on the identified region. Therefore, the posterior is sensitive to the prior specification.

On the other hand, note that this is not a consistency result: Consider two compatible combinations C_1 and C_2 with the same parameter subspace but C_1 is nested with C_2 and contains more moment inequalities than C_2. Since C_1 has smaller identified region $\Omega(\Theta, \Lambda)$, by (3.10), it may have smaller posterior, even asymptotically. Therefore, compatible combinations with more inequalities do not necessarily have larger posteriors. This result is quite different from those in regular moment selection procedures with point identification (for example, in Andrews (1999)), which is reasonable, however, in moment inequalities problems, because of three reasons:

(1) First, as we have seen, the posterior is sensitive to the choice of priors. The penalty term against selecting fewer moment inequalities in the posterior is hidden in

$\int_{\Omega(\Theta,\Lambda)} p_{\theta,\lambda}(\theta_s, EM(X,\theta)|C_S)d\theta$, which does not involve the sample size.

(2) Second, unlike the moment selection problem with over-identification by moment equalities (Andrews 1999, Andrews and Lu 2001), in moment inequalities models, compatible combinations may not be correct, meaning that if the true parameter of interest is assumed to be fixed, some combinations may still be compatible even though they do not contain the true parameter. Therefore, the true inequalities are not necessarily the maximal set of compatible inequalities.

(3) Finally, by allowing the parameter space to change, we allow for the model uncertainty. In this case, it is reasonable for the result to heavily reply on the prior beliefs of the useful parameter components, and of the corresponding moment inequalities.

In some special cases, however, it is possible that the posterior favors the maximal number of inequalities. Consider the following example:

Example 3.4.2. Suppose the selected moment conditions satisfy $EM_s(X, \theta) = \lambda_s \in [0, M]^m$, and the unselected moment conditions satisfy $EM_s^c(X, \theta) = \lambda_s^c \in [-M, M]^{p-m}$, for some large constant $M > 0$. Hence $\Lambda = [0, M]^m \times [-M, M]^{p-m}$. Suppose $q(\lambda)$, the prior of λ, is uniformly distributed on $[-M, M]^p$, hence $q(\lambda) = (2M)^{-p} I(\lambda \in [-M, M]^p)$. Then

$$p_\lambda(\lambda|C_s) = \frac{q(\lambda)I(\lambda \in \Lambda)}{q(\lambda \in \Lambda)} = \frac{2^m I(\lambda \in \Lambda)}{(2M)^p}$$

In addition, suppose $p_{\theta,\lambda}(\theta, \lambda|C_s) = p_\theta(\theta|C_s)p_\lambda(\lambda|C_s)$. We have

$$\int_{\Omega(\Theta_s,\Lambda)} p_{\theta,\lambda}(\theta, EM(X, \theta)|C_s)d\theta = P(\theta \in \Omega(\Theta_s, \Lambda)|C_s)\frac{2^m}{(2M)^p}$$

Assume that $P(\theta \in \Omega(\Theta_s, \Lambda)|C_s) > 0$ if C_s is compatible. Hence 2^m is the reward of more moment inequalities. However, such a reward term does not depend on the sample size n. Q.E.D.

The moment equality condition case in the literature is significantly different than the problem considered here. For the sake of comparison, I briefly illustrate it here. Consider p-dimensional candidate moment equalities $EM(X, \theta) = (Em_1, ..., Em_p) = 0$. Suppose we select m moment conditions $EM_s = 0$, and partition the conditions into selected and unselected pair $M = (M_s, M_s^c)$. As before, we use the limited information likelihood to construct the posterior:

$$L(X^n|\lambda,\theta,C_s) = \det(2\pi V/n)^{-1/2}\exp\left(-\frac{n}{2}(\bar{M}_s(\theta)^T, \bar{M}_s^c(\theta)^T - \lambda^T)V^{-1}\begin{pmatrix}\bar{M}_s(\theta)\\ \bar{M}_s^c(\theta)-\lambda\end{pmatrix}\right)$$

Straightforward calculation yields that

(3.11)

$$p(C_s|X^n) \approx Const \times n^{\dim(M_s)/2}\int_\Theta p_{\theta,\lambda}(\theta, EM_s^c(\theta) + \Sigma_2^T\Sigma_3 EM_s(\theta))e^{-\frac{n}{2}EM_s(\theta)^T\Sigma_1 EM_s(\theta)}d\theta$$

where $V^{-1} = \begin{pmatrix}\Sigma_1 & \Sigma_3\\ \Sigma_3^T & \Sigma_2\end{pmatrix}$. When M_s is incompatible, meaning that $\{\theta \in \Theta : EM_s(X,\theta) = 0\}$ is empty, $e^{-\frac{n}{2}EM_s(\theta)^T\Sigma_1 EM_s(\theta)} < e^{-an}$ for some $a > 0$. But when $EM_s(X,\theta)$ over-identifies some element θ in Θ, the posterior then replies on $n^{\dim(M_s)/2}$, which is a penalty term that rewards the use of more moment conditions. Note that this penalty term depends on the sample size, hence is not sensitive to the prior specification. In addition, by applying the Laplace expansion to the integrand of the right hand side, the posterior criterion (3.11) can be shown to be equivalent to Andrews(1999)'s MSC.

CHAPTER 4

Bayesian Semi-Nonparametric Conditional Moment Restricted Models

4.1. Introduction

In this chapter, I consider conditional moment restricted model

$$E(\rho(Z, g_0)|W) = 0 \tag{4.1}$$

where (Z^T, W^T) is a vector of observable random variables, and W may or may not be included in Z. Here ρ is a residual function known up to g_0. The conditional expectation is taken with respect to the conditional distribution of Z given W, assumed unknown. The parameter of interest is g_0, which is infinite dimensional. Model (4.1) is a very general setting, which encompasses many important classes of nonparametric and semiparametric models. Recently, Chen and Pouzo (2009a) relaxed the compactness assumption on the parameter space as imposed in Ai and Chen (2003), and established the consistency and the convergence rate using the penalized sieve minimum distance estimator. In addition, Chen and Pouzo (2009b) considered the root-n efficient estimation of θ_0 as well as the asymptotic normality of the estimator. Note that one of the most important special cases of conditional moment restricted model is nonparametric instrumental variable regression (Example 1.3.2). See Examples 1.3.1-1.3.3 and Section 1.3 in Chapter 1 for corresponding literature.

In the existing literature, there are generally two ways of regularization to overcome the ill-posendess. One is to restrict g_0 to a compact space, and then minimize a consistent estimate of the minimum distance criterion over some finite dimensional compact sieve space; see, e.g., Newey and Powell (2003), Ai and Chen (2003), and Blundell, Chen and Kristensen (2007). The other way is to introduce a Tikhonov regularization tunning parameter, relaxing the compactness assumption. The procedure is then equivalent to minimizing a consistent penalized estimate of the minimum distance criterion over an infinite dimensional function space; see, e.g., Chen and Pouzo (2009a, 2009b), Hall and Horowitz (2005), Darolles et al (2010). Other related works on NPIV in the literature are: Blundell, Chen and Kristensen (2007), Chernozhukov, Gagliardini and Scaillet (2008), Horowitz and Lee (2007), Florens and Simoni (2009a), among others.

I will first focus on the general setting (4.1), following the regularization approach by Newey and Powell (2003) and Ai and Chen (2003), which assumes that the parameter space is compact. The conditional moment restriction is transformed into infinite number of unconditional moment restrictions as the first step. The problem then becomes the estimation under many moment conditions, which was studied by Han and Phillips (2006). After establishing the posterior consistency in the general conditional moment restricted model setting, we focus on the nonparametric instrumental variable regression model. As an alternative regularization approach, I will also establish the posterior consistency without the compactness assumption for nonparametric instrumental variable regression. To achieve the consistency, I propose a data-size dependent objective prior for the purpose of regularization, whose variance converges to zero. This technique is very common in the literature of Bayesian inverse problem and ridge regression. Recently, Florens and Simoni (2009a), have proposed a quasi-Bayesian approach to solve the ill-posed problem. They assumed a normal error term, and achieved consistency of the *regularized posterior distribution*, regularizing an operator that defines the posterior mean.

The contributions of this chapter are in at least five senses. First of all, I construct the posterior of the nonparametric structural function, which provides a nonparametric Bayesian interpretation of the estimation of the conditional moment restricted model. Second, g_0 is not assumed to be point identified, for the reasons to be explained in Section 4.2. Therefore the consistency of the posterior of g_0 means that, asymptotically, it converges into any small neighborhood of the identified region, which extends model (4.1) to the partial identification setup (Chernozhukov, Hong and Tamer 2007, and Santos 2007). Third, there is no need to assume any specific distribution on the data generating process. Instead, we use the limited information likelihood (Kim 2002) to construct the posterior distribution for g_0. The use of the limited information likelihood is similar to the Bayesian GMM (Yin 2009), which is more straightforward for models characterized by either moment conditions or estimating equations than the common methos using Dirichlet process priors in the nonparametric Bayesian literature. I show that by imposing only a few regularity conditions on the moment functions and priors, the posterior distribution achieves the desired frequentist properties in the large sample sense. Fourth, we extend the problem of GMM with many moment conditions in Han and Phillips (2006) to nonparametric models, allowing the dimension of the parameter to increase with sample size. Fifth, I study in detail the nonparametric IV regression model, and show that, by incorporating a regularized prior to deal with the ill-posedness, the posterior distribution of the sieve approximation can still be consistent even if the parameter space is relaxed to be noncompact. In addition to these contributions, I also allow for the heterogeneity of the residual term, meaning that $E[\rho(Z, g_0)^2|W = w]$ can depend on w.

The remainder of this chapter is organized as follows: Section 4.2 constructs the posterior distribution of the conditional moment restricted model, starting by transforming the conditional restriction into infinite number of unconditional restrictions, and then constructing the limited information likelihood for the posterior, followed by deriving its frequentist properties in large sample limit. Section 4.3 applies the consistency results to the single index model. Section 4.4 studies in detail the nonparametric instrumental variable regression model, relaxing the compactness assumption on the parameter space. Section 4.5 presents a simple Monte Carlo simulation result. Finally Section 4.6 concludes with further discussions. Proofs are given in the appendix.

4.2. Conditional Moment Restricted Model

4.2.1. Limited Information Likelihood and Identification Functional

Consider a conditional moment condition

$$E[\rho(Z, g_0)|W] = 0 \tag{4.2}$$

where g_0 is the true nonparametric structural function, and is assumed to be inside some space of continuous functions Θ. For simplicity, throughout the paper, let us consider the case $W \in \mathbb{R}$, which is supported on a compact set $\mathcal{W}$. The results can be naturally generalized to multi-dimensional cases.

Following the setting of Ai and Chen (2003), let us approximate Θ by a sieve space Θ_q, which is a finite-dimensional compact parameter space spanned by sieve basis functions$\{\phi_1, ..., \phi_q\}$ such as splines, power series, wavelets or Fourier series, with $q \to \infty$ as $n \to \infty$, such that g_0 can be approximated arbitrarily well by $g_q = \sum_{i=1}^{q} b_i \phi_i$ for some coefficients $\{b_i : i = 1, ..., q\}$. Hence, instead of g_0, we construct the posterior of g_q, and show that $\|g_q - g_0\|_H \to 0$ in the posterior probability under some norm $\|.\|_H$.

As the first step, we need to transform the conditional moment restriction into unconditional moment restrictions. Suppose $\mathcal{W} = [a, b]$, with $a < b$. Let $\cup_{i=1}^{k_n} R_i^n$ be a partition of $\mathcal{W}$, where

$$R_j^n = \left[a + \frac{j-1}{k_n}(b-a), a + \frac{j}{k_n}(b-a)\right], j = 1, ..., k_n. \tag{4.3}$$

We allow $k_n \to \infty$ as $n \to \infty$. Let $X = (Z, W)$. For each j, define $m_{nj}(g, X) = \rho(Z, g)1_{(W \in R_j^n)}$, and $m_n(g, X) = (m_{n1}(g, X), ..., m_{nk_n}(g, X))^T$, which is $k_n \times 1$ vector. Then equation (4.2) implies

$$Em_n(g_0, X) = 0 \tag{4.4}$$

where the expectation is taken with respect to the joint distribution of $X = (Z, W)$. Note that $m_n(g, X)$ is $k_n \times 1$, where k_n increases as n increases to infinity. Hence (4.2) implies many moment conditions with the number of moments increasing to infinity. It is straightforward to verify that

$$V_0 \equiv Var(m_n(g_0, X)) = \begin{pmatrix} E(\rho(Z, g_0)^2 1_{W \in R_1^n}) & & 0 \\ & \ddots & \\ 0 & & E(\rho(Z, g_0)^2 1_{W \in R_{k_n}^n}) \end{pmatrix}$$

For each $g \in \Theta$, and $j = 1, ..., k_n$, write $\bar{m}_{nj}(g) = \frac{1}{n}\sum_{i=1}^{n} m_{nj}(g, X_i)$ and $\bar{m}_n(g) = (\bar{m}_{n1}(g), ..., \bar{m}_{nk_n}(g))^T$. Under some regularity conditions, for each fixed k, $\bar{m}_n(g_0)$ would satisfy the central limit theorem: for any $\alpha \in \mathbb{R}^k$, as n goes to infinity,

$$\left| P(\sqrt{n} V_0^{-1/2} \bar{m}_n(g_0) \leq \alpha) - \prod_{i=1}^{k} \Phi(\alpha_i) \right| \to 0 \tag{4.5}$$

where $\Phi(.)$ denotes the cumulative distribution function of standard normal.

We now construct the the posterior distribution of g_0. Note that the true parameter space Θ is infinite-dimensional, we thus need to parameterize g_0 and approximate it on the finite dimensional sieve space Θ_q. Therefore, instead of g_0, we construct the posterior for the approximating parameter of g_0 inside Θ_q. The asymptotic result (4.5) motivates a likelihood function on the sieve space Θ_q:

$$L(g_q) \propto \exp\left(-\frac{n}{2}\bar{m}_n(g_q)^T V_0^{-1}\bar{m}_n(g_q)\right)$$

According to Kim (2002), the function $L(g_q)$ can be more appropriately interpreted as the best approximation to the true likelihood function under the conditional moment restriction, by minimizing the Kullback-Leibler divergence, which is known as the *limited information likelihood*.

The right hand side of the likelihood function involves

$$\bar{G}(g_q) \equiv \bar{m}_n(g_q)^T V_0^{-1}\bar{m}_n(g_q) \tag{4.6}$$

Hence it is important to study the asymptotic property of $\bar{G}$ first. Define

$$G_{k_n}(g) \equiv Em_n(g, Z)^T V_0^{-1} Em_n(g, Z) \tag{4.7}$$

for all $g \in \Theta_q$. Using a similar argument of Han and Philips (2006), under some regularity conditions, we will show that

$$\sup_{g\in\Theta_q} |\bar{G}(g) - G_{k_n}(g)| \to^p 0 \tag{4.8}$$

Note that G_{k_n} still depends on the sample size n. It can be shown that it uniformly converges to some functional $G(g)$ over $g \in \Theta$, where

$$G(g) = \int_a^b \frac{[E(\rho(Z,g)|W=w)]^2}{E(\rho(Z,g_0)^2|W=w)} dF_W(w). \tag{4.9}$$

We call $G(g)$ the *identification functional* since the identification of g_0 is characterized by minimizing G. To be specific, define the *identified region* for g_0:

$$\Theta_I = \{g \in \Theta : E(\rho(Z,g)|W) = 0 \text{ for almost all } w \in [0,1]\},$$

which is assumed to be nonempty, then $\Theta_I = \arg\min_{g\in\Theta} G(g)$. If Θ_I is a singleton, then $\Theta_I = \{g_0\}$. Otherwise g_0 is *partially identified* on Θ_I (See, e.g. Santos 2007).

Throughout this section, we do not assume Θ_I is necessarily a singleton; therefore we allow g_0 to be only partially identified by the conditional moment restriction (4.2), for the following two reasons. First, when the conditional moment restriction is given by the nonparametric instrumental variable regression, the identification of g_0 depends on the completeness of the conditional distribution of $X|W$; however, the completeness assumption is hard to verify if the conditional distribution of $X|W$ does not belong to the exponential family. Severini and Tripathi (2006) explored identification issues with these models and note that point wise identification can easily fail (See Example 1.4.1 below). Another reason is that, sometimes instead of g_0 itself, we are only interested in a particular characteristic of it, say its linear functional $h(g_0)$. For example, in the nonparametric IV regression, if $g_0(x)$ is the inverse demand function, then its consumer surplus at some level x^* can be written as a functional $h(g_0) = \int_0^{x^*} g_0(x)dx - g_0(x^*)x^*$. In this case, the identification of g_0 might not be necessary (see example 1.4.1).

Severini and Tripathi (2006) have shown that without assuming g_0 to be identified, it is still possible to point identify its functional $h(g_0)$.

4.2.2. Uniform Convergence to the Identification Functional

We give the assumptions for the uniform convergence of $\bar{G}$ and G_{k_n}, as well as the posterior consistency.

Assumption 4.2.1. *(i) The data $X^n = (X_1, ..., X_n)$ are independent and identically distributed.*
(ii) There is a metric $\|.\|_H$ such that the parameter space Θ is compact under $\|.\|_H$.
(iii) The support of X (denoted by $\mathcal{X}$) is compact.

Condition (i) assumes the data are independent and identically distributed. Condition (ii) restricts the parameter space as well as the choice of the metric $\|.\|_H$. The compactness is a commonly imposed condition in the nonparametric and semiparametric statistical and econometric literature, and is satisfied when the infinite-dimensional parameter space consists of bounded and smooth functions (Gallant and Nychka 1987). In the nonparametric instrumental variable regression model, the compactness of the parameter space is a way of "regularization" to deal with the "ill-posed" problem (See Newey and Powell 2003). In this section, we impose prior condition, equicontinuity and the sieve approximation assumptions to establish the posterior consistency based on the norm $\|.\|_H$. When endowed with some specific norms, $(\Theta, \|.\|_H)$ becomes a Banach space. Specification of $\|.\|_H$ as well as the parameter space are provided in Sections 4.3 and 4.4, where we apply our results to nonparametric IV regression and the single index model. Condition (iii) requires that the support of the data be compact.

Assumption 4.2.2. *(i) For all $j = 1, 2, ..., k_n$, $P(W \in R_j^n) = O(k_n^{-1})$, where R_j^n is define by (4.3). (ii) $k_n = o(n^{2/5})$.*

Condition (i) is satisfied if W has a continuous density function. Condition (ii) requires that the number of moment conditions in $m_n(g, X)$ should not grow too fast, which is needed for the pointwise convergence of $|\bar{m}_n(g)^T V_0^{-1} \bar{m}_n(g) - Em_n(g, X)^T V_0^{-1} Em_n(g, X)|$. This condition is usually imposed in the literature of many moment condition problems (See Han and Phillips 2006).

Define $K_g(w) = E(\rho(Z, g)|W = w)$, $w \in [a, b]$, and denote $\mathcal{Z}$ as the support of Z.

Assumption 4.2.3. *(i) $E(\rho(Z, g_0)^2|W = w)$ is continuous and bounded away from zero on $w \in [a, b]$.*
(ii) $\{K_g(.) : g \in \Theta\}$ is equicontinuous on $[a, b]$.
(iii) $\rho(z, .)$ satisfies: for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\sup_{z \in \mathcal{Z}} \sup_{||g_1 - g_2||_H < \delta} |\rho(z, g_1) - \rho(z, g_2)| < \epsilon$$

Assumption 4.2.3 is used for the uniform convergence of G_{k_n} to the identification functional G over all $g \in \Theta$. It imposes restrictions on the conditional second moment of the true residual function. Since W is supported on a compact set, condition (i) also implies the uniform continuity of $E(\rho(Z, g_0)^2|W = w)$. In addition, since $E(\rho(Z, g_0)^2|W = w)$ depends on w, the residual heterogeneity is allowed. This assumption also implies that $\sup_{(g,w) \in \Theta \times [a,b]} |E(\rho(Z, g)|W)|$ is bounded. Condition (iii) guarantees that $G(g)$ is continuous on $(\Theta, \|.\|_H)$.

The following assumptions are needed for the uniform convergence of $|\bar{G} - G_{k_n}|$. Let $\lambda_{\max}(V)$ denote the largest eigenvalue of matrix V.

Assumption 4.2.4. *The sequence of random process* $\max_{1\leq j\leq k_n} \sqrt{n}|\bar{m}_{nj}(g) - Em_{nj}(g, X)|$ *is stochastic equicontinuous.*

This condition requires that the centered and rescaled moment functions should be uniformly continuous over Θ and n. The definition of *stochastic equicontinuity* can be found, for instance, in Newey and McFadden (1994), which is commonly assumed in the probability convergence theory and econometrics literature for the uniform convergence of stochastic functions. It will be shown in Section 4.3 and 4.4 that Assumption 4.2.3 and 4.2.4 are satisfied by the single index model and nonparametric instrumental variable regression model, with mild assumptions on the data generating process.

Assumption 4.2.5. *(i)* $\sup_{g_q\in\Theta_q} \lambda_{\max}(Var(m_n(g_q, X))) = O(k_n)$.
(ii) For all $g_{q_n} \in \Theta_{q_n}$, *for all* $j = 1, 2, ..., k_n$, *for all* $k_n, q_n \leq n$ *and* n, $E[m_{nj}(g, X) - E(m_{nj}(g, X))]^4 \leq B < \infty$.

Condition (i) and (ii) require that the fourth moments of $m_n(g, X) - E(m_n(g, X))$ exist and that the second moment matrix has eigenvalues no larger than $O(k_n)$. In Han and Phillips (2006), it was assumed that the eigenvalues of $Var(m_n(g, X))$ are bounded by a universal constant uniformly over g and n, and a sufficient condition for their assumption was provided, which assumed that the covariance structure is dominated:

$$\sup_{g\in\Theta} Var(m_n(g, X)) \leq aI_{k_n} + b_{k_n}b_{k_n}^T$$

where a is some large enough constant and b^{k_n} is a k_n dimensional vector such that its elements satisfy $\lim_{n\to\infty}\sum_{i=1}^{k_n} b_i^2 < \infty$ (See Han and Phillips 2006, Assumption 1). Here, since the number of moment conditions k_n grows with n, with a nonparametric structural function, it is more reasonable for the eigenvalues also increasing with n. Therefore we relax this assumption and allow the eigenvalues of $Var(m_n(g, X))$ to increase with the same rate of the number of moment conditions. The payoff would be a slower rate of k_n going to infinity.

Under these assumptions, we can show that $\bar{G}$, the power term of the limited information likelihood, converges uniformly to the identification functional.

Theorem 4.2.1. *(i) Under Assumptions 4.2.1-4.2.3, for G_{k_n} and G defined by (4.7) and (4.9) respectively,*

$$\sup_{g \in \Theta} |G_{k_n}(g) - G(g)| \to^p 0$$

(ii) Under Assumption 4.2.1-4.2.5, for $\bar{G}$ defined by (4.6), in probability

$$\sup_{g \in \Theta_{qn}} |\bar{G}(g) - G_{k_n}(g)| \to^p 0$$

4.2.3. Posterior Consistency

We use the limited information likelihood described in Section 4.2.1 as the likelihood function: for all $g_q = \sum_{i=1}^{q} b_i \phi_i \in \Theta_q$, $L(g_q) \propto \exp\left(-\frac{n}{2}\bar{m}_n(g_q)^T V_0^{-1} \bar{m}_n(g_q)\right)$. Let $p(g_q)$ be a prior distribution of the sieve approximation of g_0. Then Bayesian rule implies:

$$p(g_q|X^n) \propto p(g_q)L(g_q)$$

For any event A that is measurable with respect to the posterior distribution of, its posterior distribution is given by

$$P(A|X^n) = \int_A p(g_q|X^n)db$$

A straightforward application of Jiang and Tanner (2008, Proposition 6) renders that for any $\delta > 0$,

$$\begin{aligned} E\{P(G(g_q) - \inf_{g\in\Theta_q} G(g) > 5\delta | X^n)\} \leq\ & P(\sup_{g\in\Theta_q} |\bar{G}(g) - G(g)| \geq \delta) \\ & + \frac{e^{-2n\delta}}{P(G(g_q) - \inf_{g\in\Theta_q} G(g) < \delta)} \end{aligned} \tag{4.10}$$

By Theorem 4.2.1, $\sup_{g\in\Theta_q} |\bar{G}(g) - G(g)| \to^p 0$ in the probability distribution of X^n as $n \to \infty$. Hence $G(g_q) - \inf_{g\in\Theta_q} G(g) \to 0$ in the posterior probability of $g_q | X^n$ given that the prior probability $P(G(g_q) - \inf_{g\in\Theta_q} G(g) < \delta)$ is bounded away from zero. This requires a regularity condition on the prior.

To proceed, we need to introduce some additional notation. Given the metric structure of $(\Theta, \|.\|_H)$, for a set $A \subset \Theta$, define $d(g, A) = \inf_{a\in A} \|g - a\|_H$. For any $\delta > 0$, let $\Theta_I^\delta = \{g \in \Theta : d(g, \Theta_I) < \delta\}$, the δ- expansion of the identified region of g_0. If g_0 is point identified ($\Theta_I = \{g_0\}$), Θ_I^δ is an open ball centered at g_0 with radius δ. In addition, for two sequences a_n and b_n, write $a_n \succ b_n$ if $\frac{a_n}{b_n} \to \infty$ as $n \to \infty$.

Assumption 4.2.6. *(i) For any $\delta > 0$ there exists $c > 0$, such that fall all large enough $q = q_n$, $P(g_q \in \Theta_I^\delta) \succ e^{-cq_n}$.*
(ii) $\frac{q_n}{n} \to 0$

Condition (i) means that the prior of g_q cannot be exponentially small on the neighborhood of Θ_I. Condition (ii) imposes a restriction on q_n, the number of terms in the sieve approximation. Recall that in Section 4.2.1 we have established that $\Theta_I = \arg\min_{g\in\Theta} G(g)$. It will also be shown in the Appendix that $G : \Theta \to \mathbb{R}$ is continuous. In addition, the sieve space

Θ_q approximates Θ arbitrarily well. Therefore this assumption implies that for any $\delta > 0$, the second term of the right hand side of (4.10) is negligible.

Assumption 4.2.7. *For each $g \in \Theta$, there exists $g_q \in \Theta_q$ such that $\|g - g_q\|_H = o(1)$.*

This assumption is simply the definition of a sieve space. It is satisfied by the spaces that are spanned by commonly used sieve basis functions such as splines, power series, wavelets and Fourier series. We will specify the norm $\|.\|_H$ in the subsequent sections.

We then have the posterior consistency for the estimation of g_0:

Theorem 4.2.2 (Posterior Consistency)**.** *Under Assumptions 4.2.1-4.2.7, for any $\delta > 0$, in probability,*

$$P(g_q \in \Theta_I^\delta | X^n) \to^p 1.$$

In particular, if g_0 is point identified, then in the probability of X^n,

$$P(\|g_q - g_0\|_H < \delta | X^n) \to^p 1.$$

Let $h(g_0)$ be a linear functional of g_0, whose practical meaning may be of interest in many applications. For example, if $h(g_0) = E[g_0(X)\omega(X)]$ for some weight function ω, then with proper choices of ω, h can be used to test some special properties of g_0 such as monotonicity, convexity, etc. On the other hand, h itself may have interesting meanings. For example, when g_0 denotes the inverse demand function in nonparametric regression, $h(g_0)$ can be the consumer surplus (See Santos 2007). Severini and Tripathi (2006) have provided conditions to point identify $h_{(}g_0)$ even if g_0 itself is not identified.

Example 4.2.1. In the application of economics, let g_0 be the inverse demand function, and we are usually interested in the change of consumer surplus $\int_0^{x^*} g_0(x)dx - g_0(x^*)x^*$, with some observable x^*. Suppose the functional of interest is $h(g_0) = \int_0^{x^*} g_0(x)dx = \int g_0(x)v(x)dx$, where $v(x) = I(0 < x < x^*)$. Assume that (X, W) are supported on $[0,1]^2$, with joint density function $f_{XW}(x,w) = 3|x - w|$. By Severini and Tripathi (2006), $h(g_0)$ is point identified if there exists $p(w) \in L^2(W)$ such that $\int p(w)f_{XW}(w,x)dw = v(x)$ for almost all $x \in [0,1]$. In fact, let ϕ' denote the derivative of the standard normal density function, and denote $p_t(w) = \frac{1}{t^2}\left(\phi'(\frac{w-x^*}{t}) - \phi'(\frac{w}{t})\right)$, it can be shown that (see Polyanin and Manzhirov (1998) and Santos (2008a)) $\lim_{t\to 0}\int_0^1 p_t(w)f_{XW}(x,w)dw = v(x)$ for almost all $x \in [0,1]$. Therefore, $h(g_0)$ is point identified.

Example 4.2.2. Suppose we want to test that the unknown function g_0 is weakly increasing. Note that any weakly increasing function $g(x)$ must satisfy $\int_{-\pi}^{\pi}\sin(x)g(x)dx \geq 0$. Hence the functional of interest here is $h(g_0) = \int_{-\pi}^{\pi}\sin(x)g_0(x)dx$. Suppose the joint distribution of (X, W) is absolutely continuous, with density function $f_{XW}(x,w)$. By Severini and Tripathi (2006), $h(g_0)$ is point identified, if there exists $p(w) \in L^2(W)$ such that $\int p(w)f_{X,W}(w,x)dw = \sin(x)$ for almost all x on its support.

Theorem 4.2.2 implies a flexible way to consistently estimate h in a Bayesian approach, without identifying g_0. In the following assumption, condition (i) assumes the point identification of $h(g_0)$. A sufficient and necessary condition can be found in Severini amd Tripathi (2006). Condition (ii) requires the continuity of h, which is satisfied when $h(g_0) = E[g_0(X)\omega(X)]$ if $E|\omega(X)| < \infty$.

Assumption 4.2.8. *(i)* $\{h(g) : g \in \Theta_I\} = \{h(g_0)\}$. *(ii)* $h : (\Theta, \|.\|_H) \to \mathbb{R}$ *is continuous.*

Corollary 4.2.1. *When g_0 is not necessarily point identified, under Assumptions 4.2.1-4.2.7, for any $\delta > 0$, in probability,*

$$P(|h(g_q) - h(g_0)| < \delta | X^n) \rightarrow^p 1.$$

4.2.4. Bayesian Implementation

Note that the likelihood $L(g_q)$ is not feasible because V_0 is unknown. Therefore we can estimate V_0 by $\hat{V}$ if g_0 is identified, where

$$\hat{V} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \rho(Z_i, \hat{g})^2 1_{W_i \in R_1^n} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n}\sum_{i=1}^n \rho(Z_i, \hat{g})^2 1_{W_i \in R_k^n} \end{pmatrix}$$

and $\hat{g}$ is a “preliminary” estimator of g_0, which can be, for example, the sieve minimum distance estimator (SMD) in Ai and Chen (2003). If g_0 is not identifiable, meaning that Θ_I is not a singleton, we suggest use $k_n I$ to replace V_0^{-1}, where I is the identity matrix. Because under very weak assumptions, each diagonal element of V_0 is of order $O(1/k_n)$, hence using $k_n I$ will not affect the consistency result, whether g_0 is identified or not.

Given the basis functions $\{\phi_1, ..., \phi_q\}$, the nonparametric function can be represented by a linear combination $g_q(t) = \sum_{i=1}^q b_i \phi_i(t)$, where ϕ_i is the ith basis function which can be power series wavelets, or Fourier series. One can treat the basis coefficients $\{b_1, ..., b_q\}$ to be independent parameters, each with prior $\pi(b_j) \sim N(0, j^{-\alpha})$ for some $\alpha > 0$, where the variance is chosen such that the higher order terms has smaller variation around zero, which overcomes the potential over-fitting problem. The posterior of the sieve approximation is then given by

$$p(g_q|X^n) \propto \prod_{j=1}^{q} \pi(b_j) \exp[-\frac{n}{2}\bar{G}(g_q)]$$

One can then simulate $\{b_j\}$ from the posterior through the MCMC procedure.

4.3. Application: Single Index Model

In the single-index model,

$$\rho(Z, g_0) = Y - h_0(W\theta_0)$$

where $g_0 = (h_0, \theta_0)$. For this specific application, the parameter of interest consists of both an infinite-dimensional parameter $h \in \mathcal{H}$, and a finite dimensional parameter $\theta \in \Omega$. The parameter space is written as $\Theta = \mathcal{H} \times \Omega$. We can approximate Θ by sieve space $\Theta_q = \mathcal{H}_q \times \Omega$, where $\mathcal{H}_q$ is the sieve space approximation to $\mathcal{H}$, as $q = q_n \to \infty$. Let $\{\phi_1, ..., \phi_q\}$ be an orthonormal basis for $\mathcal{H}_q$ such that g_0 can be approximated arbitrarily well by $g_q = \sum_{i=1}^{q} b_i\phi_i$ for some coefficients $\{b_i : i = 1, ..., q\}$. Then the sieve approximation to g_0 can be written as $(\sum_{i=1}^{q} b_i\phi_i, \theta_0) \in \Theta_q$.

Define $\mathcal{M} = \{w\theta : w \in [a, b], \theta \in \Omega\} \subset \mathbb{R}$. We use the Euclidean norm for θ, and introduce the Hölder norm for h. Define

$$\|h\|_s = \sup_{t\in\mathcal{M}} |h(t)| + \sup_{t_1 \neq t_2} \frac{|h(t_1) - h(t_2)|}{|t_1 - t_2|} \tag{4.11}$$

Let $\mathcal{H} = \{h : \|h\|_s < B\}$ for some known, large positive constant B. Here $\mathcal{H}$ is a Hölder ball of order one, a space of functions $h : \mathcal{H} \to \mathcal{R}$ such that the first derivative is bounded. It is known that power series, splines, and Fourier series all can approximate functions in the Hölder ball well. For $g = (h, \theta) \in \mathcal{H} \times \Omega = \Theta$, define

$$\|g\|_H = |\theta| + \|h\|_s \tag{4.12}$$

Note that Ai and Chen (2003) used a Hölder norm with higher orders, and the norm in Newey and Powell (2003) can be simplified to higher order Hölder norm if the data Z is supported on a compact set.

We verify Assumption 4.2.2 by imposing more general conditions on the distribution of data generating process.

Assumption 4.3.1. *$g_0 = (h_0, \theta_0) \in \Theta = \mathcal{H} \times \Omega$, with norm $\|.\|_H$ which satisfies:*
(i) $\mathcal{H} = \{h : \|h\|_s \leq B\}$, where $\|.\|_s$ is defined as (4.11).
(ii) Ω is compact.
(iii) $\|.\|_H$ is defined as (4.12).

Assumption 4.3.2. *The conditional distribution of $W|Y$ and the marginal distribution of W have continuous density function $f_{W|Y}(w|y)$ and $f_W(w)$ on $[a, b]$ respectively, which satisfy:*
(i) $f_W(w)$ is bounded away from zero on $[a, b]$.
(ii) For any $\delta > 0$, there exists $d > 0$ such that

$$\sup_y \sup_{|w_1 - w_2| < d} |f_{W|Y}(w_1|y) - f_{W|Y}(w_2|y)| < \delta$$

Proposition 4.3.1. *Let $\rho(Z, g) = Y - h(W\theta)$, where $EY^2 < \infty$ and $\sup_{(h,\theta)\in\Theta} Eh(W\theta)^2 < \infty$, then Assumption 4.3.1 and 4.3.2 imply Assumptions 4.2.3 and 4.2.4.*

4.4. Nonparametric Instrumental Variable Regression

The nonparametric instrumental variable regression model is given by

$$Y = g_0(X) + \epsilon$$

where X is endogenous, which is correlated with ϵ. The parameter of interest is g_0, which is the nonparametric structural function. In addition, suppose we observe an instrumental variable $W \in [a, b]$ a.s., such that $E(\epsilon|W) = 0$. The nonparametric IV model is thus essentially a conditional moment restriction $E(Y|W) = E(g_0(X)|W)$. Let $Z = (Y, X)$, then $\rho(Z, g) = Y - g(X)$.

Define $T : \Theta \to L^2(W)$, such that $T(g) = E(g(X)|W)$, and $E(Y|W = w) \equiv \mu(w)$, then

$$Tg_0 = \mu \tag{4.13}$$

The inference on g_0 is difficult. The first difficulty comes from the identification, which depends on the invertibility of T. If T is nonsingular, in which case it has no zero eigenvalue, g_0 can be point identified by $g_0 = T^{-1}\mu$. Newey and Powell (2003) characterize the identification of g_0 in terms of the completeness of the conditional distribution of X given W. However, if the distribution of $X|W$ is not assumed to be parametric, neither the invertibility of T nor the completeness is easy to verify. See Severini and Tripathi (2006) for a detailed description of the identification issue of g_0.

Even when g_0 is identified, the second difficulty arises in estimation. As pointed out by Newey and Powell (2003) and Hall and Horowitz (2005), there is an ill-posed problem. Note that (4.13) is a Fredholm integral equation of the first kind. Since T^{-1} is not bounded, it is not continuous. Therefore, small inaccuracy in the estimation of μ can lead to large inaccuracy in the estimation of g_0, which is known as the *ill-posed problem* (Kress 1999). In the existing literature, there are generally two ways of regularization to overcome the ill-posendess. One is to restrict g_0 to a compact space, and then minimize a consistent estimate of the minimum distance criterion over some finite dimensional compact sieve space; see, e.g., Newey and Powell (2003), Ai and Chen (2003), and Blundell, Chen and Kristensen (2007). The other way is to introduce a Tikhonov regularization tunning parameter, relaxing the compactness assumption. The procedure is then equivalent to minimizing a consistent penalized estimate of the minimum distance criterion over an infinite dimensional function space; see, e.g., Chen and Pouzo (2009a, 2009b), Hall and Horowitz (2005), Darolles et al (2010), and references therein. Recently, Florens and Simoni (2009a) proposed a quasi-Bayesian approach, which regularizes an operator that defines the posterior mean of g_0, assuming a normal error term in the regression.

In this section, we assume g_0 be point identified by (4.13), and focus on the posterior distribution of g_0. We will show the posterior consistency in two approaches. The first is the natural application of the general consistency result established in Section 4.2, which assumes that g_0 lies in a known compact parameter space. Alternatively, we will relax the compactness assumption, and impose a Tikhonov regularized prior instead.

4.4.1. Case with Compactness

This approach is similar to Newey and Powell (2003) and Santos (2007), focusing on the case where g_0 is known to belong to a compact set. The posterior distribution of g is restricted to this set. This approach eliminates the ill-posed problem essentially because the inverse of an integration operator is continuous on a compact set.

Let $\cup_{j=1}^{k_n} R_i^n$ be the partition of $[a, b]$ defined in Section 2.1. Under this particular model setting, $m_{nj}(g, X) = (Y - g(X))1_{W \in R_j^n}$ for each $j = 1, ..., k_n$. Let $\bar{m}_{nj}(g)$ be the sample analog of $Em_{nj}(g, X)$, and Θ_q be the sieve space approximation to Θ. We then have the limited information likelihood:

$$L(g_q) \quad \propto \quad \exp\left(-\frac{n}{2}\bar{m}_n(g_q)^T V_0^{-1} \bar{m}_n(g_q)\right)$$

and the identification functional $G : \Theta \to \mathbb{R}$:

$$G(g) = \int_a^b \frac{[E(Y - g(X)|W = w)]^2}{E(\epsilon^2|W = w)} dF_W(w)$$

A sufficient and necessary condition for point identification of g_0 is that $G(g)$ is minimized uniquely at g_0 on Θ.

When the parameter space is compact, we assume the metric $\|.\|_H$ to be:

$$\|g\|_H = \sup_x |g(x)| + \sup_{x_1 \neq x_2} \frac{|g(x_1) - g(x_2)|}{|x_1 - x_2|} \tag{4.14}$$

The parameter space $\Theta = \{g : \|g\|_H \leq B\}$ for some known, large positive constant B. The compactness of Θ under $||.||_H$ was shown by Gallant and Nychka (1987). We verify that in this model, Assumptions 4.2.3 and 4.2.4 are satisfied with a more general assumptions on the distribution of data. Let $\mathcal{Z}$ be the support of (X, Y).

Assumption 4.4.1. *$\|.\|_H$ is defined as (4.14), and $\Theta = \{g : \|g\|_H \leq B\}$.*

Assumption 4.4.2. *The conditional distribution of $W|X, Y$ and the marginal distribution of W have continuous density functions $f_{W|X,Y}(w|x, y)$ and $f_W(w)$ on $w \in [a, b]$ respectively, which satisfy:*

(i) $f_W(w)$ *is bounded away from zero on* $[a, b]$*, and*

(ii) $\forall\epsilon > 0$*, there exists* $\delta > 0$*, such that*

$$\sup_{(x,y)\in\mathcal{Z}} \sup_{|w_1-w_2|<\delta} |f_{W|X,Y}(w_1|x, y) - f_{W|X,Y}(w_2|x, y)| < \epsilon$$

This assumption provides sufficient conditions for the continuity of $E(\epsilon^2|W = w)$ on $[a, b]$ and the equicontinuity of $\{K_g(.) : g \in \Theta\}$. In fact we have the following proposition:

Proposition 4.4.1. *Let* $\rho(Z, g) = Y - g(X)$*, where* $EY^2 < \infty$ *and* $\sup_{g\in\Theta} Eg(X)^2 < \infty$*, then Assumptions 4.4.1 and 4.4.2 imply Assumptions 4.2.3 and 4.2.4.*

The posterior consistency for the nonparametric IV regression model then follows immediately from Theorem 4.2.2, which is stated as a corollary here.

Corollary 4.4.1. *Assume that* g_0 *is point identified. Under Assumptions 4.2.1, 4.2.2, 4.2.5, 4.4.1 and 4.4.2 , for any* $\delta > 0$*, in the probability of* X^n*,*

$$P(\|g_q - g_0\|_H) < \delta|X^n) \to^p 1$$

4.4.2. Relaxing the Compactness

In this subsection, we relax the compactness assumption on the parameter space, and assume $\Theta = L^2(X)$. As has been discussed earlier, in order to achieve the posterior consistency, additional regularization procedure is needed to overcome the ill-posedness. For this purpose, for the sieve approximation of g_q, we use a regularized prior:

$$p(g_q) \propto e^{-na_n^2||g_q||^2} \qquad (4.15)$$

where $\{a_n\}_{n=1}^{\infty}$ is a sequence converging to zero as n increases, but not too fast (we require $na_n^2 \to \infty$). From the frequentist point of view, prior (4.15) is similar to the penalty term in Chen and Pouzo (2009a, 2009b). Also, the regularization scheme through a penalty term on the prior is commonly used in the Bayesian literature. For example, in the parametric case, the penalty term in ridge regression can be viewed as a regularized prior from the Bayesian perspective; see, e.g., Haitovsky and Wax (1980). In the nonparametric case, recently Florens and Simoni (2009b) specified a regularized prior distribution that is an extension of Zeller's g-prior for the regularization.

We only consider the bivariate case of (X, W), and assume both X and W are supported on $[a, b]$ without loss of generality. We still employ the limited information likelihood defined in Section 2:

$$L(g_q) \propto e^{-\frac{n}{2}\bar{m}_n(g_q)^T V_0^{-1}\bar{m}_n(g_q)} \tag{4.16}$$

where

$$\bar{m}_n(g) = \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))1_{(W_i\in R_1^n)}, ..., \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))1_{(W_i\in R_k^n)}\right)'$$

The limited information likelihood is the best approximation to the true likelihood under the moment conditions 4.2.3. Therefore the posterior distribution based on $L(g)$ has an asymptotic likelihood interpretation.

As in the compactness case, the large sample behavior of $L(g)$ depends on the limit of $\bar{m}_n(g)^T V_0^{-1}\bar{m}_n(g)$, which is $G(g) = \int_a^b \frac{[E(Y-g(X)|W=w)]^2}{E(\epsilon^2|W=w)} dF_W(w)$. By the earlier definition, $T(g)(w) = E(g(X)|W = w)$ for each $w \in [a, b]$, and the relation $T(g_0) = E(Y|W)$, we have:

$\forall g \in L^2(X)$,

$$G(g) = \int_a^b [T(g - g_0)(w)]^2 (E(\epsilon^2|W = w))^{-1} dF_W(w).$$

For any $\tilde{g} \in L^2(W)$, define

$$||\tilde{g}||_W^2 = \int_a^b \tilde{g}(w)^2 (E(\epsilon^2|W = w))^{-1} dF_W(w)$$

It then follows that $G(g) = \|T(g - g_0)\|_W^2$.

Let us assume g_0 is point identified, i.e., $G(g) = 0$ if and only if $g = g_0$. We also assume T is compact, and hence has singular value system $\{|\lambda|_j, \phi_j, \psi_j\}$, where $T\phi_j = |\lambda_j|\psi_j$, and for each $g \in L^2(X)$, we have the singular value decomposition $g = \sum_{j=1}^{\infty} b_j\phi_j + Q$, where $Q \in \mathcal{N}(T)$, the null space of T. In the case of point identification, T is nonsingular, and therefore $Q = 0$.

Let the singular values $|\lambda_j|$, j=1,2,.. be ordered such that $|\lambda_1| \geq |\lambda_2| \geq ... > 0$, converging to zero. In addition, $\{\psi_j\}$ can be orthonormalized such that

$$\int_a^b \frac{\psi_i(w)^2}{E(\epsilon^2|W = w)} dF_W(w) = 1$$
$$\int_a^b \frac{\psi_i(w)\psi_j(w)}{E(\epsilon^2|W = w)} dF_W(w) = 0, i \neq j$$

We can then write $g_0(x) = \sum_{i=1}^{\infty} g_i\phi_i(x)$. For each $q \in \mathbb{N}$, the sieve space Θ_q is defined as the space spanned by $\{\phi_1, ..., \phi_q\}$, with $q = q_n \to \infty$ as n increases. Therefore each $g_q \in \Theta_q$ has an expansion $g_q(x) = \sum_{i=1}^{q} b_i\phi_i(x)$ with coefficients $\{b_i\}_{i=1}^q$. In addition, let $\|g_0\| = \sum_{i=1}^{\infty} g_i^2$, and $\|g_q\| = \sum_{i=1}^{q} b_i^2$.

The posterior distribution of the sieve approximation of g_0 is then given by

$$p(g_q|Data) \propto \exp\left(-na_n^2\|g_q\|^2 - \frac{n}{2}\bar{m}_n(g_q)^T V_0^{-1}\bar{m}_n(g_q)\right) \tag{4.17}$$

We now give the regularity conditions for the posterior consistency.

Assumption 4.4.3. *(i) T is nonsingular.*

(ii) (X, W) has joint density function f_{XW} and marginal density functions f_X and f_W respectively, which satisfy

$$\iint \frac{f_{XW}(x,w)^2}{f_X(x)f_W(w)}dxdw < \infty \tag{4.18}$$

Assumption 4.4.4. *There exists $\alpha \in (0,1)$ such that for any constant $c > 0$,*

$$\sup_{\|g_q\|\leq c} |\bar{m}_n(g_q)^T V_0^{-1}\bar{m}_n(g_q) - G(g_q)| = o_p(n^{-\alpha})$$

Assumption 4.4.5. *(i) $a_n^2 \to 0$, and $na_n^2 \to \infty$, as $n \to \infty$.*

(ii) There exists $\{s_n\}_{n=1}^{\infty} \subset \mathbb{N}$, $s_n \to \infty$, such that $\sum_{j\geq s_n} g_j^2 = O(a_n^2/\lambda_{s_n}^2) = o(1)$, as n increases.

(iii) $n \succ q_n \succ \max\{n^{1-\alpha}, na_n^2/\lambda_{s_n}^2, \lambda_{q_n}^{-2}\}$.

Assumption 4.4.3 guarantees the point identification of g_0, which was also assumed by Hall and Horowitz (2005). Hence g_0 can be recovered by $g_0 = T^{-1}\mu$ where $\mu(w) = E(Y|W = w)$. Condition (ii) guarantees that T is a compact operator (See Carrasco, Florens and Renault (2006), Section 2.2)). Assumption 4.4.4 assumes the rate of the uniform convergence to $G(g)$ on any compact subset of $L^2(X)$. Assumption 4.4.5 imposes the rate of convergence restrictions on the regularized parameter a_n, the singular values of T, the Fourier coefficients of g_0, and the dimension of the sieve space q_n. Roughly speaking, neither a_n nor $|\lambda_n|$ should converge to zero too fast. Moreover, the Fourier coefficients of g_0 should vanish at least as fast as $O(a_n^2/\lambda_{s_n}^2)$.

One can verify that Assumption 4.4.5 is satisfied, for example, if

$$\begin{aligned} s_n &= O(n^r), && \text{for some } r > 0 \\ |\lambda_j| &= O(j^{-p}), && \text{for some } 0 < p < \min\{\frac{1}{2}, \frac{1}{2r}\} \\ a_n^2 &= O(n^{-k}), && \text{for some } 2pr < k < 1 \\ \sum_{j \geq s_n} g_j^2 &= O(n^{-(k-2pr)}), && n \succ q_n \succ \max\{n^{1-\alpha}, n^{1-(k-2pr)}\}. \end{aligned}$$

Under these stated regularity conditions, the posterior distribution (4.17) is consistent when the parameter space for g_0 is not compact.

Theorem 4.4.1. *When the parameter space of g_0 is $\Theta = L^2(X)$, under Assumptions 4.4.3, 4.4.4 and 4.4.5, assuming that the posterior distribution is given by (4.17), we have*

$$E[\|g_{q_n} - g_0\|^2 | Data] = o_p(1)$$

Before presenting the numerical examples, I would like to give some final words on the regularized prior. The variance of the prior distribution (4.15) shrinks to zero, which is in the spirit of Tikhonov regularization scheme to deal with the inverse problem. On the other hand, it has a zero mean. Note that one of the attractiveness of the Bayesian approach for nonparametric instrumental regression problem is that it can incorporate prior knowledge of the structural function in the prior distribution. In fact, this can be achieved through a nonzero mean term. For instance, suppose a priorly it is known that g_0 is concave, one can choose a known concave function g^* as the mean in the prior. Hence the regularized prior becomes

$$\log p(g) \propto -na_n^2 ||g - g^*||^2 \tag{4.19}$$

I conjecture that by using prior (4.19), the posterior of the sieve approximation is still consistent because of the shrinkage of the prior variance. In addition, the posterior mean would be concave, as the prior mean is a concave function.

Conjecture 4.4.1. *If prior (4.19) is used, where g^* is either concave (monotone), under Assumptions in Theorem 4.4.1, the posterior is consistent, and the posterior mean is also concave (monotone).*

If this conjecture actually holds, one can then incorporate the prior knowledge of g_0 in the regularized prior distribution.

4.5. Markov Chain Monte Carlo

I present a simple simulation example in this section. The simulated model is a nonparametric IV regression, designed as:

$$\begin{aligned} &y = g(x) + u = \sin(x)\exp(\sqrt{|x|}) \\ &x = w + v \end{aligned}$$

where the errors u and v and instrument w are generated as

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} \sim i.i.d.N \left(0, \begin{pmatrix} 1.09 & 0.6 & 0 \\ 0.6 & 1.09 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right)$$

One can verify that $cov(x, u) = 0.6$, and hence x is endogenous. In addition, $cov(w, u) = 0$ and $cov(w, x) = 1$, so w is a valid instrumental variable.

I generated $n = 1000$ observations for (x, w, y). All the observations of w fall inside of interval $[-4, 4]$, so we partitioned $[-4, 4]$ evenly into $k = 50, 100$ subintervals. To implement the Bayesian procedure, $V = k * I$ was used as the weight matrix in the limited information likelihood, to replace the unknown V_0. We applied the Hermite series as the basis functions of the sieve approximation, i.e., $H_1(x) = 1, H_2(x) = x$, and $H_j(x) = H_{j-1}(x)x - (j-1)H_{j-2}(x)$, with $q = 4, 5, 6$ terms (we will comment on practical choice of q_n in next section).

For Approach 1, I placed the posterior in a compact set for the purpose of regularization. I placed i,i,d, prior $b_j \sim N(0, 0.5j^{-3})$ on each of the sieve approximation coefficients. The variance gets small for large index j, so that the priors of higher order coefficients b_j gradually concentrate around zero, which is designed to deal with the potential over-fitting problem. The simulation was carried out by conducting the Metropolis algorithm with $B = 1000$ replicates. The first 300 draws were "warm-ups" and were discarded to ensure the MCMC draws became stationary. Only the coefficients in the draw that satisfy $\sum_{j=1}^{q} |b_j| \leq 100$ were kept to ensure the posterior distribution is within a compact space. The sample averages of the remaining draws were calculated, treated as the estimated posterior mean of the sieve coefficients. Both the estimated curve and the true structural function are plotted in Figure 4.1.

For Approach 2, I relaxed the compactness assumption, but used the regularized prior (4.15), with $a_n = 0.05$ (small), 0.1 (moderate), and 0.8 (large). The estimated and the true curves are plotted in Figure 4.2.

Both figures demonstrate that, when q_n is moderate, the estimated curve capture the true curve fairly well. However, when $q_n = 6$, the finite sample bias is non-negligible, which may due to the over-fitting. Comparably, this problem is not severe in Approach 1, as the prior variances for higher order coefficients are decreasing to zer. Moreover, the finite sample bias

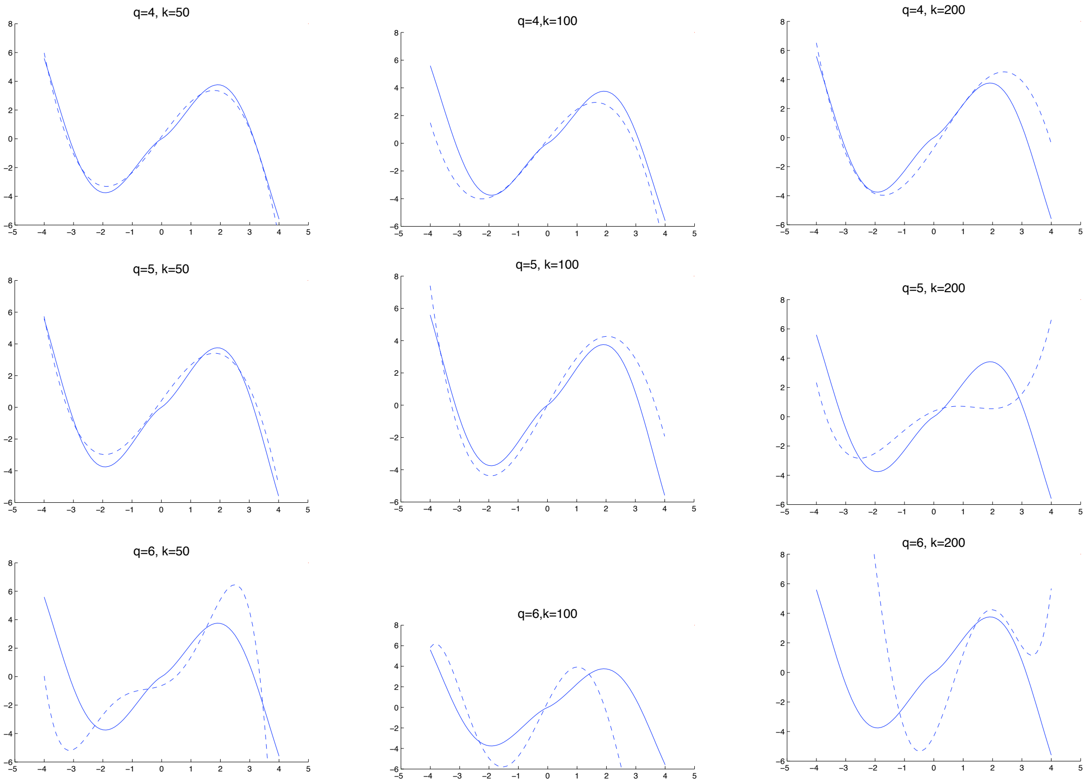


Figure 4.1. Approach 1, solid: true structural function; dashed: the estimated function

for Approach 2 also comes from the choice of a_n. Finally, our result is less sensitive to the choice of k_n.

4.6. Conclusion and Discussion

I studied the nonparametric conditional moment restricted model in a Bayesian approach, with a special focus on the frequentist properties of the posterior distribution. There was no any specific distribution assumed on the data generating process. In stead, I derived the posterior using the limited information likelihood, allowing the proposed procedure more flexible than the traditional nonparametric Bayesian approach by assuming a normal distribution on the error term, while the latter cannot avoid the risk of mis-specifying the underlying true distribution.

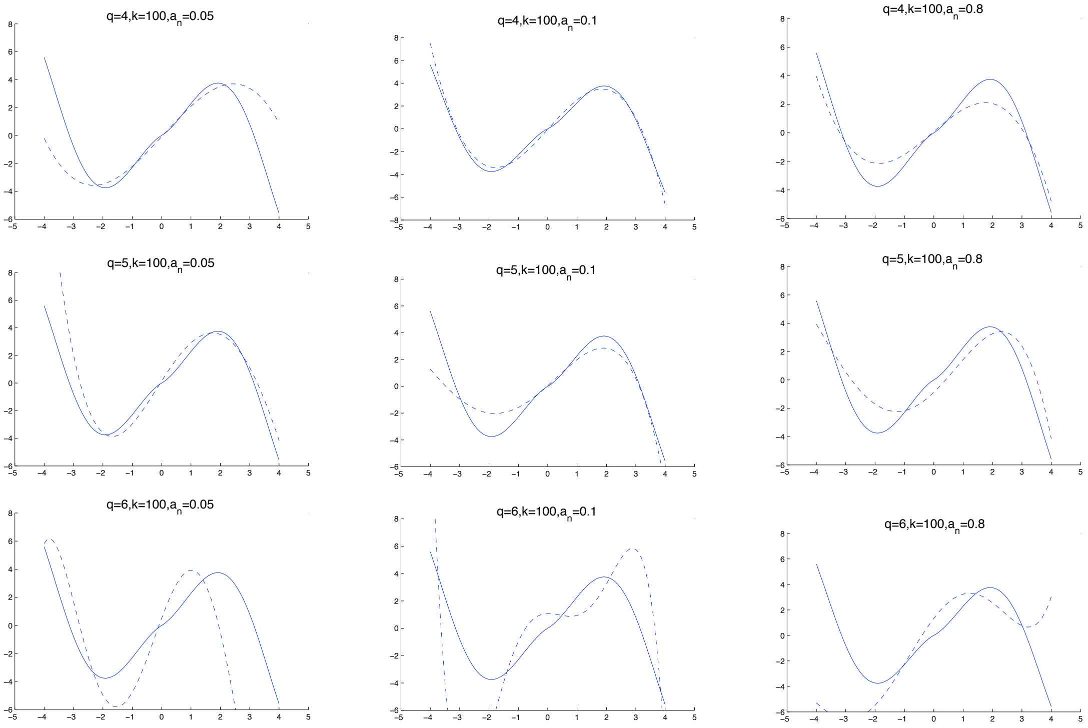


Figure 4.2. Approach 2, solid: true structural function; dashed: the estimated function

The limited information likelihood is the best approximation to the true likelihood by minimizing the Kullback-Leibler divergence. In fact, there are other alternative moment condition-based likelihood functions. For example, if we approximate the empirical distribution of the sampled data instead of the true likelihood, we end up with the empirical likelihood (Owen 1990). If additionally instead of Kullback-Leilber divergence, the chi-square distance is used as the metric, we will then obtain the generalized empirical likelihood (Imbens et al. (1998), Newey and Smith (2001) and Kitamura (2006)). It is still possible to establish the posterior consistency if these alternative moment condition-based likelihoods are used as the likelihood function, which is left as a future research direction.

In addition, our proposed Bayesian procedure allows the partial identification of g_0, which is more flexible than the traditional approach if the practical objective of interest is a linear continuous functional of g_0 instead of g_0 itself. This is because even if g_0 is not point identified, it is still possible to point identify $h(g_0)$ in many practical applications. The posterior of the functional can be constructed from the posterior of g_0, based on the MCM draws. It is also possible to incorporate the prior $\pi(h(g_0))$ for $h(g_0)$ directly, by transforming $\pi(h(g_0))$ into $p(b_i)$, the prior of the sieve coefficients in $g_q = \sum_{i=1}^{q} b_i\phi_i$, such that $p(h(g_q)) \approx \pi(h(g_0))$ where $p(h(g_q))$ is derived from $p(b_i)$. We will leave this as a future work.

The compactness of the parameter space was the key assumption to achieve the consistency for the general conditional moment restriction setting. We also showed that if the compactness assumption is relaxed, by imposing a regularized prior whose variance converges to zero asymptotically, the posterior distribution based on the limited information likelihood is still consistent in nonparametric IV regression. By imposing a regularized prior that depends on the tuning parameter a_n, the ill-posed problem in nonparametric IV regression is overcome.

In applications, our results require a priori choices of k_n, q_n and a_n, where k_n is the number of partitions of the support of W, q_n is the number of terms in the sieve approximation to g_0, and a_n is the tuning parameter in the regularized prior, to deal with the ill-posedness in nonparametric IV regression. We point out that our results are robust to the choice of k_n, however, sensitive to q_n. Although it is required q_n should diverge to infinity as n increases, it turns out that the sieve approximation with large value of q_n may suffer from over-fitting. It is for this reason Newey and Powell (2003) suggested choosing a small number for q_n, and Ai and Chen (2003) chose q_n simply by its assumed rate, which is also small. In addition, as pointed out by Chen (2007), some existing data-driven selection methods such as cross-validation, generalized cross-validation, and AIC may be used. Alternatively, it is possible to impose a Poisson prior on q_n with the mean parameter diverging to infinity, and then a proper q_n can be chosen from its posterior. The posterior consistency can also be achieved by such a procedure, and we will leave this as the future work. Finally, the tuning parameter a_n was used in our approach as well as in Hall and Horowitz (2005) to overcome the ill-posedness. Development of methods for selecting a_n is another important research topic.

CHAPTER 5

Bayesian Analysis for Classification Risk Using Empirical Likelihood

5.1. Introduction

One of the classical problem in data mining is to predict the unknown nature of a feature, by classifying the data into subgroups. Suppose $Y \in \{0,1\}$ is a binary variable to be predicted, which depends on a vector of covariates X. The classification and prediction are based on a classification rule $C(X,\theta) \in \{0,1\}$, which is a function of X and a certain action parameter θ. The action parameter θ is chosen such that under some risk function l, the expected risk $El(Y,C(X,\theta))$ is as small as possible. In the traditional classification problem, instead of minimizing the expected risk, if there are i.i.d. realizations $(Y_1,X_1),...,(Y_n,X_n)$, researchers choose θ to minimize the empirical risk

$$r_n = \frac{1}{n}\sum_{i=1}^{n} l(Y_i, C(X_i,\theta)) \tag{5.1}$$

Minimizing the empirical risk to find a good classification rule is a classical problem, which has been studied by a number of researchers, see for example, Devoye, Gyorfi and Lugosi (1996). Mohammadi and Van de Geer (2005) defined the optimal action parameter θ as the minimizer of the expected risk, and consistently estimated it by minimizing the empirical risk over a class of actions. More recently, Jiang and Tanner (2008) constructed the Gibbs posterior for the action parameter, and aimed at minimizing the risk function without modeling the data probabilistically. Other related works can be found, for example, in Koltchinskii and Panchenko (2002), among others.

In this chapter, we consider a Bayesian approach to making joint probabilistic inference on the action and the associated risk, without requiring a probability model for the underlying data generating process. This approach is more robust than the traditional likelihood-based method, which requires modeling the distribution of the data generating process. In Jiang and Tanner (2008), the Gibbs posterior is constructed from an empirical risk function, which does not require the probability distributional assumptions on the data. However, this approach lacks a Bayesian probability interpretation, in the sense that the likelihood function used for constructing the posterior is neither the true likelihood nor its approximation. Therefore, the Gibbs posterior cannot be interpreted as the posterior distribution of the action parameter in the traditional conditional probability sense. In this chapter, we overcome this difficulty by applying the empirical likelihood (Owen 1990). Let r denote the theoretical risk that satisfies

$$El(Y, C(X, \theta)) = r \tag{5.2}$$

The empirical likelihood is derived from the moment condition (5.2), as a function of (θ, r). It is well known that the empirical likelihood is the best approximation to the empirical distribution of the data in terms of the Kullback-Leibler distance subject to moment restrictions. Therefore, with a proper prior, the posterior distribution derived from the empirical likelihood has Bayesian interpretation asymptotically. Lazar (2003) provided simulation evidence in terms of the posterior coverage probability to show that the empirical likelihood can be used as a valid likelihood function for every absolutely continuous prior.

One of the most important features of the moment restriction (5.2) is that the parameters (θ, r) are only partially identified. That is, there are more than one pairs of (θ, r) in the parameter space satisfying the moment restriction. Therefore the posterior distribution of (θ, r) will not degenerate to any single point even asymptotically. In recent years, partially identified models are receiving rapid attentions in both statistics and econometrics literatures, and there are many works done in this growing area. In these models, the identified region, defined as the set of parameters that satisfy the moment restriction (5.2), becomes the object of interest. See for example, Chernozhukov, Hong and Tamer (2007). More recently, Liao and Jiang (2010) have studied the properties of the posterior distribution of the parameters in a similar setting of the moment restriction (5.2), where they used the limited information likelihood idea (Kim 2002) to construct the likelihood function. In this chapter, we show that the posterior distribution, constructed based on the empirical likelihood, has similar asymptotic properties to those described in Liao and Jiang (2010). To be specific, we will show that the joint posterior distribution for (θ, r) will be asymptotically supported on an arbitrarily small neighborhood of the curve $\{(\theta, r) : El(Y, C(X, \theta)) = r\}$. So far the posterior consistency of the empirical likelihood for partially identified models has not been formally established, while the point identified case was previously studied by Chernozhukov and Hong (2003) and Moon and Schorfheide (2004). Therefore an important contribution of this chapter is that we show the consistency of the posterior distribution, constructed based on the empirical likelihood, for the parameters that are only partially identified by the moment condition of the form (5.2).

Note that the posterior distribution $P(\theta, r|Data)$ allows us to construct the conditional posterior distribution $P(\theta|r, Data)$, which is the posterior distribution of the action to achieve a given tolerant risk level. To our best knowledge, the posterior distribution of the action conditional on the theoretical misclassification risk has not been characterized before. The consistency of $P(\theta, r|Data)$ implies that given r being controlled at a certain level $r \leq r_0$, and θ is generated from $P(\theta|r \leq r_0, Data)$, the true expected risk $El(Y, C(X, \theta)) \leq r_0 + \epsilon$ for any $\epsilon > 0$, with posterior probability approaching one, regardless of the distribution of the data generating process.

We notice that, compared to the literature on the classical empirical risk minimization (ERM) approach, there is disproportionally less work done on the inferential side of the classification problem. The classical ERM approach obtains an optimal θ by minimizing the empirical risk. However, in such an approach, neither the associated true risk nor the posterior distribution for θ that achieves such a true risk are known to us (although we may be able to find an asymptotic confidence interval for the true risk based on the minimum empirical risk). An alternative Bayesian approach is to model the probability $P(Y = 1|X = x)$ either parametrically or nonparametrically, and then use a likelihood-based posterior, see for example, Coram and Lalley (2006). In comparison, our proposed method can describe, for example, the posterior distribution of the risk associated with any action θ, and on the other hand, the conditional posterior of θ to achieve a certain tolerant risk level r_0. Therefore, instead of improving the prediction accuracy and the risk minimization, which has already been paid a huge amount of efforts in the data mining literature, our contribution is to provide a new language for probabilistic inference on both the risk and actions.

The remainder of this chapter is organized as follows. Section 5.2 will introduce the basic model framework and the posterior distribution based on the empirical likelihood. Section 5.3 presents the main results of this chapter. Section 5.4 comments the possible extensions to a more general risk function in data mining. Section 5.5 provides some simulation examples to illustrate the main ideas of this chapter and demonstrates how they are used in practice. Finally, Section 5.6 illustrates an empirical application of the credit classification using the German Credit Benchmark data.

5.2. Empirical Likelihood Posterior Distribution

Consider the following equation

$$r = E[\rho(W,\theta)|\theta] = \int \rho(W,\theta)P(dW) \tag{5.3}$$

where θ is a parameter used in the action of data mining and r is the resulting risk. Here $W = (Y,X)$ with $Y \in \{0,1\}$ being the label to be predicted and X as an input. The classification loss $\rho(W,\theta) = |Y - C(X,\theta)|$ where $C(X,\theta) \in \{0,1\}$ is a classification rule labelled by θ. The probability measure $dF(W)$ is based based on the (true) distribution of W. We assume the action parameter θ belongs to an action space Θ, and therefore the parameter space for (θ, r) is $\Theta \times [0,1]$.

We will regard (5.3) as a moment condition and base on this alone, without further modeling the distribution of W, construct a posterior distribution jointly for the action-risk parameters (θ, r). This can be done by applying the empirical likelihood generated by the moment condition. Assume that we observe a data set $D = (W_1, ..., W_n)$, which are assumed to be iid realizations of W. The empirical likelihood based on (5.3) is defined by (Owen 1990, and Qin and Lawless 1994):

$$P_{EL}(D|\theta,r) \quad = \quad \sup_{p_1,\dots,p_n} \{\prod_{i=1}^{n} p_i | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i[\rho(W_i,\theta) - r] = 0\}$$

$$= \exp\{-\max_{\mu\in\mathbb{R}}\sum_{i=1}^{n}\log\{1+\mu[\rho(W_i,\theta)-r]\}\}. \tag{5.4}$$

The empirical likelihood procedure has a good distribution interpretation using information theory. Let $\mathcal{P}$ denote the space of probability measures on the Borel $\sigma-$field on $\mathbb{R}^{\dim(W)}$. Define $\mathcal{M}(\theta,r)=\{\mu\in\mathcal{P}:\int\rho(w,\theta)d\mu(w)=r\}$. It can be verified (see Kitamura 2001) that the empirical likelihood (5.4) is the solution to

$$\inf_{P\in\mathcal{M}(\theta,r)} I(\mu_n||P)$$

where μ_n denotes the empirical measure of D, and $I(P_1||P_2)$ denotes the Kullback-Leibler distance between P_1 and P_2. Therefore, although (5.4) is not the true likelihood function based on the data, it is the best approximation to the empirical distribution of the data under the moment restriction (5.3), in terms of the Kullback-Leibler distance. This yields the Bayesian interpretation of the posterior distribution based on the empirical likelihood of this chapter. Suppose a prior distribution $\pi(\theta,r)$ is assigned in some sense, the resulting posterior then becomes (up to a normalization factor):

$$P(\theta,r|D)\propto\exp\left\{-\max_{\mu\in\mathbb{R}}\sum_{i=1}^{n}\log\{1+\mu[\rho(W_i,\theta)-r]\}\right\}\pi(r,\theta). \tag{5.5}$$

Instead of achieving a better risk minimization, in this chapter we are aiming at providing a new language for probabilistic inference on the risk and action. The posterior distribution is a flexible formalism that allows us to derive a number interesting results, for example:

(1) $P(r|\theta,D)$, which is the posterior distribution of the resulting risk achieved by a given action θ;

(2) $P(\theta|r = r_0, D)$ or $P(\theta|r \leq r_0, D)$, which is the posterior distribution of the action θ needed to achieve a risk being r_0 or at most r_0 respectively;

(3) $P(\theta \in A_1|r \leq r_0, D)/P(\theta \in A_2|r \leq r_0, D)$, which compares the posterior probabilities of two 'models' $A_{1,2}$ (or two sets of actions), in order to achieve risk at most r_0.

(4) $P(r|D)$, which is the posterior distribution of the achievable risks by all possible actions θ from the support of the prior Θ.

Note that the risk r is defined by $r = E[\rho(W, \theta)|\theta]$, which depends on the action parameter θ and the underlying distribution P_W, and therefore can be written as $r = r(\theta, P_W)$. Here r is not fixed even if θ is, because P_W is unknown and depends on the unknown parameters, and therefore, from a Bayesian point of view, is random. A possible alternative way is to put prior on (θ, P_W), and obtain the posterior of r from $r = r(\theta, P_W)$. However, it is then necessary to model P_W non-parametrically. Our proposed approach is more convenient: by putting the joint prior on $\pi(r, \theta)$ directly, we can let the data tell the functional relationship between r and θ through the joint posterior distribution. For example, in the simulation study of Section 5.5, the priors of r and θ are assumed to be independent. However, the scatterplot of the MCMC draws of (r, θ) from the posterior distribution (Figure 5.1) clearly illustrates a functional relationship between them.

5.3. Main Results

In the classification problem when $\rho = |Y - C(X, \theta)|$ and $Y, C(X, \theta) \in \{0, 1\}$, it is straightforward to verify that the empirical likelihood and the corresponding posterior distribution have explicit analytic expressions. Define the empirical risk

$$\hat{R}(\theta) = n^{-1}\sum_{i=1}^{n} |Y_i - C(X_i, \theta)|. \tag{5.6}$$

The following theorem says that, in the classification problem framework, the log-empirical likelihood function is proportional (up to the scale $-n$) to the Kullback-Leibler distance between two Bernoulli distributions with success probabilities $\hat{R}(\theta)$ and r respectively.

Theorem 5.3.1. *When $\rho = |Y - C(X,\theta)|$ and $Y, C(X,\theta) \in \{0,1\}$ $\hat{R}(\theta), r \in [0,1]$, where $\hat{R}$ is given in (5.6), then the posterior distribution for (θ, r) using the empirical likelihood is given by*

$$P(\theta, r|D) \propto \exp(-nK(\hat{R}(\theta), r))\pi(\theta, r), \tag{5.7}$$

where

$$K(p,q) = \begin{cases} p\ln(p/q) + (1-p)\ln\{(1-p)/(1-q)\}, & \textit{if } p, q \in (0,1) \\ +\infty, \quad \textit{if } p \in (0,1], q = 0, \ \textit{or } p \in [0,1), q = 1 \\ 0 \quad \textit{if } q \in [0,1), p = 0, \ \textit{or } q \in (0,1], p = 1. \end{cases} \tag{5.8}$$

PROOF. See the Appendix.

Note that (5.3) does not identify (θ, r) (nor necessarily $\theta|r$). That is to say, the posterior distribution of (θ, r) will not degenerate to any single point even asymptotically. Based on a treatment similar to Liao and Jiang (2010), however, we can derive a posterior consistency result for this partially identified framework. The following theorem is our main result of this chapter, which shows that the posterior $P(\theta, r|D)$ will cluster around the region of true parameters that satisfy the momement condition (5.3).

Theorem 5.3.2. *Consider the classification case, when* $\rho(W) = |Y - C(X,\theta)|$ *and* $Y, C(X,\theta) \in \{0,1\}$. *Denote* $R(\theta) = E[\rho(W,\theta)|\theta]$, $\hat{R} = n^{-1}\sum_{i=1}^{n} \rho(W_i, \theta)$, *and* $\eta(\theta, r) = \min\{R, 1 - R, r, 1 - r\}$. *Assume the following:*

(i) The prior $\pi(|R - r| \leq \delta, \eta \geq \tau) > 0$ *for any postive constants* δ, τ*;*

(ii) $\sup_{\theta\in\Theta} |\hat{R} - R| \rightarrow^{P*} 0$ *in the probability of* D *as* $n \rightarrow \infty$*;*

then for any $\epsilon > 0$*, we have: in the probability of D, as* $n \rightarrow \infty$

$$P(R(\theta) - \epsilon \leq r \leq R(\theta) + \epsilon|D) \rightarrow^{P*} 1.$$

Proof. See the Appendix.

Intuitively, a generalized posterior consistency theory implies that the posterior distribution should be asymptotically supported around the set of minimizers of $nK(\hat{R}(\theta), r)$, the power in the likelihood function. Since $K(\hat{R}(\theta), r)$ is the Kullback-Leibler distance between $Bernoulli(\hat{R}(\theta))$ and $Bernoulli(r)$, it is minimized when $R(\theta)$ and r are close to each other. This theorem indicates that the posterior of (r, θ) based on the empirical likelihood indeed asymptotically clusters around the curve defined by the moment restriction $\{(r, \theta) : r = R(\theta)\}$. Therefore, even though the posterior of the risk $R(\theta)$ is unknown (since it also depends on the unknown distribution P_W), we can look at the posterior of r instead.

The following corollaries describe two useful implications. The first says that if we would like to find actions to control the true risk to be at most some desired risk level r_0, then we can use actions randomly generated by the conditional posterior distribution $P(\theta|r \leq r_0, D)$, which will tend to generate actions θ that have resulting true risks $R(\theta)$ at most a little bit worse than the desired level r_0.

Corollary 5.3.1. *Suppose that* $P(r \leq r_0|D) > \xi$ *for some constant* $\xi > 0$*, then under the regularity conditions in Theorem (5.3.2), for any* $\epsilon > 0$*,* $P(E[\rho(W,\theta)|\theta] \leq r_0 + \epsilon|D, r \leq r_0) \rightarrow^{P^*} 1$ *as* $n \rightarrow \infty$*.*

Proof. Denote $R = E[\rho(W,\theta)|\theta]$. Then $P(R > r_0 + \epsilon|D, r \leq r_0) = P(R - r_0 > \epsilon, r \leq r_0|D)/P(r \leq r_0|D) \leq P(|R - r| > \epsilon|D)\xi^{-1} \rightarrow^{P^*} 0$ due to Theorem (5.3.2). Q.E.D.

Define $r^* = \inf_{\theta \in \Theta} E[\rho(W,\theta)|\theta]$, the minimum expected risk over all the actions in Θ. The next corollary states that the posterior distribution for r has no support below r^* asymptotically.

Corollary 5.3.2. *Under the regularity conditions in Theorem (5.3.2), for any* $\epsilon > 0$*,* $P(r < r^* - \epsilon|D) \rightarrow^{P^*} 0$ *as* $n \rightarrow \infty$*.*

Proof. $P(r < r^* - \epsilon|D) = P(r + \epsilon < \inf_\theta R|D) \rightarrow^{P^*} 0$ due to Theorem (5.3.2).

Finally, for the sake of numerical computation, we point out that the Beta distribution is a conjugate prior for the conditional posterior distribution of $r|\theta, D$. To be more specific, if the priors for θ and r are independent, and $\pi(r)$ is Beta(a,b), then straightforward calculation yields:

$$\begin{aligned} P(r|\theta, D) &\sim Beta(n\hat{R}(\theta) + a, n(1 - \hat{R}(\theta)) + b) \\ P(\theta|D) &\propto \pi(\theta)\hat{R}(\theta)^{-n\hat{R}(\theta)}(1 - \hat{R}(\theta))^{-n(1-\hat{R}(\theta))}B(n\hat{R}(\theta) + a, n(1 - \hat{R}(\theta)) + b) \end{aligned}$$

where $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$.

5.4. More general risk functions in data mining

It is noted that while we have focused on the classification risk, the current method and and theoretical results can be easily generalized to other risk functions of the form $R = E[\rho(W, \theta)]$, where $\rho = f(Y, A)$ with $Y \in \{0, 1\}$ and $A = A(X, \theta) \in \{0, 1\}$. (One simple example is a linear rule $A = I(X^T\theta > 0)$ parameterized by θ.) For one example in a data mining context: A marketing effort $A = I$[mail] of mailing out an advertisement with cost $c = 1$ will be based on x (including, e.g., gender, age, ethnic group, education, ...). The outcome will be $Y = I[purchase]$ where a purchase will lead to net income $g = 100$. Then one would like to maximize the expected profit $E[(gY - c)A]$ or minimize a risk $R = constant - E[(gY - c)A]$. Here up to a constant, $f(Y, A) = -(gY - c)A$, so that $f(0, 0) = f(1, 0) = 0$, $f(0, 1) = c = 1$, $f(1, 1) = c - g = -99$. Such profit-and-loss decision matrices are included in popular data mining software such as SAS Enterprise Miner. We can apply the proposed method to construct a posterior distribution jointly for the action parameter a and the resulting risk r.

5.5. Simple Monte Carlo Example

Let the data be generated from the following design

$$Y = I_{(3X - \epsilon > 0)}$$
$$X \sim N(0, 1), \epsilon \sim N(0, 3)$$

where X and ϵ are independent. We apply the classification rule $C(X, \theta) = I_{(X - \theta > 0)}$. Let $\rho(\theta) = |Y - C(X, \theta)|$, one can then show that the expected risk is given by

$$E[l(\theta)|\theta] = E_X\{[1 - \Phi(\sqrt{3}X)]I_{(X > \theta)} + \Phi(\sqrt{3}X)I_{(X \leq \theta)}\} \tag{5.9}$$

where the expectation E_X is taken with respect to the distribution of X, which is standard normal. We generated $n = 1000$ data points $(Y_1, X_1), ..., (Y_n, X_n)$. The posterior for (θ, r) were constructed according to (5.7) based on the empirical likelihood, with priors $\pi(\theta) \sim N(0, 1)$, $\pi(r) \sim U[0, 1]$, and $\pi(\theta, r) = \pi(\theta)\pi(r)$. According to Theorem 5.3.2, the posterior distribution should be clustered around the risk curve $\{(\theta, E[l(\theta)|\theta])\}$. To illustrate this matter of fact, $B = 10,000$ MCMC draws were generated from the posterior. In each step of the Metropolis algorithm, we used proposal density $\theta^t \sim N(\theta^{t-1}, 0.5)$, and $r^t \sim U[0, 1]$. The first quarter of the draws were discarded to ensure that the MCMC procedure becomes stationary.

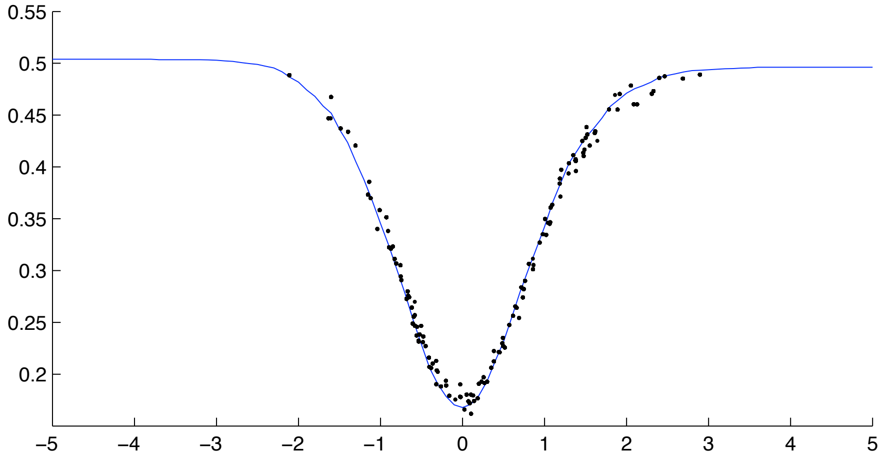
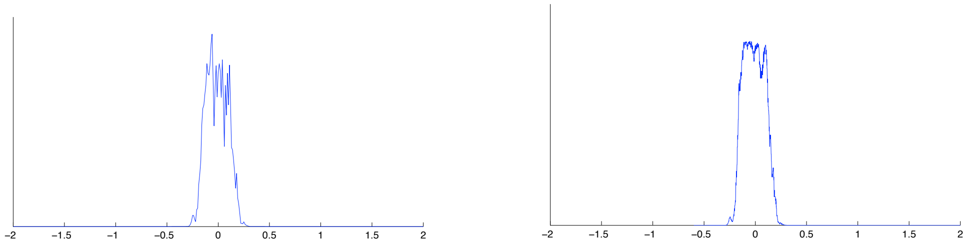


Figure 5.1. Design 1: Expected risk curve and MCMC draws

Figure 5.1 plots the expected risk curve $E[l(\theta)|\theta]$ against θ, and the scatterplot of the MCMC draws of (θ, r) from the posterior distribution. It is clearly illustrated that the MCMC draws are clustered around the true expected curve, supporting our posterior consistency result.

Our method also gives the posterior distribution of the action θ to achieve a certain risk level. For example, if we want to control the classification risk to be at most 5% minimum quantile of the posterior distribution of r, let r_0 be the 5% quantile of the MCMC draws of r, $r_0 = 0.1795$, then the desired posterior for θ is given by $P(\theta|D, r \leq r_0)$. If we want to control the classification risk to be the at the minimized empirical risk level $\min_\theta \hat{R}(\theta) = 0.1748$, then the posterior is given by $P(\theta|D, r = \min_\theta \hat{R}(\theta))$. See Figure 5.2 as the plotted posteriors for θ respectively.

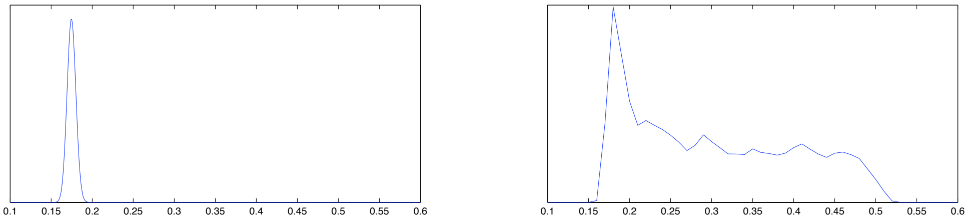
Figure 5.2. Left: $P(\theta|D, r \leq r_0)$; Right: $P(\theta|D, r = \min \hat{R})$



In addition, we can also plot $P(r|D, \theta = \arg\min_\theta \hat{R}(\theta))$, which is the posterior of the risk given the optimal action $\arg\min_\theta \hat{R}(\theta)$. From the previous discussion we know that the uniform prior is a conjugate prior for $P(r|\theta, D)$. Therefore, $P(r|D, \theta = \arg\min_\theta \hat{R}(\theta))$ is Beta$(\min \hat{R} + 1, n(1 - \min \hat{R}) + 1)$. Finally, note that in this design, $\min_\theta E[l(\theta)|\theta] = 0.168$. The marginal posterior of r is also plotted, which is obtained by numerically integrating out θ. The plot indicates that the $P(r|D)$ has not much support below the minimum expected risk. See Figure 5.3 for these plots.

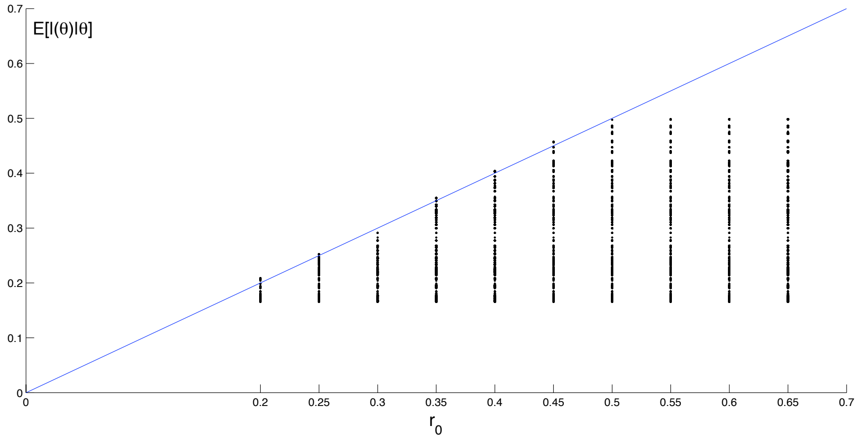
The final Figure 5.4 plots the theoretical risk $E[l(\theta_i)|\theta_i]$ versus the controlled level of risk r_0, where each θ_i is from one of the MCMC draws (θ_i, r_i), whose corresponding r_i is less than or equal to r_0, and $E[l(\theta_i)|\theta_i]$ is as given by (5.9). Therefore, one can think of the dots in Figure 5.4 with the same horizontal axis r_0 as the distributions of the theoretical risk $E[l(\theta)|\theta]$, where θ is generated from $P(\theta|r \leq r_0, D)$. We can see that almost all the dots are below the identical

Figure 5.3. Left: $P(r|D, \theta^* = \arg\min_\theta \hat{R})$; Right: $P(r|D)$



line $y = r_0$. This indicates that, once the action parameter θ is generated from the conditional posterior $P(\theta|r \leq r_0, D)$, with the risk being controlled under level r_0, the theoretical L_1 risk will also be less than or equal to r_0.

Figure 5.4. $E[l(\theta)|\theta] : \theta \sim P(\theta|r \leq r_0, D)$ versus r_0



5.6. German Credit Data: an Empirical Application

5.6.1. Data set and model specification

As an empirical example, we apply the proposed Bayesian empirical likelihood method to the credit risk classification, using the German Credit Benchmark data set provided by Asuncion and Newman (2007). The data set consists of n =1,000 past applicants and their credit rating (GOOD or BAD), which serves as the target variable Y. In addition, there are 24 attributes served as input variables, which are used as the covariates X. The attributes are either ordered categorical, such as “Credit History”, “Personal Status and Sex”, “Housing”, “Employment”, or numerical, such as “Credit Duration”, “Credit Amount”, and “Age”. See Asuncion and Newman (2007) for a complete description of the data set.

The classification rule using the jth observation is $C(X_j, \theta) = I(X_{1j}\theta_0+\theta_1+\sum_{i=2}^{24} X_{ij}\theta_i > 0)$. Here X_{ij} denotes the realization of individual j on variable $X_i, i = 1, 2, ..., 24$. Note that for the identifiability in regular binary response models, Horowitz (1992) suggested that X_1 should be a continuous variable whose coefficient $\theta_0 \in \{-1, +1\}$, i.e., X_1 is always kept in the model. Therefore, the components of the covariates are arranged so that X_{1j} denotes the jth observation of “Credit Duration”, which is a continuous variable, and it is reasonable to assume that it is related to each customer’s credit behavior. We also included an intercept term θ_1. In addition, to make sure the ranges of the covariates do not vary too much for the purpose of comparison, the continuous attributes were normalized, i.e., subtracted mean and divided by the standard deviation.

For the credit classification problem, the cost matrix is given by the following table.

Table 5.1. Cost Matrix

		Classification	
		GOOD	BAD
Target Variable	GOOD	0	1
	BAD	5	0

Note that the cost matrix is asymmetric, this is because the penalization for mis-classifying a Bad target variable Y into Good should be more severe than the vice versa. Therefore the loss function is

$$\rho(Y, X; \theta) = I(Y = Good, C(X, \theta) = Bad) + 5I(Y = Bad, C(X, \theta) = Good).$$

The method proposed in the previous sections can be applied for variable selection. Let $\psi = (\psi_1, ..., \psi_{24})$ denote a vector of selection indicators such that for each $i = 1, ..., 24$, $\psi_i = 1$ if variable X_i is selected, and $\psi_i = 0$ otherwise. We then have the "entered" parameters $\theta^{\psi} = (\theta_1\psi_1, ..., \theta_{24}\psi_{24})$. A zero component of θ^{ψ} means that the corresponding covariate is not included. The first component is set to $\psi_1 = 1$ so that the intercept θ_1 is always kept in the model. Therefore the actual classification rule based on the "entered" parameters θ^{ψ} is $C(X_j, \theta^{\psi}) = I(X_{1j}\theta_0 + \theta_1 + \sum_{i=2} X_{ij}\theta_i\psi_i > 0)$.

The log-empirical likelihood function for (θ, r, ψ) is thus given by

$$\log EL(\theta, r, \psi) = -\max_{\mu \in \mathbb{R}} \sum_{i=1}^{n} \log\{1 + \mu[\rho(Y_i, X_i, \theta^{\psi}) - r]\}.$$

Let $|\psi|$ denote the number of nonzero components of ψ, and I_k denote the $k \times k$ identity matrix. We specify the priors as follows: $\pi(\theta, r, \psi) = \pi(\theta|\psi, r)\pi(\psi, r) = \pi(\theta^{\psi}|\psi)\pi(\psi)\pi(r)$, where

$$\begin{aligned} \theta_0 = 2\gamma - 1, &\qquad \gamma \sim Binomial(1, 0.5) \\ \theta^{\psi}|\psi \sim N(0, 10I_{|\psi|}), &\qquad r \sim Uniform[0, 5] \\ \psi_i \sim Binomial(1, \lambda), i = 2, ..., 24 &\quad \psi = (1, \psi_2, ..., \psi_{24}) \end{aligned}$$

λ is a pre-specified parameter, determined by the expected number of selected covariates. The posterior distribution is then given by

$$p(\theta, \psi, r|Data) \quad \propto \quad EL(\theta, r, \psi)e^{-\frac{1}{20}\sum_{i=1}^{24}\theta_i^2\psi_i}\lambda^{|\psi|-1}(1-\lambda)^{24-|\psi|}I(0 < r < 5)$$

$$\begin{aligned}\log p(\theta, \psi, r|Data) &= \log EL - \frac{1}{20}\sum_{i=1}^{24}\theta_i^2\psi_i + (|\psi| - 1)\log\lambda + (24 - |\psi|)\log(1-\lambda) \\ &\quad + \log I(0 < r < 5) + Constant\end{aligned}$$

5.6.2. Algorithm description

The Metropolis-Hastings algorithm was conducted to obtain the MCMC draws from the posterior distribution. Similar to the algorithm proposed by Chen, Jiang and Tanner (2009), each iteration combines BETWEEN steps that propose changes between different models, with the WITHIN steps that propose changes of θ within a fixed model. These steps are given as follows: (in the algorithm below, denote $q(\theta_j)$ as the density of $N(0, 0.5)$.)

BETWEEN Step Update θ to θ' with model indices changing from ψ to ψ'.

(1) (Add/Delete) Randomly choose an index $j \in \{2, ..., 24\}$.

- If $\psi_j = 1$, propose $\psi_j' = 0$ and let $\theta_j' = 0$ with all remaining components of θ unchanged. This proposal is accepted with probability

$$\min\left\{1, \frac{p(\theta', \psi', r|Data)q(\theta_j)}{p(\theta, \psi, r|Data)}\right\}$$

- If $\psi_j = 0$, propose $\psi_j' = 1$, and generate $\theta_j' \sim N(0, 0.5)$ with all remaining components of θ unchanged. This proposal is accepted with probability

$$\min\left\{1, \frac{p(\theta', \psi', r|Data)}{p(\theta, \psi, r|Data)q(\theta_j')}\right\}$$

(2) (Swap) When $1 < |\psi| < 24$, randomly choose two indices $k, l \in \{2, ..., 24\}$, such that $\psi_k = 0$ and $\psi_l = 1$. Propose $\psi_k' = 1$ and $\psi_l' = 0$, and $\theta_k' \sim N(0, 0.5), \theta_l' = 0$. This

proposal is accepted with probability

$$\min\left\{1, \frac{p(\theta', \psi', r|Data)q(\theta_l)}{p(\theta, \psi, r|Data)q(\theta'_k)}\right\}$$

WITHIN Step Update θ' to θ^* with model indices fixed and with the nonzero values of θ' changed: For each index $j \geq 1$ such that $\psi'_j = 1$, generate $\theta^*_j \sim N(\theta'_j, 0.5)$. Generate $\gamma \sim Binomial(1, 0.5)$, and let $\theta_0 = 2\gamma - 1$. Generate $r^* \sim Uniform[0, 1]$. Accept θ^* with probability

$$\min\left\{1, \frac{p(\theta^*, \psi', r^*|Data)}{p(\theta', \psi', r|Data)}\right\}$$

Each candidate model can be indexed by a specific realization of ψ, hence denoted by M_ψ. Let r_0 denote the certain level at which the risk r is to be controlled. We compare the candidate models by the conditional posterior probability $P(M_\psi|r \leq r_0, Data)$. For each realization of ψ, let B_{ψ,r_0} denote the number of appearances of ψ in the MCMC draws whose corresponding r is less than or equal to r_0, and let B_{r_0} denote the number of draws with r less than or equal to r_0. Therefore $P(M_\psi|r \leq r_0, Data)$ can be estimated by

$$\hat{P}(M_\psi, r_0) = \frac{B_{\psi,r_0}}{B_{r_0}} \tag{5.10}$$

5.6.3. Result

The algorithm described previously was carried out for $B = 2,0000$ iterations, with $\lambda = 0.4$. The first one fifth of the draws were discarded for the MCMC procedure to warmup. The minimum sampled risk is 0.65 (note that the loss function $\rho(Y, X; \theta^\psi)$ is continuously supported on $[0, 5]$), and the 1%-percentile is about 0.68. When $r_0 \in [0.65, 0.68]$, only two sampled models have probability higher than 0.1. Table 5.3 summarizes the selected variables in the corresponding models. In addition, Figure 5.5 plots the estimated posterior probabilities $\hat{P}(M, r_0)$ (5.10) of thirty of the sampled models, as functions of r_0. As r_0 increases, the posterior probabilities of the sampled models decrease, since there are more models sampled. This demonstrates that if the level of the risk that is tolerable is large, many different types of models can achieve it.

To verify that the theoretical risk level $r_0 = 0.68$ is actually achievable, the original data set was then randomly divided into two groups: training (2/3) and validation (1/3). Both models in Table 5.3 were first fitted with the training data. For each model M_j, we generated $10,000$ MCMC samples of the corresponding selected action parameters $\{\theta^i\}_{i=1}^{10,000}$ from $P(\theta | r \leq r_0,$ training data, $M_j)$. These draws were then used to construct the classification rules with the validation data $(Y_1, X_1), ..., (Y_{nv}, X_{nv})$. The posterior expectation (based on the training data) of the empirical risk (based on the validation data) is defined as $E[\frac{1}{n_v}\sum_{k=1}^{n_v} \rho(Y_k, X_k; \theta) | r \leq r_0,$ training data, $M_j]$, which is estimated by the MCMC sample analogue:

(5.11)

$$\hat{R}_j = \frac{1}{10,000}\frac{1}{n_v}\sum_{i=1}^{10,000}\sum_{k=1}^{n_v}\rho(Y_k, X_k; \theta^i) \approx E[\frac{1}{n_v}\sum_{k=1}^{n_v}\rho(Y_k, X_k; \theta) | r \leq r_0, \text{ training data}, M_j]$$

In addition, for each $M_j, j = 1, 2$ in Table 5.3, we calculated the optimal action $\theta_{(j)}$ by minimizing the empirical risk based on the training data

$$\theta_{(j)} = \arg\min_{\theta} \frac{1}{n_c} \sum_{i \in Training} \rho(Y_i, X_i, \theta' \psi^j)$$

where ψ^j corresponds to model M_j. The optimal action was then used to generate the classification rule $C(X, \theta)$ for the validation data. The empirical risk of model j based on the validation data is given by

$$\min\text{-}ER_j = \frac{1}{n_v} \sum_{i \in Validation} \rho(Y_i, X_i; \theta'_{(j)}\psi^j) \tag{5.12}$$

To see whether model selection can lead a better prediction accuracy, we also compared $\hat{R}_j$ and ER_j with the empirical risk based on the validation data without model selection ER-ALL, where the classification rule was generated by minimizing the empirical risk of the training data, using all the 24 covariates. Table 5.2 summarizes $\hat{R}_j$ and $\min$-ER_j. We can see that $r_0 = 0.68$ is achieved by both models $M_j : j = 1, 2$. In addition, after minimizing the empirical risk using the training data, the optimal action parameter for both models indeed have led to a better empirical risk on the validation data (.682 and .679), compared to the empirical risk using all the 24 covariates (.730). On the other hand, the selected models have better interpretability of the rationale for the credit decision.

Table 5.2. Comparison of $\hat{R}$ versus r_0, $\min$-ER versus ER-ALL

Model	$\hat{R}$	r_0	$\min$-ER	ER-ALL
M_1	0.679	0.680	0.682	0.730
M_2	0.659		0.679	

$\hat{R}$, min-ER, and $ER - ALL$ represent the posterior expectation of the empirical risk after model selection (5.11) , the empirical risk after model selection (5.12), and the empirical risk without model selection. All the action parameters are calculated based on training data, and empirical risks are based on validation data.

Table 5.4 summarizes the accuracy for correctly classifying a good customer (Type I), bad customer (Type II), and the overall classification in the validation data, defined as the ratio of correct classifications (Type I, Type II and overall) over the total number of good, bad, and overall customers respectively. The classification action is obtained by using the training data, as $\arg\min_\theta \frac{1}{n_c} \sum_{i \in Training} \rho(Y_i, X_i, \theta'\psi^j)$, where ψ^j corresponds to model $M_j : j = 1, 2$ in

Figure 5.5. Estimated $P(M_\psi | r \leq r_0, Data)$ versus r_0

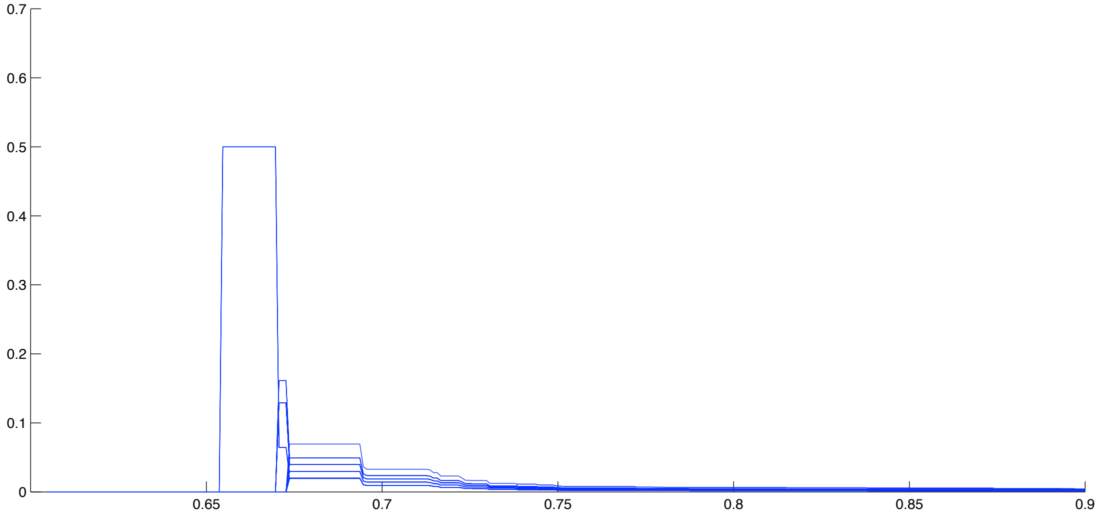


Table 5.3. Since the loss function pays five times more penalization on mis-classifying BAD customers than mis-classifying GOOD customers, the optimal classification rule under the selected models yields very good Type II prediction accuracy, which protects against accepting the BAD customers. Model M_2 appears to yield a better classification: while maintaining a high accuracy of the Type II classification, it performs much better in the Type I and overall classification.

Finally, we point out that the selected variables in the optimal models are sensitive to the choice of λ. Here λ denotes the prior expectation of the number of selected variables, which was set to 0.4 in our application. Hence a priorily we expect around 10 variables to be selected. If it is set to other values, the selected variables may be different. This fact is reasonable, since

Table 5.3. Models with $r \leq 1\%$- percentile

Model	Variables	Posterior Probability
M_1	Duration of Credit	0.5
	Credit History	
	Credit Amount	
	Present Employment Since	
	Real Estate Property	
	Age	
	Num. of Existing Credits at Bank	
	Num. of People Being Liable	
	Other Debtors/ Guarantors	
	Credit Purpose	
M_2	Duration of Credit	0.5
	Credit History	
	Credit Amount	
	Present Employment Since	
	Real Estate Property	
	Age	
	Num. of Existing Credits at Bank	
	Num. of People Being Liable	
	Other Debtors/ Guarantors	
	Telephone	

Table 5.4. Classification Accuracy

Model	Type I	Type II	Overall
M_1	0.195	0.916	0.426
M_2	0.478	0.841	0.595

different groups of variables, due to the collinearity, may have the same effect on the target variable, and therefore yield to similar classification results.

5.7. Discussion

We considered a Bayesian joint probabilistic inference on the action and the associated risk in the classification problem. The posterior probability is based on an empirical likelihood, which imposes a moment restriction relating the action to the resulting risk, but does not otherwise require a probability model for the underlying data generating process. As there is no need to assume the true likelihood function, such a Bayesian approach based on a moment-condition likelihood is attractive and promising in huge amount of application areas. It has been shown that this procedure works well when the sample size is large, since the empirical likelihood can be interpreted as the approximation to the true underlying likelihood function asymptotically. On the other hand, however, how our proposed posterior behaves in the finite sample case is still an open and interesting question, which will be left in the future studies.

Another important feature of our approach is that the parameters (θ, r) are not fully identified, i.e., the posterior density does not degenerate to a point probability mass, but asymptotically clusters around the curve $\{(\theta, r) : E[\rho(W, \theta)|\theta] = r\}$. Therefore we can generate the desired action θ from $P(\theta|r, Data)$, given a controlled level of risk r_0. We illustrated by examples how this method is used to describe the posterior of the actions to take in order to achieve a low risk, or conversely, to describe the posterior of the resulting risk for a given action. In addition, this approach can also be applied to model selection.

Recently, there has been a rapidly growing frequentist literature on the models where the parameters of interest are not fully identified, whereas the literature on Bayesian approach is comparatively much less. Therefore our method is also an important contribution to the Bayesian partial identification literature. However, there are many other important open questions to be understood from a pure Bayesian perspective. For example, how do we construct an objective prior when the parameter is not identified? In this case, nether the Jeffreys' prior (1946) nor the reference prior (Bernardo 1979) exist. The reason is that the information matrix is singular due to the lack of identification. These problems, of course, deserve further careful research.

References

[1] Andrews, D. (1992). Generic uniform convergence. *Econometric Theory*. **8** 241-257

[2] Andrews, D. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica*. 67 543-564

[3] Andrews, D. and Jia, P. (2008). Inference for parameters defined by moment inequalities: a recommended moment selection procedure. *Manuscript* . Yale University.

[4] Andrews, D. and Lu, B. (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models *Journal of Econometrics*. 101 123-164

[5] Andrews, D. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*. 78 119-157

[6] Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*. **71** 1795-1843

[7] Antoniadis A, Gredoire G. and McKeague, I. (2004). Bayesian estimation in single-index models. *Statistica Sinica*. **14** 1147-1164

[8] Asuncion, A. and Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[9] Beresteanu, A. and Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica*. **76** 763-814.

[10] Bernardo, J. (1979) Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society*, Series B. **41** 113-147.

[11] Billingsley, P. (1986). *Probability and Measure*, ch 16. Second Edition. Wiley, New York.

[12] Blundell R, Chen X. and Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica.* **75** 1613-1670

[13] Bollinger, C. and Hasselt, M. (2009). A Bayesian analysis of binary misclassification: inference in partially Identified models. *Manuscript*. University of Western Ontario.

[14] Bugni, F. (2010). Bootstrap inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set. *Econometrica*. **78** 735-753

[15] Canay, I. (2010). EL inference for partially identied models: large deviations optimality and Bootstrap validity. *Journal of Econometrics.* **156** 284-303

[16] Carrasco, M, Florens, J. and Renault, E. (2007) Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. in: J.J. Heckman and E.E. Leamer (ed.), *Handbook of Econometrics*. **VI** ch 77.

[17] Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics.* **34** 305-334.

[18] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. in: J.J. Heckman and E.E. Leamer (ed.), *Handbook of Econometrics*, **6**, chapter 76.

[19] Chen. X. and Pouzo, D. (2009a). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Manuscript*. Yale University.

[20] Chen. X. and Pouzo, D. (2009b). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*. **152** 46-60.

[21] Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics.* **115** 293-346

[22] Chernozhukov, V., Hong, H and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica.* **75** 1243-1284.

[23] Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*. **77** 1791-1828

[24] Cover, T. and Thomas, J. (1991). *Elements of Information Theory*, Wiley, New York.

[25] Coram, M. and Lalley, S. (2006). Consistency of Bayes estimators of a binary regression function. *The Annals of Statistics*. **34** 1233-1269.

[26] Darolles S, Fan Y, Florens J P. and Renault, E. (2010). Nonparametric Instrumental Regression. *Manuscript*. Toulouse School of Economics.

[27] Devroye, G., Gyorfi, L and Lugosi, G (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

[28] Dreze, J.H. (1976). Bayesian limited information analysis of the simultaneous equations model. *Econometrica*. **44** 1045-1075.

[29] Florens, J. and Simoni, A. (2009a). Nonparametric estimation of an instrumental regression: a quasi-Bayesian approach based on regularized posterior. *Manuscript.* Toulouse School of Economics.

[30] Florens, J. and Simoni, A. (2009b). Regularizing priors for linear inverse problems. *Manuscript.* Toulouse School of Economics.

[31] Gallant, A. and Nychka, D. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*. **55** 363-390

[32] Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science*. **20** 111-140.

[33] Gelfand, A. and Sahu, S. (1999). Identifiability, improper priors, and Gibbs sampling for generalized liner models. *Journal of the American Statistical Association*. **94** 247-253.

[34] Haitovsky, Y. and Wax, Y. (1980). Generalized ridge regression, least squares with stochastic prior information, and Bayesian estimators. *Applied Mathematics and Computation.* **7** 125-154.

[35] Han, C. and Phillips, P. (2006). GMM with many moment conditions. *Econometrica*. **74** 147-192

[36] Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*. **50** 1029-1054

[37] Haile, P. and Tamer, E. (2003). Inference with an incomplete model of English auctions. *Journal of Political Economy*. **111** 1-52.

[38] Hall, P. and Horowitz, J. (2005), Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*. **33** 2904-2929.

[39] Honore, B., Khan, S. and Powell, J. (2002). Quantile regression under random censoring. *Journal of Econometrics*. **109** 67-105.

[40] Horowitz, J. A smoothed maximum score estimator for the binary response model. *Econometrica*. **60** 505-531.

[41] Horowitz, J. and Lee, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*. **75** 1191-1208.

[42] Horowitz, J. and Manski, C. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of American Statistical Association*. **95** 77-84.

[43] Ichimura , H. (1993). Semiparametric least squares and weighted SLS estimation of single index models. *Journal of Econometrics*. **58** 71-120

[44] Imbens, G. and Manski, C. (2004). Confidence intervals for partially identified parameters. *Econometrica*. **72** 1845-1857.

[45] Imai, K. and van Dyk, D.A. (2004). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*. **124** 311-334.

[46] Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, London A. **186** 453-461.

[47] Jiang, W. (2009), On uniform deviations of general empirical risks with unboundedness, dependence, and high dimensionality. *Journal of Machine Learning Research*. **10** 977-996

[48] Jiang, W. and Tanner, M. (2008), Gibbs posterior for variable selection in high dimensional classification and data mining. *The Annals of Statistics*. **36** 2207-2231.

[49] Kim, J. (2002). Limited information likelihood and Bayesian analysis. *Journal of Econometrics*. **107** 175-193.

[50] Khan, S. and Tamer, E. (2009). Inference on endogenously censored regression models using conditional moment inequalities. *Journal of Econometrics*. **152** 104-119.

[51] Kitamura, Y. (2001). Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica*. **69** 1661-1672.

[52] Kitamura, Y. (2006). Empirical likelihood methods in econometrics: theory and practice. *Unpublished Manuscript* Yale University.

[53] Kitamura, Y., Tripathi, G. and Ahn, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*. **72** 1667-1714.

[54] Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*. **30** 1-50.

[55] Kovchegov, Y. and Yildiz, N. (2010). Inference in partially identified nonparametric instrumental variables models. *Manuscript*. University of Rochester.

[56] Kress, R. (1999). *Linear Integral Equation*. Springer

[57] Lazar, N. (2003). Bayesian empirical likelihood. *Biometrika*. **90** 319-326.

[58] Liao, Y. and Jiang, W. (2010). Bayesian analysis of moment inequality models. *The Annals of Statistics* **38** 275-316.

[59] Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics*. **31** 807-832.

[60] Manski, C. (2003). *Partial identification of probability distributions.* Springer Series in Statistics. Springer, New York.

[61] Manski, C. and Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*. **70** 519-547.

[62] Moon, H. and Schorfheide, F. (2004). Bayesian inference for econometric models using empirical likelihood functions. *Extended Abstract*. Econometric Society 2004 North American Winter Meetings 284.

[63] Moon, H. and Schorfheide, F. (2009a), Bayesian and frequentist inference in partially identified models. *Manuscript*. University of Pennsylvania.

[64] Moon, H. and Schorfheide, F. (2009b), Estimation with overidentifying inequality moment conditions. *Journal of Econometrics*. **153** 136-154

[65] Molinari, F. (2010), Missing treatments. *Journal of Business and Economic Statistics*. **28** 82-95

[66] McCulloch, R., Polson, N. and Rossi, P. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*. **99** 173-193.

[67] Mohammadi, L. and Van De Geer, S. (2005) Asymptotics in empirical risk minimization *Journal of Machine Learning Research*. **6** 2027-2047.

[68] Neath A. and Samaniego, F. (1997). On the efficacy of Bayesian inference for nonidentifiable models. *The American Statistician*. **51** 225-232.

[69] Newey, W. (1991), Uniform convergence in probability and stochastic equicontinuity. *Econometrica*. **59** 1161-1167.

[70] Newey, W. and McFadden, D. (1994), Large sample estimation and hypothesis testing. *Handbook of Econometrics*. **IV** ch 36

[71] Newey, W. and Powell, J. (2003), Instrumental variable estimation of nonparametric models. *Econometrica*. **71** 1565-1578.

[72] Newey, W. and Smith, R. (2001), Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*. **72** 219-255.

[73] Owen, A. (1990). Empirical likelihood for confidence regions. *The Annals of Statistics*. **18** 90-120.

[74] Pakes, A., J. Porter, K. Ho, and J. Ishii (2006). Moment inequalities and their application. *Manuscript*. Harvard University

[75] Polyanin, A. and Manzhirov, A. (1998). *Handbook of Integral Equation*. CRC Press, Boca Raton.

[76] Poirier, D. (1998). Revising beliefs in nonidentified models. *Econometric Theory*. **14** 483-509.

[77] Prakasa Rao, B.L.S. (1992). *Identifiability in Stochastic Models: Characterization of Probability Distributions.* Academic Press, London.

[78] Qin, J. and Lawless, J. (1994). Empirical Likelihood and general estimating equations. *The Annals of Statistics*. **22** 300-325.

[79] Romano, J. and Shaikh, A. (2008). Inference for Identifiable Parameters in Partially Identified Econometric Models. *Journal of Statistical Planning and Inference*. **138** 2786-2807.

[80] Rosen, A. (2008). Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. *Journal of Econometrics*. **146** 107-117.

[81] Santos, A. (2007). Inference in nonparametric instrumental variables with partial identification. *Manuscript*. University of California, San Diego.

[82] Santos, A. (2008a). Instrumental variables methods for recovering continuous linear functions. *Manuscript*. University of California, San Diego.

[83] Santos, A. (2008b). A Bootstrap procedure for Inference in nonparametric instrumental variables. *Manuscript*. University of California, San Diego.

[84] Severini, T. and Tripathi, G. (2006), Some identification issues in nonparametric linear models with endogenous regressors. *Econometric Theory*. **22** 258-278.

[85] Smith, R. (2007), Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics*. **138** 430-460.

[86] Tang, X. (2008), Essays in empirical auctions and partially identified econometric models. *Ph.D. Thesis*. Northwestern University.

[87] Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies*. **70** 147-165.

[88] Tamer, E. (2009). Partial identification in econometrics. *Manuscript*. Northwestern University.

[89] Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risk. *Proceedings of the National Academy Science USA*, **72** 20-22.

[90] Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis*. **4** 191-208.

[91] Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distribution. *Bayesian Inference and Decision Techniques: essays in honour of Bruno de Finetti.* Edited by Goel, P.K. and Zellner, A. pp. 233-243. Amsterdam, North Holland.

[92] Zellner, A. (1994). Model, prior information and Bayesian analysis. *Journal of Econometrics*. **75** 51-68.

APPENDIX A

Technical Proofs for Chapter 2

A.1. Proofs for Section 2.2: Theorem 2.2.1

Proof. Let $g(\Omega)^{-\epsilon} = \{x \in g(\Omega) : d(x, g(\Omega)^c) \geq \epsilon\}$, and $g(\Omega)^{+\epsilon} = \{x \in g(\Theta) : d(x, g(\Omega)) \leq \epsilon\}$. Let $\inf g(\Omega) = \inf_{\theta \in \Omega} g(\theta)$, and $\sup g(\Omega) = \sup_{\theta \in \Omega} g(\theta)$. $\forall \epsilon > 0$, we proceed by two steps: first show $\exists N \in \mathbb{N}$, when $n > N$, $\forall \epsilon > 0$,

$$g(\Omega)^{-\epsilon} \subset \hat{g}$$

and then show $\exists N \in \mathbb{N}$, when $n > N$, $\forall \epsilon > 0$, $\hat{g} \subset g(\Omega)^{+\epsilon}$.

Step I-1: show $g(\Omega) = [\inf g(\Omega), \sup g(\Omega)]$. Obviously, $g(\Omega) \subset [\inf g(\Omega), \sup g(\Omega)]$. On the other hand, $\forall x \in [\inf g(\Omega), \sup g(\Omega)]$, since Ω is compact, $\exists \theta_1, \theta_2 \in \Omega$, so that $g(\theta_1) \leq x \leq g(\theta_2)$. By assumptions, Ω is connected and g is continuous. By the intermediate value theorem, $\exists \theta^* \in \Omega$, $x = g(\theta^*)$. Hence $x \in g(\Omega)$.

Step I-2: show $\exists \theta^* \in A$, and a ball $B(\theta^*, R^*)$, so that $B(\theta^*, R^*) \subset \{\theta \in \Theta : g(\theta) \leq \inf_{\theta \in \Omega} g(\Omega)^{-\epsilon}\}$: In fact, $\forall \epsilon > 0$, it follows by step I-1, $g(\Omega)^{-\epsilon} = [\inf g(\Omega) + \epsilon, \sup g(\Omega) - \epsilon]$. Hence $\inf_{\theta \in \Omega} g(\Omega)^{-\epsilon} = \inf g(\Omega) + \epsilon$. Moreover, $\exists \theta_1 \in \Omega$, $g(\theta_1) < \inf g(\Omega) + \epsilon$. By the continuity of g, there exists a ball $B(\theta_1, R)$, such that $\forall \omega \in B(\theta_1, R)$, $g(\omega) < \inf g(\Omega) + \epsilon$. Hence $B(\theta_1, R) \subset \{\theta \in \Theta : g(\theta) \leq \inf_{\theta \in \Omega} g(\Omega)^{-\epsilon}\}$.

If $\theta_1 \in A$, then let $\theta^* = \theta_1$, $R^* = R$. If $\theta_1 \in \Omega \backslash A$, since A is dense in Ω, $B(\theta_1, \frac{R}{2}) \cap A \neq \phi$. Pick up $\theta_2 \in A \cap B(\theta_1, \frac{R}{2})$, $\forall \theta \in B(\theta_2, \frac{R}{4})$, then $d(\theta, \theta_1) \leq d(\theta, \theta_2) + d(\theta_2, \theta_1) \leq \frac{R}{4} + \frac{R}{2} < R$.

Hence $\theta \in B(\theta_1, R)$. It follows that $B(\theta_2, \frac{R}{4}) \subset B(\theta_1, R) \subset \{\theta \in \Theta : g(\theta) \leq \inf_{\theta \in \Omega} g(\Omega)^{-\epsilon}\}$, and $\theta_2 \in A$. Let $\theta^* = \theta_2$, $R^* = \frac{R}{4}$.

Step I-3: show $g(\Omega)^{-\epsilon} \subset \hat{g}$ for large n: By condition 2 in Theorem 2.2.1, for θ^*, there exists R_{θ^*}, and $N \in \mathbb{N}$, such that when $\rho < R_{\theta^*}$ and $n > N$, $P(\theta \in B(\theta^*, \rho)|X^n) > \pi_n$ w.p.a.1. Let $R_1 = \min\{R_{\theta^*}, R^*\}$, then $B(\theta^*, R_1) \subset \{\theta \in \Theta : g(\theta) \leq \inf_{\theta \in \Omega} g(\Omega)^{-\epsilon}\}$. Hence when $n > N$, $\forall x \in g(\Omega)^{-\epsilon}$,

$$F_g(x) = P(g(\theta) \leq x|X^n) \geq P(g(\theta) \leq \inf g(\Omega)^{-\epsilon}|X^n) \geq P(\theta \in B(\theta^*, R_1)|X^n) > \pi_n$$

Hence $x \geq F_g^{-1}(\pi_n)$. Likewise we can show $x \leq F^{-1}(1 - \pi_n)$. Therefore $g(\Omega)^{-\epsilon} \subset [F_g^{-1}(\pi_n), F_g^{-1}(1 - \pi_n)]$.

Step II: show for large n, $\hat{g} \subset g(\Omega)^{+\epsilon}$: Step I-1 implies $g(\Omega)^{+\epsilon} = [\inf g(\Omega) - \epsilon, \sup g(\Omega) + \epsilon]$. $\forall x \in [g(\Omega)^{+\epsilon}]^c$, either $x < \inf g(\Omega) - \epsilon$, or $x > \sup g(\Omega) + \epsilon$. If $x < \inf g(\Omega) - \epsilon$, then $\{\theta \in \Theta : g(\theta) \leq x\} \subset \{\theta \in \Theta : g(\theta) \leq \inf g(\Omega) - \epsilon\}$. In addition, since g is continuous on Θ, $\exists \delta > 0$ such that when $d(\theta, \Omega) \leq \delta$, $g(\theta) > \inf g(\Omega) - \epsilon$. Therefore $\forall \theta \in \{\theta : g(\theta) \leq \inf g(\Omega) - \epsilon\}$, $d(\theta, \Omega) > \delta$, which implies $\{\theta : g(\theta) \leq \inf g(\Omega) - \epsilon\} \subset (\Omega^c)^{-\delta}$. By condition 1 in the theorem, $\exists N \in \mathbb{N}$, when $n > N$, $P(\theta \in (\Omega^c)^{-\delta}|X^n) < \pi_n$ w.p.a.1. It follows that

$$P(g(\theta) \leq x|X^n) \leq P(g(\theta) \leq \inf g(\Omega) - \epsilon|X^n) \leq P(\theta \in (\Omega^c)^{-\delta}|X^n) < \pi_n$$

Hence $x \leq F_g^{-1}(\pi_n)$. If $x > \sup g(\Omega) + \epsilon$, by a similar argument we can show $x \geq F_g^{-1}(1 - \pi_n)$. Therefore, for $n > N$, if $x \in [F_g^{-1}(\pi_n), F_g^{-1}(1 - \pi_n)]$, then $x \in g(\Omega)^{+\epsilon}$. This implies $\hat{g} \subset g(\Omega)^{+\epsilon}$.

Combining Step I, II, since ϵ is arbitrary, $d_H(\hat{g}, g(\Omega)) \to 0$ in probability. □

A.2. Proofs for Section 2.3

Throughout the proof, we denote $\emptyset$ as the empty set, and $\mu(A)$ as the Lebesgure measure of set A.

A.2.1. Proof of Lemma 2.3.1

Proof. Recall that $(\Omega^c)^{-\epsilon} = \{\theta : d(\theta, \Omega) \geq \epsilon\}$, which is compact. $\forall \theta \in (\Omega^c)^{-\epsilon}$, $\min_j \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} < 0$. $\exists \theta^* \in (\Omega^c)^{-\epsilon}$ so that $\sup_{\theta \in (\Omega^c)^{-\epsilon}} \min_j \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} = \min_j \frac{Em_j(X,\theta^*)}{\sqrt{v_{jj}}} < 0$. Let

$$\delta = - \sup_{\theta \in (\Omega^c)^{-\epsilon}} \min_j \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} > 0$$

then $\forall \theta \in (\Omega^c)^{-\epsilon}$, $\min_j \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} \leq -\delta < -\frac{\delta}{2}$, which implies $(\Omega^c)^{-\epsilon} \subset A_{\frac{\delta}{2}}$. Hence $P(\theta \in (\Omega^c)^{-\epsilon} | X^n) \leq P(\theta \in A_{\delta/2} | X^n) = o_p(a_n)$.

A.2.2. Proof of Theorem 2.3.1

The following lemma is useful.

Lemma A.2.1. *With probability 1,*

$$P(Z \geq 0) \geq 1 - p \cdot \Phi\left(-\sqrt{n} \min_j \left\{ \frac{\bar{m}_j(\theta) - \frac{(V\psi)_j}{n}}{\sqrt{v_{jj}}} \right\}\right) \tag{A.1}$$

$$P(Z \geq 0) \leq \Phi\left(\sqrt{n} \min_j \left\{ \frac{\bar{m}_j(\theta) - \frac{(V\psi)_j}{n}}{\sqrt{v_{jj}}} \right\}\right) \tag{A.2}$$

Proof. Let $Z = (Z_1, ..., Z_p)^T$.

(B.1):

$$P(Z \geq 0) = 1 - P(\cup_{j \leq p} Z_j < 0) \geq 1 - \sum_{j=1}^{p} P(Z_j < 0) \geq 1 - \sum_{j=1}^{p} \Phi\left(-\sqrt{n}\frac{\bar{m}_j(\theta) - (V\psi)_j/n}{\sqrt{v}_{jj}}\right)$$

$$\geq 1 - p \cdot \Phi\left(-\sqrt{n}\min_j\left\{\frac{\bar{m}_j(\theta) - (V\psi)_j/n}{\sqrt{v_{jj}}}\right\}\right)$$

(B.2):

$$P(Z \geq 0) \leq \min_j P(Z_j \geq 0) = \Phi\left(\sqrt{n}\min_j\left\{\frac{\bar{m}_j(\theta) - \frac{(V\psi)_j}{n}}{\sqrt{v_{jj}}}\right\}\right)$$

Proof of Theorem 2.3.1

Proof. (1) According to Lemma 2.3.1, it suffices to show that, for any $\delta > 0$, $P(\theta \in A_\delta | X^n) = o_p(e^{-\alpha n})$, for some $\alpha > 0$. Define

$$\hat{A}_\delta = \left\{\theta : \min_j \frac{\bar{m}_j(X,\theta)}{\sqrt{v_{jj}}} < -\delta\right\}$$

Then

$$\begin{aligned} P(\theta \in A_\delta | X^n) &\propto \int_{A_\delta} p(\theta)L(\theta)d\theta \\ &= \int_{A_\delta \cap \hat{A}_\delta} p(\theta)L(\theta)d\theta + \int_{A_\delta \cap \hat{A}_\delta^c} p(\theta)L(\theta)d\theta \\ &\leq \int_{\hat{A}_\delta} p(\theta)L(\theta)d\theta + \int_{A_\delta \cap \hat{A}_\delta^c} p(\theta)L(\theta)d\theta \end{aligned}$$

$$\begin{aligned} A_\delta \cap \hat{A}_\delta^c &= \left\{\theta : \min_j \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} < -\delta\right\} \bigcap \left\{\theta : \frac{\bar{m}_i(X,\theta)}{\sqrt{v_{ii}}} \geq -\delta, i = 1, ..., p\right\} \\ &= \left(\bigcup_{j=1}^{p} \left\{\theta : \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} < -\delta\right\}\right) \bigcap \left\{\theta : \frac{\bar{m}_i(X,\theta)}{\sqrt{v_{ii}}} \geq -\delta, i = 1, ..., p\right\} \end{aligned}$$

$$\begin{array}{rl} = & \bigcup_{j=1}^{p}\left(\left\{\theta : \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} < -\delta\right\}\bigcap\left\{\theta : \frac{\bar{m}_i(X,\theta)}{\sqrt{v_{ii}}} \geq -\delta, i = 1, ..., p\right\}\right) \\ = & \cup_{j=1}^{p} A_j \end{array}$$

where

$$A_j = \left\{\theta : \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} < -\delta\right\}\bigcap\left\{\theta : \frac{\bar{m}_i(X,\theta)}{\sqrt{v_{ii}}} \geq -\delta, i = 1, ..., p\right\}$$

By weak law of large number, $A_j \to \phi$. Hence $\mu(A_j) = 0$, for any j. Then $\mu(A_\delta \cap \hat{A}_\delta^c) = \mu(\cup_j A_j) \leq \sum_j \mu(A_j) = 0$ w.p.a.1. Thus w.p.a.1, $P(\theta \in A_\delta | X^n) \leq Const \int_{\hat{A}_\delta} p(\theta)L(\theta)d\theta$. In addition, w.p.a.1, for some $\epsilon > 0$,

$$\begin{array}{rcl} L(\theta) & = & P(Z \geq 0)e^{-\psi^T\bar{m}(\theta)+\frac{1}{2n}\psi^T V\psi}\prod_i \psi_i \\ & \leq & Const \cdot P(Z \geq 0)e^{\|\psi\|(\sup_{\theta\in\Theta}\|Em(X,\theta)\|+\epsilon)+\epsilon} \\ & \leq & Const \cdot \Phi\left(\sqrt{n}\min_j \frac{\bar{m}_j(X,\theta)}{\sqrt{v_{jj}}} + O_p(\frac{1}{\sqrt{n}})\right) \end{array}$$

Therefore w.p.a.1,

$$\begin{array}{rcl} P(\theta \in A_\delta | X^n) & \leq & Const \cdot \int_{\hat{A}_\delta} p(\theta)\Phi\left(\sqrt{n}\min_j \frac{\bar{m}_j(X,\theta)}{\sqrt{v_{jj}}} + O_p(\frac{1}{\sqrt{n}})\right) d\theta \\ & \leq & Const \cdot \Phi(-\delta\sqrt{n} + O_p(\frac{1}{\sqrt{n}})) \\ & \leq & Const \cdot \Phi(-\frac{\delta}{2}\sqrt{n}) \\ & = & o_p(e^{-\frac{\delta^2}{8}n}) \end{array}$$

(2) For any integer $k > 0$, define

$$\Omega_k = \left\{\theta : \min_j \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} > \frac{1}{k}\right\}$$

$$\Omega_\infty = \left\{ \theta : \min_j Em_j(X,\theta) > 0 \right\}$$

For any $\Xi \subset \Omega$, for any integer $k > 0$ and $\forall \epsilon > 0$, by Lemma A.2.1, w.p.a.1,

$$\begin{aligned}
&\int_\Xi p(\theta)L(\theta)d\theta \geq \int_{\Xi\cap\Omega_k} p(\theta)L(\theta)d\theta \\
\geq\ & Const \int_{\Xi\cap\Omega_k} p(\theta)\left(1 - p\cdot\Phi\left(-\sqrt{n}\min_j \frac{\bar{m}_j(\theta) - (V\psi)_j/n}{\sqrt{v}_{jj}}\right)\right) d\theta \\
\geq\ & Const \int_{\Xi\cap\Omega_k} p(\theta)\left(1 - p\cdot\Phi\left(-\sqrt{n}\min_j \frac{Em_j(X,\theta) - \epsilon - (V\psi)_j/n}{\sqrt{v}_{jj}}\right)\right) d\theta \\
\geq\ & Const \int_{\Xi\cap\Omega_k} p(\theta)\left(1 - p\cdot\Phi\left(-\sqrt{n}\min_j \frac{Em_j(X,\theta)}{2\sqrt{v}_{jj}}\right)\right) d\theta \\
\geq\ & Const \int_{\Xi\cap\Omega_k} p(\theta)\left(1 - p\cdot\Phi\left(-\frac{\sqrt{n}}{2k}\right)\right) d\theta \\
\geq\ & \frac{Const}{2} P(\theta \in \Xi\cap\Omega_k)
\end{aligned}$$

Note that $\Omega_k \subset \Omega_{k+1}$, and $\bigcup_{k=1}^\infty \Omega_k = \Omega_\infty$. Hence $\lim_{k\to\infty} P(\Xi\cap\Omega_k) = P(\Xi\cap\Omega_\infty)$, which implies that, for some constant $C > 0$, w.p.a.1, $\int_\Xi p(\theta)L(\theta)d\theta \geq CP(\Xi\cap\Omega_\infty)$.

$$\begin{aligned}
P(\theta \in \Xi\cap\Omega_0) &= P(\Xi\cap\Omega_0) = P(\Xi\cap\Omega) - P(\Xi\cap(\Omega/\Omega_0)) \\
&= P(\Xi) - P(\Xi\cap(\Omega/\Omega_0)) \geq P(\Xi) - P(\Omega/\Omega_0) > 0
\end{aligned}$$

where $P(\Omega/\Omega_0) = P(\min_j Em_j(X,\theta) = 0) = 0$.

□

A.2.3. Proof of Theorem 2.3.2

Proof. In Theorem 2.2.1, let $A = int(\Omega)$, dense in Ω. $\forall\omega \in int(\Omega)$, $\exists R > 0$, such that $B(\omega, R) \subset \Omega$. Since $\pi_n \to 0$ but $P(\theta \in B(\omega,R)|X^n)$ is bounded away from 0 according to theorem 3.1-2, hence for large n, $P(\theta \in B(\omega, R)|X^n) > \pi_n$. Therefore, by Theorem 2.2.1,

$$[F_g^{-1}(\pi_n), F_g^{-1}(1-\pi_n)] \to g(\Omega) \text{ in p.}$$

□

A.2.4. Proof of Theorem 2.3.3

The following lemmas are useful:

Lemma A.2.2. *In probability,*

$$\limsup_{n\to\infty} \max_{\theta\in\Theta} \ln p(\theta|X^n) < \infty \tag{A.3}$$

$\forall\epsilon > 0$,

$$\liminf_{n\to\infty} \inf_{\theta\in\Omega^{-\epsilon}} p(\theta|X^n) > 0 \tag{A.4}$$

Proof. (A.3): For some $\epsilon > 0$,

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} L(\theta) \leq \prod_j \psi_j e^{\|\psi\|(\sup_{\theta\in\Theta}\|Em(X,\theta)\|+\epsilon)+\epsilon} < \infty$$

Thus

$$\limsup_{n\to\infty} \max_{\theta\in\Theta} \ln p(\theta|X^n) = Const \limsup_{n\to\infty} \max_{\theta\in\Theta} \ln p(\theta)L(\theta) \leq C{\cdot}\ln(\sup_{\theta\in\Theta} p(\theta){\cdot}\limsup_{n\to\infty} \sup_{\theta\in\Theta} L(\theta)) < \infty$$

(A.4): $\forall\epsilon > 0$,

$$\liminf_{n\to\infty} \inf_{\theta\in\Omega^{-\epsilon}} L(\theta) \quad \geq \quad Const \cdot \liminf_{n\to\infty} \inf_{\theta\in\Omega^{-\epsilon}} P(Z_\theta \geq 0)e^{-\|\psi\|\cdot(\sup_{\theta\in\Theta}\|Em(X,\theta)\|+\epsilon)}$$

$$\geq \quad C \cdot \liminf_{n\to\infty} \inf_{\theta\in\Omega^{-\epsilon}} P(Z_\theta \geq 0) > 0$$

Here C denotes a positive constant. The last inequality follows since $Z_\theta \sim N_p(\bar{m}(\theta) - V\psi/n, V/n)$, and $\Omega^{-\epsilon} \subset \Omega$, and $Em(X,\theta) \geq 0$ on Ω.

Lemma A.2.3. *In probability,*

(1) $\forall\epsilon > 0$,

$$\limsup_{n\to\infty} \sup_{\theta\in\Omega^{-\epsilon}} |\max_{\omega\in\Theta} \ln p(\omega|X^n) - \ln p(\theta|X^n)| < \infty$$

(2) *If* $\epsilon_n \prec n$, $\forall\epsilon > 0$,

$$\frac{\epsilon_n}{\inf_{\theta\in(\Omega^c)^{-\epsilon}} |\ln p(\theta|X^n)|} \to 0$$

Proof. (1) For each n,

$$\sup_{\theta\in\Omega^{-\epsilon}} |\max_{\omega\in\Theta} \ln p(\omega|X^n) - \ln p(\theta|X^n)| = \max_{\theta\in\Theta} \ln p(\theta|X^n) - \inf_{\theta\in\Omega^{-\epsilon}} \ln p(\theta|X^n)$$

The result follows immediate from Lemma A.2.2.

(2) w.p.a.1, $\ln p(\theta|X^n) < 0$ on $(\Omega^c)^{-\epsilon}$, hence

$$\begin{aligned}
\inf_{\theta\in(\Omega^c)^{-\epsilon}} |\ln p(\theta|X^n)| &= -\sup_{\theta\in(\Omega^c)^{-\epsilon}} \ln p(\theta|X^n) \\
&\geq -Const \cdot \ln \sup_{\theta\in(\Omega^c)^{-\epsilon}} L(\theta) \\
&\geq -C \cdot \ln \sup_{\theta\in(\Omega^c)^{-\epsilon}} P(Z_\theta \geq 0) \\
&\geq -C \cdot \ln \sup_{\theta\in(\Omega^c)^{-\epsilon}} \Phi\left(\sqrt{n} \min_j \frac{\bar{m}_j(\theta) - (V\psi)_j/n}{\sqrt{v_{jj}}}\right)
\end{aligned}$$

As shown in the proof of Lemma 2.3.1, there exists some $\delta > 0$ such that $(\Omega^c)^{-\epsilon} \subset A_\delta$, where $A_\delta = \{\theta : \min_j \frac{Em_j(X,\theta)}{\sqrt{v_{jj}}} < -\delta\}$. Thus w.p.a.1,

$$\begin{aligned} \inf_{\theta \in (\Omega^c)^{-\epsilon}} |\ln p(\theta|X^n)| &\geq -C \cdot \ln \sup_{\theta \in A_\delta} \Phi\left(\sqrt{n} \min_j \frac{\bar{m}_j(\theta) - (V\psi)_j/n}{\sqrt{v_{jj}}}\right) \\ &\geq -C \cdot \ln \sup_{\theta \in A_\delta} \Phi\left(\sqrt{n} \min_j \frac{\bar{m}_j(\theta)}{2\sqrt{v_{jj}}}\right) \\ &\geq -C \cdot \ln \Phi(-\frac{\delta}{2}\sqrt{n}) \\ &\geq -C_1 \cdot n + C_2 \ln n + C_3 \end{aligned}$$

where $C_1 > 0, C_2, C_3$ denote finite constants. This implies $\inf_{\theta \in (\Omega^c)^{-\epsilon}} |\ln p(\theta|X^n)| = O_p(n)$. □

Proof of Theorem 2.3.3

$\forall \epsilon > 0$, since $\epsilon_n \to \infty$, then by Lemma A.2.3-(1), $\exists N \in \mathbb{N}$, when $n > N$, for any $\theta \in \Omega^{-\epsilon}$,

$$\max_{\omega \in \Theta} \ln p(\omega|X^n) - \ln p(\theta|X^n) < \epsilon_n \quad w.p.a.1.$$

Therefore when $n > N$, $\Omega^{-\epsilon} \subset A_n$, which implies $\limsup_{n\to\infty} \sup_{\theta \in \Omega} d(\theta, A_n) \leq \epsilon$.

On the other hand, let $M = \liminf_{n\to\infty} \max_{\theta\in\Theta} \ln p(\theta|X^n)$. By (A.3) in Lemma A.2.2, $M < \infty$. Moreover, by (A.4),

$$M \geq \liminf_{n\to\infty} \inf_{\theta \in \Omega^{-\epsilon}} \ln p(\theta|X^n) \geq \ln \liminf_{n\to\infty} \inf_{\theta \in \Omega^{-\epsilon}} p(\theta|X^n) > -\infty$$

Hence $M \in \mathbb{R}$, and by the definition of M, $\exists N_1 \in \mathbb{N}$, when $n > N_1$,

$$\max_{\theta \in \Theta} \ln p(\theta|X^n) > M - \epsilon$$

In addition, $\forall\theta \in (\Omega^c)^{-\epsilon}$, $p(\theta|X^n) \to 0$ in probability. Thus for large n, $\ln p(\theta|X^n) < 0$ on $(\Omega^c)^{-\epsilon}$. $\exists N_2 \in \mathbb{N}$, when $n > N_2$,

$$\inf_{\theta\in(\Omega^c)^{-\epsilon}} |\ln p(\theta|X^n)| = -\sup_{\theta\in(\Omega^c)^{-\epsilon}} \ln p(\theta|X^n) > \epsilon_n - (M - \epsilon)$$

where the inequality is followed by lemma A.2.3-(2). Therefore when $n > N_2$.

$$\sup_{\theta\in(\Omega^c)^{-\epsilon}} \ln p(\theta|X^n) < -\epsilon_n + (M - \epsilon) \tag{A.5}$$

However, when $n > \max\{N_1, N_2\}$, $\forall\theta \in A_n = \{\theta : \max_{\omega\in\Theta} \ln p(\omega|X^n) - \ln p(\theta|X^n) \leq \epsilon_n\}$, $\ln p(\theta|X^n) \geq \max_{\omega\in\Theta} \ln p(\omega|X^n) - \epsilon_n > M - \epsilon - \epsilon_n$. Compared to (A.5), $\theta \notin (\Omega^c)^{-\epsilon}$. In other words, $d(\theta, \Omega) < \epsilon$. It follows that

$$\limsup_{n\to\infty} \sup_{\theta\in A_n} d(\theta, \Omega) \leq \epsilon$$

Since ϵ is arbitrary, $d_H(A_n, \Omega) \to 0$ in probability.

□

A.3. Proofs for Section 2.4

Define $A_\delta = \{\theta : Em_2(X,\theta)^T V_2^{-1} Em_2(X,\theta) > \delta\}$, and

$$A_2 = \left\{\theta : Em_2(X,\theta) = 0, \min_j Em_{1j}(X,\theta) < 0\right\}$$

Lemma A.3.1. $\forall\delta > 0$, *for some* $a > 0$

$$\int_{A_\delta\cup A_2} p(\theta)L(\theta)d\theta = o_p(e^{-an})$$

Proof. Define $\hat{A}_\delta = \{\theta : \bar{m}_2(\theta)^T V_2^{-1} \bar{m}_2(\theta) > \delta\}$,

$$\begin{aligned}\int_{A_\delta} p(\theta)L(\theta)d\theta &= \int_{A_\delta \cap \hat{A}_\delta} p(\theta)L(\theta)d\theta + \int_{A_\delta \cap \hat{A}_\delta^c} p(\theta)L(\theta)d\theta \\ &\leq \int_{\hat{A}_\delta} p(\theta)L(\theta)d\theta + \int_{A_\delta \cap \hat{A}_\delta^c} p(\theta)L(\theta)d\theta\end{aligned}$$

$A_\delta \cap \hat{A}_\delta^c = \{\theta : Em_2(X,\theta)^T V_2^{-1} Em_2(X,\theta) > \delta\} \cap \{\theta : \bar{m}_2(\theta)^T V_2^{-1} \bar{m}_2(\theta) \leq \delta\} \to \phi$ w.p.a.1. Hence for large n, $\mu(A_\delta \cap \hat{A}_\delta^c) = 0$. Then $\exists N$, when $n > N$, w.p.a.1,

$$\int_{A_\delta} p(\theta)L(\theta)d\theta \leq \int_{\hat{A}_\delta} p(\theta)\frac{\prod_i \psi_i}{\sqrt{\det(V_2)}} e^{-\frac{n}{2}\bar{m}_2(\theta)^T V_2^{-1} \bar{m}_2(\theta) - \psi^T(\Sigma_1^{-1}\Sigma_3^T \bar{m}_2(\theta) + \bar{m}_1(\theta)) + \frac{1}{2n}\psi^T \Sigma_1^{-1}\psi} d\theta$$

For some $\epsilon > 0$, for large n,

$$e^{-\psi^T(\Sigma_1^{-1}\Sigma_3^T \bar{m}_2(\theta) + \bar{m}_1(\theta)) + \frac{1}{2n}\psi^T \Sigma_1^{-1}\psi} \leq e^{\|\psi\|(\sup_{\theta\in\Theta}\|\Sigma_1^{-1}\Sigma_3^T Em_2(X,\theta) + Em_1(X,\theta)\| + \epsilon) + \epsilon} < \infty$$

Thus for some positive constant C, and large n,

$$\int_{A_\delta} p(\theta)L(\theta)d\theta \leq C \cdot \int_{\hat{A}_\delta} p(\theta)e^{-\frac{n}{2}\bar{m}_2(\theta)^T V_2^{-1} \bar{m}_2(\theta)} d\theta \leq C \cdot e^{-\frac{\delta}{2}n}$$

In addition, $\mu(A_2) = 0$ and $p(\theta)L(\theta)$ is bounded on Θ, hence

$$\int_{A_\delta \cup A_2} p(\theta)L(\theta)d\theta \leq \int_{A_\delta} p(\theta)L(\theta)d\theta + \int_{A_2} p(\theta)L(\theta)d\theta = O_p(e^{-\frac{\delta}{2}n})$$

□

Lemma A.3.2. $\forall \delta > 0$, *for some* $a > 0$,

$$\int_{(\Omega^c)^{-\delta}} p(\theta)L(\theta)d\theta = o_p(e^{-an})$$

Proof. : $\forall \theta \in (\Omega^c)^{-\delta}$, then either $\exists \delta(\theta) > 0, \theta \in A_{\delta(\theta)}$, or $\theta \in A_2$, hence $\theta \in A_{\delta(\theta)} \cup A_2$. Thus $(\Omega^c)^{-\delta} \subset \bigcup_{\theta \in (\Omega^c)^{-\delta}} [A_{\delta(\theta)} \cup A_2]$. Note that $(\Omega^c)^{-\delta} = \{\theta : d(\theta, \Omega) \geq \delta\}$ is compact, hence $\exists \ \{A_{\delta 1} \cup A_2, ..., A_{\delta N} \cup A_2\} \subset \{A_{\delta(\theta)} \cup A_2 : \theta \in (\Omega^c)^{-\delta}\}$ such that

$$(\Omega^c)^{-\delta} \subset \bigcup_{i=1}^{N} [A_{\delta i} \cup A_2]$$

Let $\delta^* = \min\{\delta_i, i = 1, ..., N\}$. For $a > b > 0$, $A_a \subset A_b$, hence $(\Omega^c)^{-\delta} \subset A_{\delta^*} \cup A_2$. Therefore

$$\int_{(\Omega^c)^{-\delta}} p(\theta) L(\theta) d\theta \leq \int_{A_{\delta *} \cup A_2} p(\theta) L(\theta) d\theta = o_p(a^{-an})$$

□

Lemma A.3.3. *If Z_θ follows $N_r(\bar{m}_1(\theta) + \Sigma_1^{-1} \Sigma_3^T \bar{m}_2(\theta) - \frac{1}{n} \Sigma_1^{-1} \psi, \frac{1}{n} \Sigma_1^{-1})$, then $\forall \omega \in \Xi$, $\exists R > 0$, in probability*

$$\liminf_{n \to \infty} \inf_{\theta \in B(\omega, R)} P(Z_\theta \geq 0) > 0$$

Proof. Let $\xi_n(\theta) = \bar{m}_1(\theta) + \Sigma_1^{-1} \Sigma_3^T \bar{m}_2(\theta) - \frac{1}{n} \Sigma_1^{-1} \psi$. $\forall \omega \in \Xi$, $Em_1(X, \omega) > 0$. Since $Em_1(X, \theta)$ is continuous on Θ, there exist $\epsilon > 0$, and an open ball $B(\omega, R_1)$, such that $\inf_{\theta \in B(\omega, R_1)} Em_1(X, \theta) > \epsilon$, where the inequality is taken coordinately. Moreover, $Em_2(X, \omega) = 0$; hence by the continuity of $Em_2(X, .)$, $\exists R < R_1$ such that $\sup_{\theta \in B(\omega, R)} |\Sigma_1^{-1} \Sigma_3^T Em_2(X, \theta)| < \epsilon$, where $|.|$ denotes the absolute value, taken coordinately. Therefore, $\inf_{\theta \in B(\omega, R)} (Em_1(X, \theta) + \Sigma_1^{-1} \Sigma_3^T Em_2(X, \theta)) > 0$. $\exists N$, when $n > N$, w.p.a.1, coordinately.

$$\inf_{\theta \in B(\omega, R)} (\bar{m}_1(\theta) + \Sigma_1^{-1} \Sigma_3^T \bar{m}_2(\theta) - \frac{1}{n} \Sigma_1^{-1} \psi) = \inf_{\theta \in B(\omega, R)} \xi_n(\theta) > 0$$

Let σ_{1j}^2 denote the jth diagonal element in Σ_1^{-1}, and $\xi_{nj}(\theta)$ denote the jth element of $\xi_n(\theta)$. Then

$$\begin{aligned}
\inf_{\theta\in B(\omega,R)} P(Z_\theta \geq 0) &\geq 1 - r\cdot\Phi\left(-\sqrt{n}\inf_{\theta\in B(\omega,R)}\min_j \frac{\xi_{nj}(\theta)}{\sqrt{\sigma_{1j}^2}}\right)\\
&\geq 1 - r\cdot\Phi\left(-\sqrt{n}\min_j \frac{\inf_{\theta\in B(\omega,R)}\xi_{nj}(\theta)}{\sqrt{\sigma_{1j}^2}}\right)\\
&>_n 0
\end{aligned}$$

Lemma A.3.4. *For any $\beta_n \to \infty$, $\forall\omega\in\Xi$, $\exists R>0$, $\forall\delta<R$, w.p.a.1,*

$$\int_{B(\omega,\delta)} p(\theta)L(\theta)d\theta \succ \frac{1}{\beta_n}n^{-d/2}$$

where $d = \dim(\theta)$.

Proof. $\forall\omega\in\Xi$, it can be shown that (using Lemma A.3.3), $\exists R>0$, and a positive constant C, such that $\int_{B(\omega,R)} p(\theta)L(\theta)d\theta \geq C\int_{B(\omega,R)} p(\theta)e^{-\frac{n}{2}\bar{m}_2(\theta)^T V_2^{-1}\bar{m}_2(\theta)}d\theta$. For deterministic V_2^{-1}, and a vector α, write weighted norm $\|\alpha\|_V^2 = \alpha^T V_2^{-1}\alpha$. Then we have

$$\frac{1}{2}\|\bar{m}_2(\theta)\|_V^2 \leq \|Em_2(X,\theta)\|_V^2 + \|\bar{m}_2(\theta) - Em_2(X,\theta)\|_V^2$$

By Assumption 2.4.3, for any $\beta_n \to \infty$, choose $\beta_n^{\|V^{-1}\|^{-1}}$, so that

$$e^{-n\|\bar{m}_2(\theta)-Em_2(X,\theta)\|_V^2} \geq e^{-\ln\beta_n^{\|V^{-1}\|^{-1}}\|V^{-1}\|} = \beta_n^{-\|V^{-1}\|^{-1}\cdot\|V^{-1}\|} = \frac{1}{\beta_n}$$

For some constant $\alpha > 1$, let $U = \{\theta : \|Em_2(X,\theta)\|_V^2 < \frac{\ln\alpha}{n}\}$. By Assumption 2.4.4, $\exists R' < R$, such that for any $0 < \delta < R'$, $\inf_{\theta\in B(\omega,\delta)} p(\theta) > 0$. Then

$$\int_{B(\omega,\delta)} p(\theta)L(\theta)d\theta \geq C\int_{B(\omega,\delta)\cap U} \frac{1}{\alpha\beta_n} p(\theta)d\theta \geq \frac{Const}{\beta_n}\mu(B(\omega,\delta)\cap U)$$

To derive a lower bound for the Lebesgue measure of $B(\omega,\delta)\cap U$, note that $Em_2(X,\theta)$ is Lipschitz continuous, and $Em_2(X,\omega) = 0$, $\exists\ \lambda > 0$, such that $\forall\theta \in B(\omega,\delta)$, $\|Em_2(X,\theta)\|^2 \leq \lambda\|\theta-\omega\|^2$. Then

$$\begin{aligned}\|Em_2(X,\theta)\|_V^2 &\leq \|Em_2(X,\theta)\|^2 \cdot \|V^{-1}\|^2 \\ &\leq \lambda\|V^{-1}\|^2 \cdot \|\theta-\omega\|^2 \\ &\leq \lambda\|V^{-1}\|^2|\theta-\omega|_\infty^2\end{aligned}$$

where $|\theta-\omega|_\infty = \max_j|\theta_j-\omega_j|$. Hence $\{\theta : |\theta-\omega|_\infty^2 < \frac{\ln\alpha}{\lambda\|V^{-1}\|^2 n}\} \subset U$. Moreover, for large enough n, $\{\theta : |\theta-\omega|_\infty^2 < \frac{\ln\alpha}{\lambda\|V^{-1}\|^2 n}\} \subset B(\omega,\delta)$, thus $\mu(B(\omega,\delta)\cap U) \geq \mu(\{\theta : |\theta-\omega|_\infty^2 < \frac{\ln\alpha}{\lambda\|V^{-1}\|^2 n}\}) = (2\sqrt{\frac{\ln\alpha}{\lambda\|V^{-1}\|^2}})^d n^{-d/2}$. Hence $\int_{B(\omega,\delta)} p(\theta)L(\theta)d\theta \succ \frac{1}{\beta_n}n^{-d/2}$.

Proof of Theorem 2.4.1

(1) Let $\beta_n = n^{d/2}$. Lemma A.3.4 implies that $\int_\Theta p(\theta)L(\theta)d\theta \succ n^{-d}$. Thus by Lemma A.3.2, for some $\alpha > 0$,

$$P(\theta\in(\Omega^c)^{-\delta}|X^n) = \frac{\int_{(\Omega^c)^{-\delta}} p(\theta)L(\theta)d\theta}{\int_\Theta p(\theta)L(\theta)d\theta} \prec \frac{o_p(e^{-\alpha n})}{n^{-d}} = o_p(e^{-\frac{\alpha}{2}n})$$

(2) The result follows immediately from Lemma A.3.4 and that $\int_\Theta p(\theta)L(\theta)d\theta$ is bounded. Q.E.D.

Proof of Theorem 2.4.2: It follows immediately from Theorem 2.4.1 and Theorem 2.3.2.

APPENDIX B

Technical Proofs for Chapter 3

B.1. Proof of Lemma 3.2.1

I proceed by proving $(ii) \Rightarrow (iii) \Rightarrow (i) \Rightarrow (ii)$.

$(ii) \Rightarrow (iii)$: By (ii), there exists θ_1 such that $EM_s(X, \theta_1) \geq 0$. Let $\lambda_1 = EM_s(X, \theta_1)$, then $(EM_s(X, \theta_1) - \lambda_1)^T V_0 (EM_s(X, \theta_1) - \lambda_1) = 0$. (iii) follows immediately.

$(iii) \Rightarrow (i)$: $\forall V_0 > 0$, if (iii) holds, then there exists θ_n and $\lambda_n \geq 0$ such that $(EM_s(X, \theta_n) - \lambda_n)^T V_0 (EM_s(X, \theta_n) - \lambda_n) = o(1)$. Since $V_0 > 0$, we have $EM_s(X, \theta_n) - \lambda_n = o(1)$, which implies (i).

$(i) \Rightarrow (ii)$: Since C_s is compatible, there exist $\{\theta_n\}$ and $\lambda_n \geq 0$ for all n, such that $\|EM_s(X, \theta_n) - \lambda_n)\| = o(1)$. Since $EM_s(X, \theta)$ is continuous on Θ_s and Θ_s is compact, $\{EM_s(X, \theta_n)\}_{n \geq 1}$ is bounded, which implies that $\{\lambda_n\}_{n \geq 1}$ is also bounded. Therefore there exist subsequences $\{\theta_{n_k}\}$ and $\{\lambda_{n_k}\}$ and $\lambda_0 \geq 0$, θ_0, such that $\theta_{n_k} \to \theta_0$, and $\lambda_{n_k} \to \lambda_0$. Since Θ is compact, $\theta_0 \in \Theta_s$. It follows from continuity that $EM_s(X, \theta_{n_k}) \to EM_s(X, \theta_0)$. Therefore, for all $\epsilon > 0$, for all large enough k,

$$||EM_s(X, \theta_0) - \lambda_0|| \leq ||EM_s(X, \theta_0) - EM_s(X, \theta_{n_k})|| + ||EM_s(X, \theta_{n_k}) - \lambda_{n_k}|| + ||\lambda_{n_k} - \lambda_0|| < \epsilon$$

Since ϵ is arbitrary, $EM_s(X, \theta_0) = \lambda_0$ with $\theta_0 \in \Theta_s$ and $\lambda_0 \geq 0$, which yields (ii). Q.E.D.

B.2. Proof of Theorem 3.4.1

Let

$$L(X^n|\theta_s,\lambda_s,C_s)=\int_{\mathbb{R}^{p-m}}L(X^n|\theta_s,\lambda,C_s)p(\lambda_s^c|C_s)d\lambda_s^c \tag{B.1}$$

$$L(X^n|\theta_s,C_s)=\int_{[0,\infty)^m}L(X^n|\theta_s,\lambda_s,C_s)p(\lambda_s|C_s)d\lambda_s \tag{B.2}$$

For any candidate C_s,

$$\begin{aligned}
p(C_s|X^n) &\propto p(C_s)\iiint_{\Theta_s\times[0,\infty)^m\times\mathbb{R}^{p-m}}L(X^n|\theta_s,\lambda,C_s)p(\theta_s|C_s)p(\lambda_s|C_s)p(\lambda_s^c|C_s)d\theta_s d\lambda_s d\lambda_s^c\\
\text{(B.3)}\qquad &\propto p(C_s)\int_{\Theta_s}\int_{\lambda_s\geq 0}L(X^n|\theta_s,\lambda_s,C_s)p(\lambda_s|C_s)d\lambda_s d\theta_s\\
\text{(B.4)}\qquad &\propto p(C_s)\int_{\Theta_s}L(X^n|\theta_s,C_s)p(\theta_s|C_s)d\theta_s
\end{aligned}$$

A tedious calculation shows that

$$L(X^n|\theta_s,\lambda_s,C_s)=(2\pi)^{-p/2}\det(S)^{-1/2}\exp\left\{-\frac{1}{2}(\bar{M}_s(\theta)-\lambda_s,\bar{M}_s^c(\theta))S^{-1}\begin{pmatrix}\bar{M}_s(\theta)-\lambda_s\\ \bar{M}_s^c(\theta)\end{pmatrix}\right\}$$

where

$$S=\frac{V}{n}+\begin{pmatrix}0&0\\0&\Sigma\end{pmatrix}\text{ write }S^{-1}=n\begin{pmatrix}\Sigma_1&\Sigma_3\\ \Sigma_3^T&\Sigma_2\end{pmatrix}$$

Assume $\|V\|=O(1),\|\Sigma\|=O(1)$. It follows that $\|\Sigma_1\|=O(1)$, $\|\Sigma_3\|\leq O(n^{-1})$, and $\|\Sigma_2\|=O(n^{-1})$.

When C_s is incompatible, since Σ_1 is positive definite, hence there exits $\tau>0$, w.p.a.1, $\inf_{\lambda_s\geq 0,\theta_s\in\Theta_s}(\bar{M}_s(\theta_s)-\lambda_s)^T\Sigma_1(\bar{M}_s-\lambda_s)>2\tau$. By (B.3), there exists a constant $C>0$, w.p.a.1, $p(C_s|X^n)\leq Cn^{m/2}e^{-n\tau}\int_{\Theta_s}\exp(-\frac{n}{2}\bar{M}_s^{cT}(\theta_s)\Sigma_3^T\bar{M}_s(\theta_s))\int_{\lambda_s\geq 0}e^{-\frac{n}{2}\bar{M}_s{}^c(\theta_s)\Sigma_3^T\lambda_s}p(\lambda_s|C_s)d\lambda_s d\theta_s$.

Since Θ is compact, $Em(X, \theta)$ is continuous on Θ, $\|\Sigma_3\| \leq O(n^{-1})$, and $p(\lambda_s|C_s)$ is exponential with fixed parameter ψ, $p(C_s|X^n) = O_p(n^{m/2}e^{-n\tau})$, which is exponentially small.

When C_s is compatible, we can then calculate (B.2), and use (B.4):

$$L(X^n|\theta_s, C_s) = (2\pi)^{(m-p)/2}n^{(p-m)/2}\det(V_2 + n\Sigma)^{-1/2}P(Z_\theta \geq 0)e^{\tau(\theta)}\prod \psi_i, \text{ where}$$

- V_2 is the lower diagonal block of V.
- $Z_\theta \sim N_m(\bar{M}_s(\theta) + \Sigma_1^{-1}\Sigma_3^T\bar{M}_s^c(\theta) - \frac{1}{n}\Sigma_1^{-1}\psi, \frac{\Sigma_1^{-1}}{n})$
- $\tau(\theta) = -\frac{n}{2}\bar{M}_s^c(\theta)(V_2 + n\Sigma)^{-1}\bar{M}_s^c(\theta) - \psi^T[\Sigma_1^{-1}\Sigma_3\bar{M}_s^c(\theta) + \bar{M}_s(\theta)] + \frac{1}{2n}\psi^T\Sigma_1^{-1}\psi$

Note that $\det(V_2 + n\Sigma) = O(n^{p-m})$, hence for some constant $C = O(1)$, we can write $L(X^n|\theta_s, C_s) = CP(Z_\theta \geq 0)e^{\tau(\theta)}$.

Lemma B.2.1. *For any $C_s = (M_s, \Theta_s)$, let $\Omega_s = \{\theta \in \Theta_s : EM_s(X, \theta) \geq 0\}$ (for simplicity, write (M_s, Θ_s) to represent (M_{s_1}, Θ_{s_2})). In probability,*

$$\text{(B.5)} \qquad \lim_{n\to\infty} p(C_s|X^n) = Cp(C_s)\int_{\Omega_s} p(\theta_s|C_s)e^{-\psi^T EM_s(\theta) - \frac{1}{2}EM_s^c(\theta)\Sigma^{-1}EM_s^c(\theta)}d\theta$$

Proof. $p(C_s|X^n) = Cp(C_s)\int_{\Theta_s} p(\theta|C_s)p(Z_\theta \geq 0)e^{\tau(\theta)}d\theta$. Let RHS denote the right-hand side of (B.5).

$$|p(C_s|X^n) - RHS| \leq Cp(C_s)\int_{\Theta_s} p(\theta|C_s)\left|1_{\Omega_s}e^{-\psi^T EM_s(\theta) - \frac{1}{2}EM_s^c(\theta)\Sigma^{-1}EM_s^c(\theta)} - p(Z_\theta \geq 0)e^{\tau(\theta)}\right| d\theta$$

Denote $\Delta(\theta) = p(\theta|C_s)\left|1_{\Omega_s}e^{-\psi^T EM_s(\theta) - \frac{1}{2}EM_s^c(\theta)\Sigma^{-1}EM_s^c(\theta)} - p(Z_\theta \geq 0)e^{\tau(\theta)}\right|$. Write $|p(C_s|X^n) - RHS| \leq Cp(C_s)\left(\int_{U_1}\Delta(\theta)d\theta + \int_{U_2}\Delta(\theta)d\theta + \int_{U_3}\Delta(\theta)d\theta\right)$, where

$$U_1 = \{\theta \in \Theta_s : EM_s(X, \theta) > 0\}$$

$$U_2 = \{\theta \in \Theta_s : EM_s(X,\theta) \geq 0, Em_j(X,\theta) = 0 \text{ for some } m_j \in M_s\}$$

$$U_3 = \{\theta \in \Theta_s : \text{ for some } m_j \in M_s, Em_j(X,\theta) < 0\}$$

We look at the integrations on U_i, $i = 1, 2, 3$ subsequently.

U_1**:** Note that $\Omega_s = \{\theta \in \Theta_s : EM_s(X,\theta) \geq 0\}$, and $Z_\theta \sim N_m(\bar{M}_s(\theta)+\Sigma_1^{-1}\Sigma_3^T\bar{M}_s^c(\theta)-\frac{1}{n}\Sigma_1^{-1}\psi, \frac{\Sigma_1^{-1}}{n})$. Note that $\|\Sigma_1\| = O(1)$, and $\|\Sigma_3\| \leq O(n^{-1})$. For any $\epsilon > 0$, by uniform weak law of large number, w.p.a.1, $\sup_{\theta \in U_1} |P(Z_\theta \geq 0) - 1| < \epsilon$. Hence for large n w.p.a.1, $\sup_{\theta\in U_1} \left|1_{\Omega_s} e^{-\psi^T EM_s(\theta) - \frac{1}{2}EM_s^c(\theta)\Sigma^{-1}EM_s^c(\theta)} - p(Z_\theta \geq 0)e^{\tau(\theta)}\right| < \epsilon$. It follows that

$$\int_{U_1} \Delta(\theta)d\theta \leq \epsilon \int_{U_1} p(\theta|C_s)d\theta \leq \epsilon$$

U_2**:** The Lebesgue measure of $U_2 = 0$.

U_3**:** $\forall\theta \in U_3$, $1_{\theta\in\Omega_s} = 0$, hence $\Delta(\theta) = p(\theta|C_s)P(Z_\theta \geq 0)e^{\tau(\theta)}$. $\forall\epsilon > 0$, w.p.a.1, $\sup_{U_3} P(Z_\theta \geq 0) < \epsilon$, thus for large n, $\int_{U_3} \Delta(\theta)d\theta \leq \epsilon \int_{U_3} p(\theta|C_s)e^{\tau(\theta)}d\theta \leq \epsilon O_p(1)$.

Therefore $|p(C_s|X^n) - RHS| < Const \cdot p(C_s)\epsilon$, w.p.a.1, with arbitrarily small ϵ. □

Now back to the proof of Theorem 3.4.1. When C_s is compatible, by the previous lemma, w.p.a.1, $p(C_s|X^n) \geq Cp(C_s)e^{-\frac{1}{2}\sup_{\theta\in\Theta_s}(\|EM(X,\theta_s)\|^2\|\Sigma^{-1}\|+\|\psi\|\|EM(X,\theta_s)\|)}P(\theta \in \Omega_s|C_s)$. Assuming $P(\theta \in \Omega_s|C_s) > 0$, it follows that $p(C_s|X^n) = Const \times p(C_s)$ for some $Const > 0$.

B.3. Proof of Theorem 3.4.2

Let W be a p-dimensional random vector whose conditional distribution given (X^n, θ) is $N_p(\bar{M}(\theta), V/n)$. It then follows that $L(X^n|\theta_s, \lambda, C_s)$ is equal to the density function of $W|X^n, \theta$. Therefore $\int_{\lambda\in\Lambda} L(X^n|\theta_s, \lambda, C_s)p(\theta_s, \lambda|C_s)d\lambda = E_W[p(\theta_s, W|C_S)I(W \in \Lambda)|X^n, \theta]$.

Define $Z = \sqrt{n}V^{-1/2}(W - \bar{M}(\theta))$. It then follows that $Z|X^n, \theta \sim N_p(0, I)$. Hence

$$\begin{aligned}
&E_W[p(\theta_s, W|C_S)I(W \in \Lambda)|X^n, \theta] \\
&= E_Z[p_{\theta,\lambda}(\theta_s, n^{-1/2}V^{1/2}Z + \bar{M}(\theta)|C_S)I(n^{-1/2}V^{1/2}Z + \bar{M} \in \Lambda)|X^n, \theta] \\
&\equiv \Xi_n(X^n, \theta)
\end{aligned}$$

Let $\Omega(\Theta, \Lambda) = \{\theta \in \Theta : EM(X, \theta) \in \Lambda\}$, and $E_{X^n}(.)$ denote the expectation operator taken with respect to the distribution of X^n.

Lemma B.3.1. *For any $x^n = (x_1, .., x_n)$ in the support of X^n,*
$\lim_n \int_\Theta \Xi_n(x^n, \theta)d\theta = \int_{\Omega(\Theta,\Lambda)} p_{\theta,\lambda}(\theta_s, EM(X, \theta)|C_S)d\theta.$

Proof. Define $A_n(\theta, z) = p_{\theta,\lambda}(\theta_s, n^{-1/2}V^{1/2}z + \bar{M}(\theta)|C_S)I(n^{-1/2}V^{1/2}z + \bar{M} \in \Lambda)$. Hence $\Xi_n(X^n, \theta) = E_Z[A_n(\theta, Z)|X^n, \theta] = \int A_n(\theta, z) \prod \phi(z_i)dz$, where $\phi(z_i)$ is the probability density function of standard normal distribution, and $z = (z_1, ..., z_p)$. By Assumption 3.4.2, for all fixed x^n, $A_n(\theta, z) \leq g(\theta)$ for all z, and $\int\int g(\theta) \prod \phi(z_i)dzd\theta < \infty$. Since $p_{\theta,\lambda}(\theta, \lambda)$ is continuous with respect to λ, $\lim_n A_n(\theta, z) = p_{\theta,\lambda}(\theta, EM(X, \theta)|C_s)I(EM(X, \theta) \in \Lambda)$. Apply the dominated convergence theorem with fixed x^n:

$$\lim_n \int_\Theta \Xi_n(x^n, \theta)d\theta = \lim_n \int\int A_n(\theta, z) \prod \phi(z_i)dzd\theta = \int_{\Omega(\Theta,\Lambda)} p_{\theta,\lambda}(\theta_s, EM(X, \theta)|C_S)d\theta$$

Q.E.D.

Finally, since

$$p(C_s|X^n) \propto p(C_s) \iint_{\Theta_s, \lambda \in \Lambda} L(X^n|\theta_s, \lambda, C_s)p(\theta_s, \lambda|C_s)d\theta_s d\lambda = p(C_s) \int_{\Theta_s} \Xi_n(X^n, \theta_s)d\theta_s$$

and $\forall x^n$, $p(C_s|X^n = x^n) \leq Const \times p(C_s) \int g(\theta)d\theta$, we can apply the dominated convergence theorem again to obtain:

$$E_{X^n}\left|p(C_s|X^n) - Const \times p(C_s)\int_{\Omega(\Theta,\Lambda)} p_{\theta,\lambda}(\theta_s, EM(X,\theta)|C_S)d\theta\right| \rightarrow 0$$

which implies $p(C_s|X^n) \rightarrow^p Const \times p(C_s)\int_{\Omega(\Theta,\Lambda)} p_{\theta,\lambda}(\theta_s, EM(X,\theta)|C_S)d\theta$ in probability. Q.E.D.

APPENDIX C

Technical Proofs for Chapter 4

C.1. Proofs for Section 4.2

C.1.1. Proof of Theorem 4.2.1

(i) By definition, $G_{k_n}(g) = Em_n(g,Z)^T V_0^{-1} Em_n(g,Z)$. Since V_0 is diagonal, it is straightforward to show that

$$G_{k_n}(g) = \sum_{j=1}^{k_n} \frac{[E(\rho(Z,g)1_{W\in R_j^n})]^2}{E(\rho(Z,g_0)^2 1_{W\in R_j^n})} = \sum_{j=1}^{k_n} \int_{R_j^n} \frac{[E(\rho(Z,g)1_{W\in R_j^n})]^2}{E(\rho(Z,g_0)^2 1_{W\in R_j^n}) P(W\in R_j^n)} dF_W(w)$$

Also,

$$G(g) = \int_a^b \frac{[E(\rho(Z,g)|W=w)]^2}{E(\rho(Z,g_0)^2|W=w)} dF_W(w) = \sum_{j=1}^{k_n} \int_{R_i^n} \frac{[E(\rho(Z,g)|W=w)]^2}{E(\rho(Z,g_0)^2|W=w)} dF_W(w)$$

It follows that for all $g\in\Theta$

$$\begin{aligned}
|G_{k_n}(g) - G(g)| \quad &\leq \quad \sum_{j=1}^{k_n} \int_{R_j^n} \left| \frac{[E(\rho(Z,g)1_{W\in R_j^n})]^2}{E(\rho(Z,g_0)^2 1_{W\in R_j^n}) P(W\in R_j^n)} - \frac{[E(\rho(Z,g)|W=w)]^2}{E(\rho(Z,g_0)^2|W=w)} \right| dF_W(w) \\
&\leq \quad \sup_{1\leq j\leq k_n} \sup_{w\in R_j^n} \left| \frac{[E(\rho(Z,g)1_{W\in R_j^n})]^2}{E(\rho(Z,g_0)^2 1_{W\in R_j^n}) P(W\in R_j^n)} - \frac{[E(\rho(Z,g)|W=w)]^2}{E(\rho(Z,g_0)^2|W=w)} \right| \\
&\leq \quad \sup_{1\leq j\leq k_n} \sup_{w\in R_j^n} \left| \frac{E(\rho(Z,g)1_{W\in R_j^n})}{\sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n}) P(W\in R_j^n)}} + \frac{E(\rho(Z,g)|W=w)}{\sqrt{E(\rho(Z,g_0)^2|W=w)}} \right| \\
&\times \quad \sup_{1\leq j\leq k_n} \sup_{w\in R_j^n} \left| \frac{E(\rho(Z,g)1_{W\in R_j^n})}{\sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n}) P(W\in R_j^n)}} - \frac{E(\rho(Z,g)|W=w)}{\sqrt{E(\rho(Z,g_0)^2|W=w)}} \right| \\
&= \quad A(g)\times B(g), \text{ say.}
\end{aligned}$$

Show $\sup_{g\in\Theta} A(g) < \infty$:

$$\begin{aligned}
\sup_{g\in\Theta} A(g) &\leq \sup_{g\in\Theta}\sup_{1\leq j\leq k_n}\sup_{w\in R_j^n}\left\{\frac{|E(\rho(Z,g)1_{W\in R_j^n})|}{\sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n})P(W\in R_j^n)}} + \frac{|E(\rho(Z,g)|W=w)|}{\sqrt{E(\rho(Z,g_0)^2|W=w)}}\right\} \\
&\leq \sup_{g\in\Theta}\sup_{1\leq j\leq k_n}\frac{|E(\rho(Z,g)1_{W\in R_j^n})|/P(W\in R_j^n)}{\sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n})/P(W\in R_j^n)}} + \sup_{g\in\Theta}\sup_{w\in[0,1]}\frac{|E(\rho(Z,g)|W=w)|}{\sqrt{E(\rho(Z,g_0)^2|W=w)}}
\end{aligned}$$

By Assumption 4.2.1 and 4.2.2, $\sup_{g\in\Theta}\sup_{w\in[0,1]}|E(\rho(Z,g)|W=w)| < \infty$, and $\inf_{w\in[0,1]} E(\rho(Z,g_0)^2|W=w) > 0$, hence the second term is finite. As for the first term, note that for each n,

$$\sup_{g\in\Theta}\sup_{1\leq j\leq k_n}\frac{|E(\rho(Z,g)1_{W\in R_j^n})|}{P(W\in R_j^n)} = \sup_{g\in\Theta}\sup_{1\leq j\leq k_n}\frac{|\int_{R_j^n} E(\rho(Z,g)|W=w)dF_W(w)|}{P(W\in R_j^n)}$$

$$\leq \sup_{g\in\Theta}\sup_{w\in[0,1]}|E(\rho(Z,g)|W=w)| < \infty$$

Also for each R_j^n in the partition and each n,

$$\frac{E(\rho(Z,g_0)^2 1_{W\in R_j^n})}{P(W\in R_j^n)} = \frac{\int_{R_j^n} E(\rho(Z,g_0)^2|W=w)dF_W(w)}{P(W\in R_j^n)} \geq \inf_{w\in[0,1]} E(\rho(Z,g_0)^2|W=w) > 0$$

These yield that $\sup_{g\in\Theta} A(g) < \infty$. It is left to show $\sup_{g\in\Theta} B(g) \to 0$.

$$\begin{aligned}
\sup_{g\in\Theta} B(g) = &\quad \sup_{g}\sup_{1\leq j\leq k_n}\sup_{w\in R_j^n}\left|\frac{E[\rho(Z,g)1_{W\in R_j^n}]/P(W\in R_j^n)}{\sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n})/P(W\in R_j^n)}} - \frac{E(\rho(Z,g)|W=w)}{\sqrt{E(\rho(Z,g_0)^2|W=w)}}\right| \\
\leq &\quad \sup_{g}\sup_{1\leq j\leq k_n}\sup_{w\in R_j^n}\left|\frac{E[\rho(Z,g)1_{W\in R_j^n}]/P(W\in R_j^n) - E(\rho(Z,g)|W=w)}{\sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n})/P(W\in R_j^n)}}\right| \\
+ &\quad \sup_{g}\sup_{1\leq j\leq k_n}\sup_{w\in R_j^n}\left|\frac{E(\rho(Z,g)|W=w)}{\sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n})/P(W\in R_j^n)}} - \frac{E(\rho(Z,g)|W=w)}{\sqrt{E(\rho(Z,g_0)^2|W=w)}}\right| \\
= &\quad C + D,\ \text{say}
\end{aligned}$$

We have shown that the denominator in C is bounded below by $\sqrt{\inf_w E(\rho(Z,g_0)^2|W=w)}$, which is bounded away from zero. Hence

$$\begin{aligned}
C &\leq Const \sup_g \sup_{1\leq j\leq k_n} \sup_{w\in R_j^n} \left| \frac{E[\rho(Z,g)1_{W\in R_j^n}]}{P(W\in R_j^n)} - E(\rho(Z,g)|W=w) \right| \\
&= Const \sup_g \sup_{1\leq j\leq k_n} \sup_{w\in R_j^n} P(W\in R_j^n)^{-1} | \int_{R_j^n} E(\rho(Z,g)|W=t)dF_W(t) - \\
& \quad \int_{R_j^n} E(\rho(Z,g)|W=w)dF_W(t)| \\
&\leq Const \sup_{j\leq k_n} P(W\in R_j^n)^{-1} \int_{R_j^n} \sup_g \sup_{w\in R_j^n} |K_g(t)-K_g(w)|dF_W(t) \\
&\leq Const \times \delta, \text{ for any } \delta>0, \text{ and large } n.
\end{aligned}$$

The first equality is due to $E[\rho(Z,g)1_{W\in R_j^n}] = \int_{R_j^n} E(\rho(Z,g)|W=t)dF_W(t)$, and $E(\rho(Z,g)|W=w)P(W\in R_j^n) = \int_{R_j^n} E(\rho(Z,g)|W=w)dF_W(t)$. The second inequality follows by putting $\sup_g \sup_{w\in R_j^n}$ inside and rewrite $E(\rho(Z,g)|W)$ into $K_g(.)$. The last inequality follows by Assumption 4.2.2, and that $\{K_g(.) : g\in\Theta\}$ is uniformly equicontinuous on Θ given Θ is compact. Also, for large enough n, the size of R_j^n is arbitrarily small. In addition, for any R_j^n, we have $\int_{R_j^n} dF_W(t) = P(W\in R_j^n)$.

Finally,

$$D \leq \sup_{g,w} |E(\rho(Z,g)|W=w)|$$

$$\times \sup_{j\leq k_n} \sup_{w\in[0,1]} \left| \frac{1}{\sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n})/P(W\in R_j^n)}} - \frac{1}{\sqrt{E(\rho(Z,g_0)^2|W=w)}} \right|$$

Note that $\sup_{g,w} |E(\rho(Z,g)|W=w)| < \infty$, and

$$\frac{1}{\sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n})/P(W\in R_j^n)}} - \frac{1}{\sqrt{E(\rho(Z,g_0)^2|W=w)}}$$

$$
\begin{aligned}
&= \frac{\sqrt{E(\rho(Z,g_0)^2|W=w)} - \sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n})/P(W\in R_j^n)}}{\sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n})E(\rho(Z,g_0)^2|W=w)/P(W\in R_j^n)}} \\
&\leq Const \times \frac{E(\rho(Z,g_0)^2|W=w) - E(\rho(Z,g_0)^2 1_{W\in R_j^n})/P(W\in R_j^n)}{\sqrt{E(\rho(Z,g_0)^2|W=w)} + \sqrt{E(\rho(Z,g_0)^2 1_{W\in R_j^n})/P(W\in R_j^n)}} \\
&\leq \frac{Const}{P(W\in R_j^n)} \times \int_{R_j^n} [E(\rho(Z,g_0)^2|W=w) - E(\rho(Z,g_0)^2|W=t)]dF_W(t)
\end{aligned}
$$

By Assumption 4.2.1, $E(\rho(Z,g_0)^2|W=w)$ is uniformly continuous on $[a,b]$. Hence for any $\delta > 0$, there exists N, when $n > N$, for any R_j^n, $|R_j^n|$ is small enough, and $\int_{R_j^n} |E(\rho(Z,g_0)^2|W = w) - E(\rho(Z,g_0)^2|W=t)|dF_W(t) < \delta P(W\in R_j^n)$. Hence $D < Const\times\delta$. Since δ is arbitrary, $\sup_{g\in\Theta)} B(g) \to 0$.

(ii) Define $\xi_{nj}(g,X) = m_{nj}(g,X) - Em_{nj}(g,X)$, $j = 1, ..., k_n$, $\xi_n(g,X) = (\xi_{n1}, ..., \xi_{nk_n})^T = m_n(g,X) - Em_n(g,X)$, and $\bar{\xi}_n(g) = \frac{1}{n}\sum_{i=1}^n \xi_n(g,X_i) = \bar{m}_n(g) - Em_n(g,X) = (\bar{\xi}_{n1}, ..., \bar{\xi}_{nk_n})^T$. Then

$$
\begin{aligned}
\bar{G}(g) - G_{k_n}(g) &= [Em_n(g,X) + \bar{\xi}_n(g)]^T V_0^{-1}[Em_n(g,X) + \bar{\xi}_n(g)] - Em_n(g,X)^T V_0^{-1} Em_n(g,X) \\
&= 2Em_n(g,X)^T V_0^{-1}\bar{\xi}_n(g) + \bar{\xi}_n(g)^T V_0^{-1}\bar{\xi}_n(g) \\
&= W_n^1(g) + W_n^2(g), \text{ say}
\end{aligned}
$$

It suffices to show $W_n^i(g) \to 0$ uniformly on Θ in probability, $i = 1, 2$.

For $W_n^1(g)$: Step I: show $W_n^1(g) \to^p 0$ for each g. In fact, $EW_n^1(g) = 0$ since $E\bar{\xi}_n(g) = 0$ for each $g \in \Theta$. Now for any g,

$$Var(W_n^1(g)) = E(W_n^1(g)^2) = 4Em_n(g,X)^T V_0^{-1} E(\bar{\xi}_n(g)\bar{\xi}_n(g)^T) V_0^{-1} Em_n(g,X)$$

Note that $E(\bar{\xi}_n(g)\bar{\xi}_n(g)^T) = Var(\bar{\xi}_n(g)) = \frac{1}{n}Var(\xi_n(g)) = \frac{1}{n}Var(m_n(g,X))$. Hence by Assumption 4.2.3, $E(\bar{\xi}_n(g)\bar{\xi}_n(g)^T) = O(\frac{k_n}{n})$. In addition,

$$\begin{aligned}\lambda_{\max}(V_0^{-1}) &= \frac{1}{\min_{1\leq j\leq k_n} E(\rho(Z,g_0)^2 1_{W\in R_j^n})} = \frac{1}{\min_{1\leq j\leq k_n} E(E(\rho(Z,g_0)^2|W)1_{W\in R_j^n})} \\ &\leq \frac{1}{\min_{1\leq j\leq k_n} P(W\in R_j^n)} \\ &= O(k_n)\end{aligned}$$

The inequality and the last equality are due to Assumption 4.2.2 (i) and Assumption 4.2.4 (i). Therefore, $Var(W_n^1(g)) = O(\frac{k_n^2}{n})G_{k_n}(g) = O(\frac{k_n^2}{n}) = o(1)$.

Step II, show $\{W_n^1(g)\}$ is stochastic equicontinuous on Θ. In fact,

$$\begin{aligned}|W_n^1(g)| &= 2|Em_n(g,X)^T V_0^{-1}\bar{\xi}_n(g)| \leq \frac{2}{\sqrt{n}}\|Em_n(g,X)^T V_0^{-1}\|\|\sqrt{n}\bar{\xi}_n(g)\| \\ &= \frac{2}{\sqrt{n}}(Em_n(g,X)^T V_0^{-2} Em_n(g,X))^{\frac{1}{2}}(\sum_{j=1}^{k_n}(\sqrt{n}\bar{\xi}_{nj}(g))^2)^{\frac{1}{2}} \\ &\leq \frac{2}{\sqrt{n}}(\lambda_{\max}(V_0^{-1})G_{k_n}(g))^{\frac{1}{2}} \times \sqrt{k_n}\left(\frac{1}{k_n}\sum_{j=1}^{k_n}(\sqrt{n}\bar{\xi}_{nj}(g))^2\right)^{\frac{1}{2}} \\ &\leq O\left(\frac{k_n}{\sqrt{n}}\right)\max_{1\leq j\leq k_n}|\sqrt{n}\bar{\xi}_{n_j}(g)|\end{aligned}$$

By Assumption 4.2.3(iii), $\max_{1\leq j\leq k_n}|\sqrt{n}\bar{\xi}_{n_j}(g)|$ is stochastic equicontinuous. Hence $\{W_n^1(g)\}$ is stochastic equicontinuous.

Step III: By Newey and McFadden (1994) Lemma 2.8, it follows from that $\{W_n^1(g)\}$ is stochastic equicontinuous and pointwise converges to zero, and that Θ is compact, $W_n^1(g) \to 0$ uniformly on Θ.

For $W_n^2(g)$, we proceed in the same arguments. Step I: show $W_n^2(g)$ converges pointwise to zero. In fact

$$\begin{aligned} W_n^2(g) &= \bar{\xi}_n(g)^T V_0^{-1} \bar{\xi}_n(g) = \sum_{j=1}^{k_n} \frac{\bar{\xi}_{nj}(g)^2}{E(\rho(Z,g_0)^2 1_{W\in R_j^n})} \\ &\leq Const \sum_{j=1}^{k_n} \frac{\bar{\xi}_{nj}(g)^2}{P(W \in R_j^n)} = O(k_n) \sum_{j=1}^{k_n} \left(\frac{1}{n} \sum_{i=1}^{n} \xi_{nj}(g, X_i) \right)^2 \\ &= O(1) \left[\left(\frac{k_n}{n^2} \right) \sum_{j=1}^{k_n} \sum_{i=1}^{n} \xi_{nj}^2(g, X_i) + \left(\frac{k_n}{n^2} \right) \sum_{j=1}^{k_n} \sum_{i \neq l} \xi_{nj}(g, X_i) \xi_{nj}(g, X_l) \right] \\ &= O(1)[W_{n1}(g) + W_{n2}(g)], \text{ say} \end{aligned}$$

Step I-1: show $W_{n1}(g) \rightarrow^p 0$ for each g. It suffices to show $E|W_{n1}(g)|^2 \rightarrow^p 0$ for any g. In fact,

$$E|W_{n1}(g)|^2 = \frac{k_n^2}{n^4} E \left(\sum_{j=1}^{k_n} \sum_{i=1}^{n} \xi_{nj}^2(g, X_i) \right)^2 \leq \frac{k_n^4 n^2}{n^4} B = O\left(\frac{k_n^4}{n^2} \right) = o(1)$$

where the inequality and the last equality are due to Assumption 4.2.5 (ii) and Assumption 4.2.2 (ii).

Step I-2: show $W_{n2}(g) \rightarrow^p 0$ for each g. Since $\forall j$ and g, $E\xi_{nj(g,X)} = 0$, and $\xi_{nj}(g, X_i)$ and $\xi_{nj}(g, X_l)$ are independent if $i \neq l$, thus $EW_{n2}(g) = 0$. To show $Var(W_{n2}(g)) \rightarrow 0$ for each $g \in \Theta$, we apply the following result:

Lemma C.1.1. *If $X_1, ..., X_n$ are uncorrelated with means all equal to zero, then*

$$Var(\sum_{i\neq j} X_i^T X_j) = 2 \sum_{i \neq j} tr[E(X_i X_i^T) E(X_j X_j^T)]$$

Proof. Straightforward but tedious calculations yield that both left and right hand side equal $2\sum_{i\neq j} E(X_i^T X_j)^2$. □

Back to Step I-2: Applying Lemma A.1 to $W_{n2}(g) = \frac{k_n}{n^2}\sum_{i\neq l}\xi_n(g,X_i)^T\xi_n(g,X_l)$, and noting that $tr(AB) \leq \sqrt{tr(A^2)tr(B^2)}$, we have

$$\begin{aligned} Var(W_{n2}(g)) &= \frac{2k_n^2}{n^4}\sum_{i\neq j} tr[E(\xi_n(g,X_i)\xi_n(g,X_i)^T)E(\xi_n(g,X_j)\xi_n(g,X_j)^T)] \\ &\leq \frac{2k_n^2}{n^4}\sum_{i\neq j}\sqrt{tr[E^2(\xi_n(g,X_i)\xi_n(g,X_i)^T)] \times tr[E^2(\xi_n(g,X_j)\xi_n(g,X_j)^T)]} \end{aligned}$$

For each $i = 1, ..., n$ and $g \in \Theta$, $E(\xi_n(g,X_i)\xi_n(g,X_i)^T) = Var(\xi_n(g,X_i)) = Var(m_n(g,X)) = O(k_n)$, hence $tr[E^2(\xi_n(g,X_i)\xi_n(g,X_i)^T)] = O(k_n^3)$. Therefore $Var(W_{n2}(g)) = O(k_n^5/n^2) = o(1)$.

Step II: show $\{W_n^2(g)\}$ is stochastic equicontinuous on Θ. In fact,

$$\begin{aligned} W_n^2(g) &= \bar{\xi}_n(g)^T V_0^{-1}\bar{\xi}_n(g) = \sum_{j=1}^{k_n}\frac{\bar{\xi}_{nj}(g)^2}{E(\rho(Z,g_0)^2 1_{W\in R_j^n})} \\ &\leq Const\sum_{j=1}^{k_n}\frac{\bar{\xi}_{nj}(g)^2}{P(W\in R_j^n)} = O(k_n)\sum_{j=1}^{k_n}\bar{\xi}_{nj}(g)^2 \\ &= O\left(\frac{k_n^2}{n}\right)\frac{1}{k_n}\sum_{j=1}^{k_n}[\sqrt{n}\bar{\xi}_{nj}(g)]^2 \\ &\leq O\left(\frac{k_n^2}{n}\right)\max_{1\leq j\leq k_n}|\sqrt{n}\bar{\xi}_{nj}(g)|^2 \end{aligned}$$

Note that if $\{|Q_n(g)|\}$ is stochastic equicontinuous, then so is $\{Q_n(g)^2\}$. Therefore $\{W_n^2(g)\}$ is stochastic equicontinuous. Q.E.D.

C.1.2. Proof for Theorem 4.2.2

The following lemmas are useful.

Lemma C.1.2. *$G : \Theta \to \mathbb{R}$ is continuous*

Proof. By Assumption 4.2.3, there exists $\sigma^2 > 0$ such that $\inf_{w\in[0,1]} E(\rho(Z,g_0)^2|W = w) \geq \sigma^2 > 0$. In addition, given $\Theta \times [0,1]$ is compact, by Assumption 4.2.3 (ii), $\sup_{(g,w)\in\Theta\times[0,1]} |E(\rho(Z,g)|W = w)| < \infty$. Hence for any $g_1, g_2 \in \Theta$,

$$\begin{aligned} |G(g_1) - G(g_2)| &\leq \int_a^b \frac{|E[\rho(Z,g_1)|W=w]^2 - E[\rho(Z,g_2)|W=w]^2|}{E(\rho(Z,g_0)^2|W=w)} dF_W(w) \\ &\leq \frac{2}{\sigma^2} \sup_{(g,w)\in\Theta\times[0,1]} |E[\rho(Z,g)|W=w]| \\ &\quad \times \int_a^b E(|\rho(Z,g_1) - \rho(Z,g_2)||W=w) dF_W(w) \\ &\leq Const \sup_{z\in\mathcal{Z}} |\rho(z,g_1) - \rho(z,g_2)| \end{aligned}$$

The continuity follows since $\{\rho(z,.) : z \in \mathcal{Z}\}$ is equicontinuous on Θ.

Lemma C.1.3. *For any* $\delta > 0$, $\liminf_{n\to\infty} P(G(g_q) - \inf_{g\in\Theta_{q_n}} G(g) < \delta) \succ e^{-cq_n}$

Proof. Write $q = q_n$ (depending on n). By Lemma C.1.2, G is continuous on Θ. Hence for any $\epsilon > 0$, there exists $\delta > 0$, for $g_q \in \Theta_I^\delta$, $|G(g_q) - \sup_{g\in\Theta_I} G(g)| < \epsilon$. Note that $\Theta_I = \arg\min_{g\in\Theta} G(g)$, hence $G(g_q) - \inf_{g\in\Theta} G(g) < \epsilon$. In addition, for any q, $G(g_q) - \inf_{g\in\Theta_q} G(g) \leq G(g_q) - \inf_{g\in\Theta} G(g) < \epsilon$, which implies $\{g_q \in \Theta_I^\delta\} \subset \{G(g_q) - \inf_{g\in\Theta_q} G(g) < \epsilon\}$. Then Assumption 4.2.6, implies that for any $\epsilon > 0$, for all large q, $P(G(g_q) - \inf_{g\in\Theta_q} G(g) < \epsilon) \geq P(g_q \in \Theta_I^\delta) \succ e^{-cq_n}$.

Proof of Theorem 4.2.2:

The posterior of g_q^0 is given by

$$p(g_q|X^n) \propto p(g_q)\exp\left(-\frac{n}{2}\bar{G}(g_q)\right)$$

A straightforward application of Jiang and Tanner (2008, Proposition 6) renders that $\forall\delta > 0$,

$$\begin{aligned}P(G(g_q) - \inf_{g\in\Theta_q} G(g) > 5\delta|X^n) &\le P(\sup_{g\in\Theta_q} |\bar{G}(g) - G(g)| \ge \delta) \\ &+ \frac{e^{-2n\delta}}{P(G(g_q) - \inf_{g\in\Theta_q} G(g) < \delta)}\end{aligned} \tag{C.1}$$

By Theorem 4.2.1, $\sup_{g\in\Theta_q} |\bar{G}(g) - G(g)| \to 0$ in the probability distribution of X^n as $n \to \infty$. Also, Lemma C.1.3 implies that for any $\delta > 0$,

$$\frac{e^{-2n\delta}}{P(G(g_q) - \inf_{g\in\Theta_q} G(g) < \delta)} \le e^{-2n\delta + cq_n} = o(1)$$

provided that $q_n/n \to 0$. It then follows from C.1 that $G(g_q) - \inf_{g\in\Theta_q} G(g) \to 0$ in the posterior probability. Note that $\inf_{g\in\Theta} = G(g_0)$, and $\inf_{\Theta_q} G(g) \ge G(g_0)$, $g_q^0 \in \Theta_q$, which implies $0 \le G(g_q^0) - \inf_{g\in\Theta_q} G(g) \le G(g_q^0) - G(g_0)$, hence by the triangular inequality,

$$\begin{aligned} G(g_q) - \inf_{g\in\Theta} G(g) &\le G(g_q) - \inf_{g\in\Theta_q} G(g) + |\inf_{g\in\Theta_q} G(g) - G(g_q^0)| + |G(g_q^0) - G(g_0)| \\ &\le G(g_q) - \inf_{g\in\Theta_q} G(g) + 2(G(g_q^0) - G(g_0)) \end{aligned}$$

Since $\|g_q^0 - g_0\| \to 0$ and $G(g)$ is continuous on Θ, $G(g_q) - \inf_{g\in\Theta} G(g) \to 0$ in the posterior probability, meaning that in the probability of the distribution of X^n, for any $\epsilon > 0$,

$$P(G(g_q) - \inf_{g\in\Theta} G(g) > \epsilon | X^n) \to 0 \tag{C.2}$$

By definition $\Theta_I = \arg\min_\Theta G(g)$, hence for any $\delta > 0$, $\epsilon_0 \equiv \inf_{g\in\Theta\cap(\Theta_I^\delta)^c} G(g) - \inf_{g\in\Theta} G(g) > 0$. C.2 then yields with probability approaching 1, $G(g_q) < \inf_{g\in\Theta\cap(\Theta_I^\delta)^c} G(g)$, which implies $g_q \in \Theta_I^\delta$ with the posterior probability approaching 1. Q.E.D.

C.2. Proofs for Section 4.3

C.2.1. Proof for Proposition 4.3.1

Proof. (i) When $\rho(Z, g) = Y - h(W\theta)$, then

$$\begin{aligned}
& |E(\rho(Z, g_0)^2|W = w_1) - E(\rho(Z, g_0)^2|W = w_2)| \leq \frac{1}{f_W(w_1)f_W(w_2)} \\
& \times \int |(y - h_0(w_1\theta_0))^2 f_W(w_2) f_{W|Y}(w_1|y) - (y - h_0(w_2\theta_0))^2 f_W(w_1) f_{W|Y}(w_2|y)| dF_Y(y) \\
\leq \; & Const \cdot [\int (y - h_0(w_1\theta_0))^2 |f(w_1|y) - f(w_2|y)| dF_Y(y) \\
& + |f_W(w_1) - f_W(w_2)| \int (y - h_0(w_1\theta_0))^2 f_{W|Y}(w_2|y) dF_Y(y) \\
& + |h_0(w_1\theta_0) - h_0(w_2\theta_0)| \int f_{W|Y}(w_2|y)|2y + h_0(w_1\theta_0) + h_2(w_2\theta_0)| dF_Y(y) \\
\leq \; & A + B + C, say
\end{aligned}$$

For any $\delta > 0$, $\|w_1 - w_2\|$ is small implies $A < \delta$, provided that $\forall \epsilon > 0$, $\exists \delta$, such that $\sup_y \sup_{\|w_1 - w_2\| < \delta} |f_{W|Y}(w_1|y) - f_{W|Y}(w_2|y)| < \epsilon$, and that $\sup_t |h(t)| < \infty$. Also $B < \delta$ provided that f_W is continuous on $[a, b]$, and $\sup_{(w,y)\in\mathcal{Z}} f_{W|Y}(w|y) < \infty$. Finally, $C < \delta$ follows from the continuity of $h_0(.)$.

(ii) For any $w_1 \neq w_2$, $\sup_{g\in\Theta} |E(\rho(Z, g)|W = w_1) - E(\rho(Z, g)|W = w_2)| = \sup_{(h,\theta)\in\Theta} |h(w_1\theta) - h(w_2\theta)|$. Since $h \in \mathcal{H}$ is compact under Hölder norm, there exists a constant such that for any $h \in \mathcal{H}$, and $t_1 \neq t_2$,

$$\frac{|h(t_1) - h(t_2)|}{|t_1 - t_2|} \leq Const$$

Hence $\sup_{h\in\mathcal{H}} |h(w_1\theta) - h(w_2\theta)| \leq Const \cdot |w_1 - w_2| \|\theta\|$. Note that Const does not depend on h and θ is in a compact set, it follows that for any $\epsilon > 0$, exists δ, as long as $|w_1 - w_2| < \delta$, $\sup_{(h,\theta)\in\Theta} |h(w_1\theta) - h(w_2\theta)| < \epsilon$, which yields the equicontinuity of $\{E(\rho(Z, g)|W = .) : g \in \Theta\}$.

(iii) For any w_1, w_2 and h_1, h_2, $|\rho(z, g_1) - \rho(z, g_2)| = |h_1(w\theta_1) - h_2(w\theta_2)| \leq |h_1(w\theta_1) - h_1(w\theta_2)| + |h_1(w\theta_2) - h_2(w\theta_2)| \leq \|h_1\|_s |w| \|\theta_1 - \theta_2\| + \|h_1 - h_2\|_s$. This is because,

$$|h_1(w\theta_1) - h_1(w\theta_2)| \leq \sup_{t_1 \neq t_2} \frac{|h_1(t) - h_1(t_2)|}{|t_1 - t_2|} |w\theta_1 - w\theta_2| \leq \|h_1\|_s |w\theta_1 - w\theta_2|$$

and $|h_1(w\theta_2) - h_2(w\theta_2)| \leq \sup_t |h_1(t) - h_2(t)|$. Recall that $\mathcal{H} = \{h : \|h\|_s < B\}$ for some known constant B. Hence as long as $\|\theta_1 - \theta_2\| < \epsilon/2B$, and $\|h_1 - h_2\| < \epsilon/2$, $\sup_w |h_1(w\theta_1) - h_2(w\theta_2)| \leq \epsilon$. Q.E.D.

C.2.2. Proof for Proposition 4.3.1: Assumption 4.2.4

Lemma C.2.1. *For all* $j = 1, ..., k_n$, *let* $\bar{\xi}_{nj}(g) = \bar{m}_{nj}(g) - Em_{nj}(g, X)$, *then Assumption 4.2.4 holds on* $(\Theta, \|.\|_H)$ *if* $\forall \epsilon > 0$, $\exists \delta > 0$, *such that*

$$\limsup_{n \to \infty} \max_{j \leq k_n} P\left(\sup_{\|g_1 - g_2\|_H \leq \delta} \sqrt{n} |\bar{\xi}_{nj}(g_1) - \bar{\xi}_{nj}(g_2)| > \epsilon \right) < \epsilon \tag{C.3}$$

Proof. For any $g_1, g_2 \in \Theta$, let j_1 and j_2 be such that

$$\max_{1 \leq j \leq k_n} \sqrt{n} |\bar{m}_{nj}(g_1) - Em_{nj}(g_1, X)| = \sqrt{n} |\bar{m}_{nj_1}(g_1) - Em_{nj_1}(g_1, X)|.$$

$$\max_{1 \leq j \leq k_n} \sqrt{n} |\bar{m}_{nj}(g_2) - Em_{nj}(g_1, X)| = \sqrt{n} |\bar{m}_{nj_2}(g_2) - Em_{nj_2}(g_2, X)|.$$

Let $Q_n(g) = \max_{1 \leq j \leq k_n} \sqrt{n} |\bar{m}_{nj}(g) - Em_{nj}(g, X)|$, then

$$Q_n(g_1) = \sqrt{n} |\bar{\xi}_{nj_1}(g_1)|, \quad Q_n(g_2) = \sqrt{n} |\bar{\xi}_{nj_2}(g_2)|$$

By definition, $Q_n(g_1) \geq \sqrt{n} |\bar{\xi}_{nj_2}(g_1)|$, and $Q_n(g_2) \geq \sqrt{n} |\bar{\xi}_{nj_1}(g_2)|$. It follows that $\sqrt{n} |\bar{\xi}_{nj_2}(g_1)| - Q_n(g_2) \leq Q_n(g_1) - Q_n(g_2) \leq Q_n(g_1) - \sqrt{n} |\bar{\xi}_{nj_1}(g_2)|$. Note that if $y \leq x \leq z$, then

$|x| \leq |y| + |z|$. Hence by the triangular inequality,

$$|Q_n(g_1) - Q_n(g_2)| \leq \sqrt{n}|\bar{\xi}_{nj_1}(g_1) - \bar{\xi}_{nj_1}(g_2)| + \sqrt{n}|\bar{\xi}_{nj_2}(g_1) - \bar{\xi}_{nj_2}(g_2)|$$

By the definition in Andrews (1992), $Q_n(g)$ is stochastic equicontinuous if (C.3) holds. Q.E.D.

Now $Em_{nj}(g, X) = E[(Y - h(W\theta))1(W \in R_j)] = E[1(W \in R_j)E(Y - h(W\theta)|W)]$. Since $E(Y - h(W\theta)|W = w)$ is continuous on $w \in [a, b]$ for all h by Proposition 4.3.1, therefore $Em_{nj}(g, X) \leq \sup_w E(Y - h(W\theta)|W = w)P(W \in R_j) = O(1/k)$ for all g. By weak law of large number, $k\bar{m}_{nj}(g) = O_p(kEm_{nj}(g, X)) = O_p(1)$. Hence by Proposition 3 in Jiang (2009), for any small $\gamma > 0$, for compact Θ, $\sup_{g\in\Theta} |k\bar{m}_{nj}(g) - kEm_{nj}(g, X))| = o_p(n^{-1/2+\gamma})$, which implies implies $\sqrt{n} \sup_{g\in\Theta} \bar{\xi}_{nj}(g) = O_p(n^{\gamma}/k) = o_p(1)$ for all $j \leq k$. Therefore for all $j \leq k$, for large enough n,

$$P\left(\sup_{\|g_1 - g_2\|_H \leq \delta} \sqrt{n}|\bar{\xi}_{nj}(g_1) - \bar{\xi}_{nj}(g_2)| > \epsilon\right) \leq P\left(\sup_{g\in\Theta} \sqrt{n}\bar{\xi}_{nj}(g) > \frac{\epsilon}{2}\right) < \epsilon$$

Q.E.D.

C.3. Proofs for Section 4.4

C.3.1. Proof for Proposition 4.4.1

Proof. (i) First note that for all $w \in [0, 1]$, and $\rho(Z, g) = Y - g(X)$

$$\begin{aligned} E(\rho(Z, g_0)^2|W = w) &= \int (y - g_0(x))^2 dF_{Y,X|W}(y, x|w) \\ &= \int \frac{(y - g_0(x))^2}{f_W(w)} f_{W|Y,X}(w|y, x)dF_{Y,X}(y, x) \end{aligned}$$

Hence for any w_1, w_2,

$$\begin{aligned} & |E(\rho(Z,g_0)^2|W=w_1) - E(\rho(Z,g_0)^2|W=w_2)| \\ \le & \int (y-g_0(x))^2 \left|\frac{f_{W|Y,X}(w_1|y,x)}{f_W(w_1)} - \frac{f_{W|Y,X}(w_2|y,x)}{f_W(w_2)}\right| dF_{Y,X}(y,x) \\ \le & \frac{1}{f_W(w_1)f_W(w_2)}\int (y-g_0(x))^2\times \\ & |f_{W|Y,X}(w_1|y,x)f_W(w_2) - f_{W|Y,X}(w_2|y,x)f_W(w_1)|dF_{Y,X}(y,x) \end{aligned} \tag{C.4}$$

Note that $|f_{W|Y,X}(w_1|y,x)f_W(w_2) - f_{W|Y,X}(w_2|y,x)f_W(w_1)| \le |f_{W|Y,X}(w_1|y,x)f_W(w_2) - f_{W|Y,X}(w_2|y,x)f_W(w_2)| + |f_{W|Y,X}(w_2|y,x)f_W(w_2) - f_{W|Y,X}(w_2|y,x)f_W(w_1)| \le \sup_{w\in[0,1]} f_W(w)\cdot |f_{W|Y,X}(w_1|y,x) - f_{W|Y,X}(w_2|y,x)| + \sup_{w\in[0,1]} f_{W|Y,X}(w,|y,x)\cdot|f_W(w_2) - f_W(w_1)| \le Const\cdot \sup_{x,y}|f_{W|Y,X}(w_1|y,x) - f_{W|Y,X}(w_2|y,x)| + Const\cdot|f_W(w_1) - f_W(w_2)|$.

By Assumption 4.4.2, $\forall\delta > 0$, $\exists d > 0$, such that when $|w_1 - w_2| < d$, $|f_{W|Y,X}(w_1|y,x)f_W(w_2) - f_{W|Y,X}(w_2|y,x)f_W(w_1)| < \delta$ for all x, y. In addition, f_W is bounded away from zero. Hence C.4 $\le Const\cdot E(\epsilon^2)\delta$, which yields the continuity of $E(\rho(Z,g_0)^2|W = w)$ on $[0,1]$.

(ii) For all w_1, w_2, and $g \in \Theta$,

$$\begin{aligned} |K_g(w_1) - K_g(w_2)| &= |E(Y-g(X)|W=w_1) - E(Y-g(X)|W=w_2)| \\ &\le \int |y-g(x)| \left|\frac{f_{W|Y,X}(w_1|y,x)}{f_W(w_1)} - \frac{f_{W|Y,X}(w_2|y,x)}{f_W(w_2)}\right| dF_{Y,X}(y,x) \end{aligned}$$

Using a similar argument as in (i), we see that for any $\delta > 0$, as long as $|w_1 - w_2| < d$ for some d, $|K_g(w_1) - K_g(w_2)| < \delta\cdot Const\cdot E|Y - g(X)|$. Then Assumption 4.2.3 (ii) follows provided that $EY^2 < \infty$ and $g \in \Theta$.

(iii) Condition (iii) follows immediately given that $\rho(Z, g) = Y - g(X)$, and $\sup_x |g_1(x) - g_2(x)| = \|g_1 - g_2\|_H$.

Finally, Assumption 4.2.4 can be proved using the same argument in the proof of Proposition 4.3.1, using Lemma C.2.1. Q.E.D.

C.3.2. Proof of Theorem 4.4.1

First, we apply the singular representation of T. Since T is compact, let $\{|\lambda_i|, \phi_j, \psi_j\}_{i\geq 1}$ be the singular system of T, where $|\lambda_1|, |\lambda_2|...$, are the singular values of T such that $T\phi_i = |\lambda_i|\psi_i$. For any $g \in L^2(X)$, we have the singular value decomposition $g(x) = \sum_{i=1}^{\infty} b_i\phi_i(x) + Q$, where $Q \in \mathcal{N}(T)$. Since we assume in this case g_0 is identified, therefore the null space $\mathcal{N}(T) = \{0\}$. Hence $Q = 0$. Also, $(Tg)(w) = \sum_{i=1}^{\infty} b_i|\lambda_i|\psi_i(w)$. Define $T^* : L^2(X) \to L^2(X)$, such that for any $g = \sum_{i=1}^{\infty} b_i\phi_i(x)$,

$$(T^*g)(x) = \sum_{i=1}^{\infty} b_i|\lambda_i|\phi_i(x)$$

with norm $||T^*g||^2 = \sum_{i=1}^{\infty} b_i^2|\lambda_i|^2$. In addition, the inner product of $L^2(W)$ is defined as

$$< \psi_i, \psi_j >= \int \psi_i(w)\psi_j(w)E(\epsilon^2|w)^{-1}dF_W(w)$$

and $\{\psi_i\}$ are orthonormalized such that $< \psi_i, \psi_j >= \delta_{ij}$, the Kronecker delta. Therefore the norm of Tg is

$$\|Tg\|_W^2 = \int_a^b [(Tg)(w)]^2(E[\epsilon^2|W = w])^{-1}dF_W(w) = \sum_{i=1}^{\infty} b_i^2\lambda_i^2 = ||T^*g||$$

Therefore $G(g) = ||T(g - g_0)||_W^2 = \|T^*(g - g_0)\|^2$.

Let $g_q = \sum_{i=1}^q b_i\phi_i(x)$. The proof is carried out by deriving an upper bound of $\int \|g_q - g_0\|^2 L(g_q)p(g_q)db$, and a lower bound of $\int L(g_q)p(g_q)db$. For some sequence $\epsilon_n^2 = O(a_n^4/\lambda_{s_n}^2) = o(1)$, define

$$A_n = \{(b_1, ..., b_q) \in \mathbb{R}^q : G(g_q) + a_n^2\|g_q\|^2 \leq \epsilon_n^2\}$$

The following lemmas are useful:

Lemma C.3.1.

$$\int_{A_n} \|g_q - g_0\|^2 L(g_q)p(g_q)db = O_p\left(\left(\frac{\epsilon_n^2}{a_n^2} + \frac{a_n^2}{\lambda_{s_n}^2} + \sum_{j\geq s_n} g_j^2\right)(\frac{\epsilon_n}{a_n})^{q_n}\right)$$

Proof. Note that on A_n, $G(g_q) + a_n^2\|g_q\|^2 \leq \epsilon_n^2$. Let $\xi = g_q - g_0 + a_n(T^* + a_n)^{-1}g_0$. Hence

$$\begin{aligned}\|(T^* + a_n)\xi\|^2 &= \|(T^* + a_n)(g_q - g_0) + a_n g_0\|^2 = \|T^*(g_q - g_0) + a_n g_q\|^2 \\ &\leq 2(\|T^*(g_q - g_0\|^2 + a_n^2\|g_q\|^2) \leq 2\epsilon_n^2\end{aligned}$$

Let $\xi(x) = \sum_{i=1}^\infty \eta_i\phi_i(x)$ be the spectrum expansion of ξ. $\|(T^* + a_n)\xi\|^2 = \sum_{i=1}^\infty(|\lambda_i| + a_n)^2\eta_i^2 \geq a_n^2\|\xi\|^2$. Therefore, $||\xi||^2 \leq 2\epsilon^2/a_n^2$. It follows that

$$\begin{aligned}\|g_q - g_0\|^2 &\leq 2a_n^2\|(T^* + a_n)^{-1}g_0\|^2 + 4\frac{\epsilon_n^2}{a_n^2} = 2\sum_{j=1}^\infty\left(\frac{a_n}{|\lambda_j| + a_n}\right)^2 g_j^2 + \frac{4\epsilon_n^2}{a_n^2} \\ &\leq \frac{2a_n^2}{\lambda_{s_n}^2}\|g_0\|^2 + 2\sum_{j\geq s_n} g_j^2 + \frac{4\epsilon_n^2}{a_n^2}\end{aligned}$$

In addition, $A_n \subset A_n^* = \{b \in \mathbb{R}^{q_n} : a_n^2\|b\|^2 \leq \epsilon_n^2\}$. We have

$$\int_{A_n} \|g_q - g_0\|^2 L(g_q)p(g_q)db \quad \propto \quad \int_{A_n} \|g_q - g_0\|^2 e^{-na_n\|g_q\|^2 - \frac{n}{2}\bar{m}_n(g_q)^T V_0^{-1}\bar{m}_n(g_q)}db$$

$$\begin{aligned} &\leq \int_{A_n} \|g_q - g_0\|^2 db \leq \left(\frac{2a_n^2}{\lambda_{s_n}^2}\|g_0\|^2 + 2\sum_{j\geq s_n} g_j^2 + \frac{4\epsilon_n^2}{a_n^2}\right)\mu(A_n^*) \\ &= O_p\left(\left(\frac{2a_n^2}{\lambda_{s_n}^2}\|g_0\|^2 + 2\sum_{j\geq s_n} g_j^2 + \frac{4\epsilon_n^2}{a_n^2}\right)(\frac{\epsilon_n}{a_n})^{q_n}\right) \end{aligned}$$

Q.E.D.

Let $T^*g_0(x) = \sum_{i=1}^{\infty} \mu_i\phi_i(x)$ be the spectrum expansion of $T^*g_0(x)$. For $q = q_n$, define

$$\begin{aligned} A_{1c} &= \{b \in \mathbb{R}^q : a_n^2\|g_q\|^2 > \epsilon_n^2\} \\ A_{2c} &= \{b \in \mathbb{R}^q : \|T^*(g_q - g_0)\|^2 + a_n^2\|g_q\|^2 > \epsilon_n^2, a_n^2\|g_q\|^2 \leq \epsilon_n^2\} \end{aligned}$$

Recall that $A_n^c = \{b \in \mathbb{R}^q : \|T^*(g_q - g_0)\|^2 + a_n^2\|g_q\|^2 > \epsilon_n^2\}$, which then follows that $A_n^c \subset A_{1c} \cup A_{2c}$. We evaluate $A_{1,2c}$ respectively.

Lemma C.3.2.

$$\int_{A_{2c}} \|g_q - g_0\|^2 L(g_q)p(g_q)db = O_p\left(\exp(n^{1-\alpha} - n\epsilon_n^2)\left(\frac{\epsilon_n^2}{a_n^2}\right)^{q/2}\right)$$

Proof. $A_{2c} \subset Ball(0, |\epsilon_n|/a_n)$, a ball centered at zero with radius $|\epsilon_n|/a_n = o(1)$. By Assumption 4.4.4, we have uniform convergence on A_{2c}:

$\sup_{g_q\in A_{2c}} |\bar{m}_n(g_q)^T V_0^{-1}\bar{m}_n(g_q) - G(g_q)| < n^{-\alpha}$. Therefore

$$\begin{aligned} &\int_{A_{2c}} \|g_q - g_0\|^2 L(g_q)p(g_q)db \leq \int_{A_{2c}} \|g_q - g_0\|^2 \exp(-nG(g_q) - na_n^2\|g_q\|^2 \\ &+n|G(g_q) - \bar{m}(g_q)'V_0^{-1}\bar{m}(g_q)|)db \\ &\leq \exp(n^{1-\alpha} - n\epsilon_n^2)\int_{A_{2c}} \|g_q - g_0\|^2 db \\ &\leq 2\exp(n^{1-\alpha} - n\epsilon_n^2)\left(\|g_0\|^2\mu(Ball(0, |\epsilon_n|/a_n)) + \int_{Ball(0,|\epsilon_n|/a_n)} \|g_q\|^2 db\right) \end{aligned}$$

$$
\begin{aligned}
&\le 2\exp(n^{1-\alpha} - n\epsilon_n^2)\left(\frac{\epsilon_n^2}{a_n^2}\right)^{q/2}\left(\|g_0\|^2 + \frac{\epsilon_n^2}{a_n^2}\right) \\
&= O_p\left(\exp(n^{1-\alpha} - n\epsilon_n^2)\left(\frac{\epsilon_n^2}{a_n^2}\right)^{q/2}\right)
\end{aligned}
$$

The equality follows from that $\epsilon_n^2/a_n^2 = O(a_n^2/\lambda_{s_n}^2) = o(1)$. Q.E.D.

Lemma C.3.3. *For some constant* $c > 0$*,*

$$\int_{A_{1c}} \|g_q - g_0\|^2 L(g_q)p(g_q)db = O_p\left(e^{-n\epsilon_n^2}(\frac{\pi}{n\lambda_q^2})^{q/2}\left(1 + \frac{q}{n\lambda_q^2}\right)\right)$$

Proof. $\int_{A_{1c}} \|g_q - g_0\|^2 L(g_q)p(g_q)db \le e^{-n\epsilon_n^2} \int_{A_{1c}} \|g_1 - g_0\|^2 e^{-n\bar{m}(g_b)^T V_0^{-1}\bar{m}(g_b)}db$. Define the following matrix notion:

$$
\begin{aligned}
H &= (h_{ij})_{n\times n}, \qquad \text{where } h_{ij} = \sum_{l=1}^{k} I(W_i \in R_l, W_j \in R_l) \\
\Psi_q &= (\phi_i(X_j))_{q\times n}, \qquad y = (Y_1, ..., Y_n)^T, \quad V = \Psi_q H \Psi_q^T
\end{aligned}
$$

Write $g_q(x) = \sum_{i=1}^{q} b_i\phi_i(x)$, and $\bar{m}_j(g_q) = n^{-1}\sum_{i=1}^{n}(Y_i - g_q(X_i))I(W_i \in R_j)$. It is then straightforward to verify that $\sum_{j=1}^{k} \bar{m}_j(g_q)^2$ has the following matrix representation:

$$
\begin{aligned}
\sum_{j=1}^{k} \bar{m}_j(g_q)^2 = \quad & \frac{1}{n^2}(b^T - y^T H\Psi_q^T V^{-1})V(b - V^{-1}\Psi_q Hy) \\
& + \frac{1}{n^2}(y^T Hy - y^T H\Psi_q^T V^{-1}\Psi_q Hy) \\
& \ge \frac{1}{n^2}(b^T - y^T H\Psi_q^T V^{-1})V(b - V^{-1}\Psi_q Hy)
\end{aligned}
$$

Note that $\bar{m}(g_q)^T V_0^{-1}\bar{m}(g_q) = \sum_{j=1}^{k} \bar{m}_j(g_q)^2 E[\epsilon^2 I(W \in R_j)]^{-1}$, and for all $j = 1, ..., k$, $E[\epsilon^2 I(W \in R_j)] = O(k^{-1})$. Thus there exist two positive constants c_1 and c_2 such that

$c_1 k \sum_{j=1}^{k} \bar{m}_j(g_q)^2 \leq \bar{m}(g_q)^T V_0^{-1} \bar{m}(g_q) \leq c_2 k \sum_{j=1}^{k} \bar{m}_j(g_q)^2$. Therefore, without loss of generality, we will only consider the simplified case there there exists a universal constant c such that $\bar{m}(g_q)^T V_0^{-1} \bar{m}(g_q) = ck \sum_{j=1}^{k} \bar{m}_j(g_q)^2$.

Let $Z = (Z_1, ..., Z_q)$ be a multivariate normal random vector, such that $Z \sim MVN(\mu, \Sigma)$, where $\mu = y^T H \Psi_q^T V^{-1}$, and $\Sigma = \frac{n}{2ck} V^{-1}$. The joint density function of $(Z_1, ..., Z_q)$ is $f_Z(b_1, ..., b_q) = (2\pi)^{-q/2} \det(\Sigma)^{-1/2} e^{-\frac{1}{2}(b-\mu)^T \Sigma^{-1}(b-\mu)}$. Hence $e^{-cnk \sum_{j=1}^{k} \bar{m}_j(g_q)^2} = (2\pi)^{q/2} \det(\Sigma)^{1/2} f_Z(b_1, ..., b_q$ which yields

$$
\begin{aligned}
&\int_{A_{1c}} \|g_q - g_0\|^2 L(g_q) p(g_q) db \leq e^{-n\epsilon_n^2} \int_{A_{1c}} \|g_q - g_0\|^2 e^{-n\bar{m}(g_b)^T V_0^{-1} \bar{m}(g_b)} db \\
&\leq \exp(-n\epsilon_n^2) \int_{A_{1c}} \|g_q - g_0\|^2 e^{-cnk \sum_{j=1}^{k} \bar{m}_j(g_q)^2} db \\
&\leq \exp(-n\epsilon_n^2)(2\pi)^{q/2} \det(\Sigma)^{1/2} \int_{A_{1c}} \|g_q - g_0\|^2 f_Z(b_1, ..., b_q) db \\
&\leq \exp(-n\epsilon_n^2)(2\pi)^{q/2} \det(\Sigma)^{1/2} \left(2\|g_0\|^2 + 2 \int_{A_{1c}} \sum_{i=1}^{q} b_i^2 f_Z(b_1, ..., b_q) db \right) \\
&\leq \exp(-n\epsilon_n^2)(2\pi)^{q/2} \det(\Sigma)^{1/2} \left(2\|g_0\|^2 + 2 \sum_{i=1}^{q} E[Z_i^2 | Data] \right)
\end{aligned}
$$

Let $\tilde{Z}_1, .., \tilde{Z}_q$ be independent normal variables such that $\tilde{Z}_i \sim N(\mu_i/|\lambda_i|, (2n\lambda_i^2)^{-1})$. The joint density function of $(\tilde{Z}_1, ..., \tilde{Z}_q)$ is

$f_{\tilde{Z}}(b_1, ..., b_q) = (n/\pi)^{q/2} \prod_{i=1}^{q} |\lambda_i| \exp(-n \sum_{i=1}^{q} (b_i|\lambda_i| - \mu_i)^2)$. Since $G(g_q) = \|T^*(g_q - g_0)\|^2 = \sum_{i=1}^{q} (b_i|\lambda_i| - \mu_i)^2 + \sum_{i \geq q} \mu_i^2$, hence $e^{-nG(g_q)} \propto f_{\tilde{Z}}(b_1, ..., b_q)$. In addition, $e^{-n\bar{m}(g_q)^T V_0^{-1} \bar{m}(g_q)} \propto f_Z(b_1, ..., b_q)$, and note that $\forall g_q \in \Theta_q$, we have pointwise convergence $\lim_{n \to \infty} \bar{m}(g_q)^T V_0^{-1} \bar{m}(g_q) = G(g_q)$. Hence it follows that $E(Z_i^2 | Data) = O_p(E(\tilde{Z}_i^2 | Data)) = O_p(\frac{\mu_i^2}{\lambda_i^2} + \frac{1}{2n\lambda_i^2})$, and $\det(\Sigma)^{1/2} = O_p((2n)^{-q/2} \prod_{i=1}^{q} |\lambda_i|^{-1})$. We have

$$
\int_{A_{1c}} \|g_q - g_0\|^2 L(g_q) p(g_q) db
$$

$$\begin{aligned}&\le O_p\left(e^{-n\epsilon_n^2}(\pi/n)^{q/2}\prod_{i=1}^{q}|\lambda_i|^{-1}\left(\|g_0\|^2+\sum_{i=1}^{q}\left(\frac{\mu_i^2}{\lambda_i^2}+\frac{1}{2n\lambda_i^2}\right)\right)\right)\\&= O_p\left(e^{-n\epsilon_n^2}(\frac{\pi}{n\lambda_q^2})^{q/2}\left(1+\frac{q}{n\lambda_q^2}\right)\right)\end{aligned}$$

where the last equality follows from that $\sum_{i=1}^{q}\mu_i^2/\lambda_i^2 \le \sum_{i=1}^{\infty}\mu_i^2/\lambda_i^2 = \|g_0\|^2 = O(1)$, and that $\lambda_1^2 \ge \lambda_2^2 \ge ... \ge \lambda_q^2$. Q.E.D.

Lemma C.3.4.

$$\int L(g_q)p(g_q)db \ge O_p\left(\exp(-n^{1-\alpha}-q/2-na_n^2)\left(\frac{q}{2n}\right)^{q/2}\right)$$

PROOF. A result to be used later is,

$$\max_{c_n\ge 1}\frac{1}{c_n}\left(\frac{\ln c_n}{n}-a_n^2\right)^{q/2} = e^{-(q/2+na_n^2)}\left(\frac{q}{2n}\right)^{q/2}$$

with optimal $c_n^* = e^{q/2+na_n^2}$. Define $\|T^*\| = \sup_{\|g\|=1}\|T^*g\|^2$, $\eta_n(b) = n\|T^*\|^2\|g_q-g_0\|^2 + na_n\|g_q\|^2$, and $B_n = \{b\in\mathbb{R}^q : \eta_n(b) < \ln c_n^*\}$. It can be shown that $B_n \subset \mathbb{R}^q$ is a ball centered at $a = (a_1, ..., a_q)$ with radius r, where

$$a_i = \frac{\|T^*\|^2 g_i}{a_n^2+\|T^*\|^2} = O(1),$$

$$r^2 = \frac{1}{na_n^2+n\|T^*\|^2}\left(\ln c_n^* - \frac{n\|T^*\|^2a_n^2}{a_n^2+\|T^*\|^2}\right) = O\left(\frac{\ln c_n^*}{n}-a_n^2\right)$$

Therefore $\mu(B_n) = O_p(r^q)$. Note that with optimal c^*, $r^2 = O(q/n) = o(1)$. By Assumption 4.4.4, $\sup_{g_q\in B_n}|G(g_q)-\bar{m}(g_q)^TV_0^{-1}\bar{m}(g_q)| < n^{-\alpha}$. We have

$$\int L(g_q)p(g_q)db = \int \exp(-n\bar{m}(g_q)^TV_0^{-1}\bar{m}(g_q) - na_n^2\|g_q\|^2)db$$

$$
\begin{aligned}
&\geq \int_{B_n} \exp(-nG(g_q) - na_n^2\|g_q\|^2 - n|G(g_q) - \bar{m}(g_q)^T V_0^{-1}\bar{m}(g_q)|)db \\
&\geq \exp(-n^{1-\alpha}) \int_{B_n} \exp(-n\|T^*(g_q - g_0)\|^2 - na_n^2\|g_q\|^2)db \\
&\geq \exp(-n^{1-\alpha}) \int_{B_n} \exp(-\eta_n(b))db \\
&\geq \exp(-n^{1-\alpha})c_n^{*-1}\mu(B_n) \\
&\geq O_p\left(\exp(-n^{1-\alpha})\frac{1}{c_n^*}\left(\frac{\ln c_n^*}{n} - a_n^2\right)^{q/2}\right) \\
&= O_p\left(\exp(-n^{1-\alpha} - q/2 - na_n^2)\left(\frac{q}{2n}\right)^{q/2}\right)
\end{aligned}
$$

Q.E.D.

Proof for Theorem 4.4.1

$$
\int \|g_q - g_0\|^2 p(g_q|Data)db = \frac{\int_{A_n}\|g_q - g_0\|^2 L(g_q)p(g_q)db + \int_{A_n^c}\|g_q - g_0\|^2 L(g_q)p(g_q)db}{\int L(g_q)p(g_q)db}
$$

By Lemma C.3.1, C.3.4, Assumption 4.4.5, and the definition of ϵ_n^2, $\int_{A_n}\|g_q - g_0\|^2 L(g_q)p(g_q)db = O_p(\epsilon/a_n)^{q+2}$. Therefore

$$
\frac{\int_{A_n}\|g_q - g_0\|^2 L(g_q)p(g_q)db}{\int L(g_q)p(g_q)db} = O_p\left(\left(\frac{2e\epsilon_n^2 n}{a_n^2 q}\right)^{q/2}\left(\frac{\epsilon_n^2}{a_n^2}\right)\right) = o_p(1)
$$

The last equality follows from that $\epsilon_n^2/a_n^2 = O(a_n^2/\lambda_{s_n}^2) = o(1)$, and $\epsilon_n^2 n/(qa_n^2) = O(na_n^2/q\lambda_s^2) = o(1)$, since $q \succ na_n^2/\lambda_s^2$ by Assumption 4.4.5.

By Lemma C.3.2,

$$
\begin{aligned}
&\frac{\int_{A_{2c}}\|g_q - g_0\|^2 L(g_q)p(g_q)db}{\int L(g_q)p(g_q)db} = O_p\left(\exp(2n^{1-\alpha} + q/2 + na_n^2 - n\epsilon_n^2)\left(\frac{2n\epsilon_n^2}{qa_n^2}\right)^{q/2}\right) \\
&= O_p\left(\frac{2e\epsilon_n^2 n}{a_n^2 q}\right)^{q/2} = o_p(1)
\end{aligned}
$$

By Lemma C.3.3,

$$\frac{\int_{A_{1c}} \|g_q - g_0\|^2 L(g_q)p(g_q)db}{\int L(g_q)p(g_q)db} = O_p\left(\exp(n^{1-\alpha} + q/2 + na_n^2 - n\epsilon_n^2)\left(\frac{2\pi}{q\lambda_q^2}\right)^{q/2}(1+\frac{q}{n\lambda_q^2})\right)$$

$$\le O_p\left(\left(\frac{2\pi e}{q\lambda_q^2}\right)^{q/2}(1+\frac{q}{n\lambda_q^2})\right) = o_p(1)$$

The last equality follows from the fact that: at one hand, $2\pi e/(q\lambda_q^2) = o(1)$ since $q \succ \lambda_q^2$ by assumption. On the other hand, define $d_n = \left(\frac{2\pi e}{q\lambda_q^2}\right)^{q/2}\frac{q}{n\lambda_q^2}$. It's left to show $d_n = o(1)$. In fact, since $q = q_n \to \infty$, then for some constant $C > 0$,

$$\begin{aligned}
\ln d_n &= -\left(\frac{q}{2} - 1\right)\ln q - \left(\frac{q}{2} + 1\right)\ln\lambda_q^2 + \frac{q}{2}C - \ln n \\
&= O\left(-\frac{q}{2}\ln q - \frac{q}{2}\lambda_q^2 + \frac{q}{2}C - \ln n\right) \\
&\le O(-q\ln q - q\ln\lambda_q^2 + qC) = O(-q(\ln q - C) - q\ln\lambda_q^2) \\
&= O(-q\ln q - q\ln\lambda_q^2) = O(-q\ln q\lambda_q^2)
\end{aligned}$$

We have $-q\ln q\lambda_q^2 \to -\infty$ since $q \succ \lambda_q^{-2}$. Therefore $\ln d_n \to -\infty$, which implies $d_n = o(1)$. Q.E.D.

APPENDIX D

Technical Proofs for Chapter 5

D.1. Proof of Theorem 5.3.1

Proof. Define $L(\mu) = \sum_{i=1}^{n} \log[1 + \mu(|Y_i - C(X_i, \theta)| - r)]$. Let n_1 be the number of i such that $|Y_i - C(X_i, \theta)| = 0$, and n_2 be the number of i such that $|Y_i - C(X_i, \theta)| = 1$. Note that $\hat{R} = n^{-1} \sum_{i=1}^{n} |Y_i - C(X_i, \theta)| = n^{-1} n_2$, and $n_1 + n_2 = n$. Since $|Y_i - C(X_i, \theta)| \in \{0, 1\}$, we have

$$
\begin{aligned}
L(\mu) &= \sum_{i:|Y_i - C(X_i,\theta)|=0} \log(1 - \mu r) + \sum_{i:|Y_i - C(X_i,\theta)|=1} \log(1 + \mu(1 - r)) \\
&= n_1 \log(1 - \mu r) + n_2 \log(1 + \mu(1 - r))
\end{aligned}
$$

Differentiating $L(\mu)$ and setting to zero, it is straightforward to verify that the optimal $\mu^* = \frac{n_2 - nr}{nr(1-r)}$, with $\max_\mu L(\mu) = (n - n_2) \log \frac{n - n_2}{n(1-r)} + n_2 \log \frac{n_2}{nr}$. Replacing n_2 by $n\hat{R}$ yields $\max_\mu L(\mu) = nK(\hat{R}, r)$. Finally, we verify that the second derivative $L''(\mu^*) = -\frac{n^3(1-r)^2 r^2}{n_2 n_1} < 0$ when $r \in (0, 1)$. Q.E.D.

D.2. Proof of Theorem 5.3.2

The following lemmas establish some relationships between the expression in the log empirical likelihood and the square (or absolute value) distances. The first lemma is for the special case of classification and the second one is more general.

Lemma D.2.1. *For $p, q \in [0, 1]$, and $K(p, q)$ as defined in (5.8),*

$$0.5(q-p)^2 \le K(p,q) \le \{\min(p,q,1-p,1-q)\}^{-2} 0.5(p-q)^2. \tag{D.1}$$

Proof. This is straightforward by a second order Taylor expansion of $-\ln(1+\delta_{1,2})$, where $\delta_1 = q/p - 1$ and $\delta_2 = (1-q)/(1-p) - 1$. Q.E.D.

Lemma D.2.2. *Let $m_i,\ i = 1, ..., n$, be iid random variables. Suppose positive constants δ^+, δ^-, M are such that $P(m_1 > \delta^+) > 0$, $P(m_1 < -\delta^-) > 0$, and $P(|m_1| \le M) = 1$. Then, with probability P^* at least $1 - \pi_n$ for some $\pi_n > 0$ exponentially small in n, we have*

$$(\min\{\delta^+, \delta^-\})^{-1}|\bar{m}| \ge \max_{\mu}\left\{\overline{\ln(1+\mu m)}\right\} \ge (7/36)M^{-2}(\bar{m})^2, \tag{D.2}$$

where overlines represent sample averages.

Proof of Theorem 5.3.2

Denote $R = E^*\rho(W, a)$ and $\Delta = \sup_a |\hat{R} - R|$, then

$P^*[|R - r| > \epsilon | D] \le \int I(|R-r| > \epsilon) I(\Delta \le \epsilon/2) e^{-nK(\hat{R},r)} d\pi / \int e^{-nK(\hat{R},r)} d\pi + I(\Delta > \epsilon/2)$. The numerator of the first term is less than $e^{-n\epsilon^2/8}$ since $|\hat{R} - r| \ge |R - r| - \Delta > \epsilon/2$ and this implies $K(\hat{R}, r) > \epsilon^2/8$ due to a previous lemma.

The denominator is bounded by

$\int e^{-nK(\hat{R},r)} d\pi \ge \int I(|R-r| \le \delta) I(\Delta \le \delta/2) I(\eta \ge \tau) e^{-nK(\hat{R},r)} d\pi$

$\ge e^{-n(\tau-\delta/2)^{-2}(9/8)\delta^2} \pi(|R-r| \le \delta, \eta \ge \tau) I(\Delta \le \delta/2)$, where $\eta = \min(R, 1-R, r, 1-r)$ and δ and τ are some positive constants. Here we used again a previous lemma to bound $K(\hat{R}, r) \le \{\min(\hat{R}, 1-\hat{R}, r, 1-r)\}^{-2} 0.5(\hat{R}-r)^2 \le (\tau - \delta/2)^{-2} 0.5(\hat{R}-r)^2 \le (\tau - \delta/2)^{-2} 0.5(\delta + \delta/2)^2$.

Combining these we obtain: the event $\Delta \leq \min\{\delta/2, \epsilon/2\}$ implies the event

$$P(|R-r| > \epsilon | D) \leq \frac{e^{-n\epsilon^2/8+(9n/8)(\delta/(\tau-\delta/2))^2}}{p(|R-r| \leq \delta, \eta \geq \tau)}.$$

Note that $p(|R-r| \leq \delta, \eta \geq \tau) > 0$ by assumption. Choose constants τ and δ suitably, then the right hand side can be made arbitrarily close to zero (and exponentially small in n). This happens with P^*, the probability in D being at least $P^*(\Delta \leq \min\{\delta/2, \epsilon/2\})$, which converges to 1 by assumption. Q.E.D.

Vita

Yuan Liao

PERSONAL INFORMATION

- Date of Birth: October 1st, 1982
- Place of Birth: Beijing, China
- Marital Status: Single

EDUCATION

- 2005 B.S. in Mathematics, Tsinghua University, Beijing
- 2010 Ph.D. (expected) in Statistics, Northwestern University, Evanston, IL

PUBLICATION

- Bayesian analysis in moment inequality models (with Wenxin Jiang) (2010) *The Annals of Statistics*. **38** 275-316.

PROFESSIONAL AFFILIATIONS

- Member of American Statistical Association. 2006-Present
- Member of Econometric Society. 2009-Present