ESSAYS IN SEMIPARAMETRIC ECONOMETRICS AND PANEL DATA ANALYSIS

by

Martin Burda

BSc. (joint), Uppsala University, 1999

Bc. (joint), University of Economics, Prague, 2000

M.A., McGill University, 2001

Submitted to the Graduate Faculty of

the Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

ARTS AND SCIENCES

This dissertation was presented

by

Martin Burda

It was defended on

April 25, 2007

and approved by

Mehmet Caner, Department of Economics

Jean-François Richard, Department of Economics

Daniel Berkowitz, Department of Economics

Roman Liesenfeld, Professor, University of Kiel, Germany

Dissertation Advisors: Mehmet Caner, Department of Economics,

Jean-François Richard, Department of Economics

ESSAYS IN SEMIPARAMETRIC ECONOMETRICS AND PANEL DATA ANALYSIS

Martin Burda, PhD

University of Pittsburgh, 2007

Limited dependent variable (LDV) panel data models pose substantial challenges in maximum likelihood estimation. The likelihood function in such models typically contains multivariate integrals that are often analytically intractable. To overcome such problem in a panel probit model with unobserved individual heterogeneity and autocorrelated errors, in Chapter 1 - co-authored with Roman Liesenfeld and Jean-François Richard - we perform classical and Bayesian analysis of the model based on the Efficient Importance Sampling (EIS) technique (Richard and Zhang, 2006). We apply our method to the product innovation activity of a panel of German manufacturing firms in response to imports and foreign direct investment confirming their positive effects. Nonetheless, our key coefficient estimates are smaller than found in previous literature which can be explained by our flexible model assumptions. The remaining two chapters present my work on new estimation methods for models based on conditional moment restrictions. Such models are frequently stipulated by economic theory but only a few estimators based directly on them have so far been analyzed in the literature. Indeed, estimation of parameters therein poses a difficult ill-posed inverse problem. Rather, these models are typically converted into unconditional moment restrictions that are easier to handle. However, such conversion results in a loss of information compared to the original specification. Using the information-theoretic framework of so-called Generalized Minimum Contrast (GMC) estimation, in Chapter 2 I propose a new class of estimators based directly on conditional moment restrictions that encompasses the entire GMC family. Moreover, I show that previous literature covering a few special cases of the GMC class use an arbitrary

uniform weighting scheme over the space of exogenous variables that can be improved upon with optimal local weighting. All currently available GMC estimators are based on moments containing finite-dimensional Euclidean parameters. To alleviate a potential misspecification problem resulting from strong parametric assumptions, in Chapter 3 I propose a new Sieve-based Locally Weighted Conditional Empirical Likelihood (SLWCEL) estimator containing also infinite dimensional unknown functions, thus extending a special case of Chapter 2 to the semiparametric environment. Much of Chapter 3 is devoted to analysis of SLWCEL's asymptotic properties.

TABLE OF CONTENTS

PREFACE x

1.0 INTRODUCTION 1

2.0 PANEL DATA PROBIT MODEL 5

2.1 EMPIRICAL EXAMPLE 8

2.2 EXISTING PANEL PROBIT MODEL SPECIFICATIONS 9

2.2.1 Model 1: Pooled Probit 10

2.2.2 Model 2: Panel Probit with Autocorrelated Errors 10

2.2.3 Model 3: Random Parameters Model 11

2.2.4 Model 4: Latent Class, Finite Mixture Model 12

2.3 ALTERNATIVE PANEL PROBIT MODEL 13

2.3.1 EIS Evaluation of the Likelihood 14

2.3.2 Bayesian MCMC Approach Based on EIS 16

2.4 EMPIRICAL RESULTS 17

2.5 CONCLUSION 21

APPENDIX 26

Appendix 2.1: EIS Likelihood Evaluation 26

Appendix 2.2: Sampling from Posterior Densities 29

3.0 LOCALLY WEIGHTED GENERALIZED MINIMUM CONTRAST ESTIMATION UNDER CONDITIONAL MOMENT RESTRICTIONS 34

EXISTING METHODS 37

Information-theoretic Approaches to Estimation 37

Unconditional Moment Restrictions 38

Conditional Moment Restrictions 38
CONDITIONAL MOMENT RESTRICTIONS: ALTERNATIVE ESTIMATION METHOD 39
Stochastic Environment 40
Information-theoretic Model of the Conditional GMC Problem 40
Locally Weighted Conditional Empirical Likelihood 44
Other GMC Class Members 47
SIMULATION 48
CONCLUSION 50
4.0 SIEVE-BASED EMPIRICAL LIKELIHOOD 52
SIEVE-BASED CONDITIONAL EMPIRICAL LIKELIHOOD 55
CONSISTENCY 59
CONVERGENCE RATES 61
ASYMPTOTIC NORMALITY 65
CONCLUSION 70
APPENDIX 71
Appendix 4.1: Proofs of Main Results 71
Appendix 4.2: Auxiliary Results 79
Consistency 79
Convergence Rates 89
Asymptotic Normality 90
Appendix 4.3: Sieve Conditional Variance Proof 92
5.0 BIBLIOGRAPHY 98

LIST OF TABLES

2.1 Panel Probit - Models 1-3 19
2.2 Panel Probit - Model 4 20
2.3 Panel Probit - EIS-SML and EIS-MCMC 22
3.1 Simulation Results 50

LIST OF FIGURES

2.1 Marginal Posterior Densities 23
2.2 Descriptive Histograms for the Data 24
2.3 Autocorrelation Functions of the Parameters 25
3.1 Sample Simulated Data 49
3.2 Plot of sigma against x 51

PREFACE

I would like to thank my advisors Mehmet Caner and Jean-François Richard for their guidance and inspiration that has nurtured me throughout my graduate career. Special thanks also goes to Dan Berkowitz, Roman Liesenfeld, Nese Yildiz and Dharmarajan Hariharan for helpful comments and discussions. Finally, I would like to express my deep gratitude to my parents and my fiancée Vera for their eternal love and patience. Without any of you, this project would not have been possible.

1.0 INTRODUCTION

Limited dependent variable (LDV) panel data models pose substantial challenges in maximum likelihood estimation. The likelihood function in such models typically contains multivariate integrals that are often analytically intractable. This obstacle is usually overcome with the use of simulation methods that replace integrals with computationally inexpensive Monte Carlo (MC) estimates. Highly accurate smooth probability simulators are indispensable for a successful implementation of MC estimators. In particular, the recently developed Efficient Importance Sampling (EIS) technique (Richard and Zhang, 2000, 2006) has been found highly competitive with previous alternatives.

In Chapter 1 − co-authored with Roman Liesenfeld and Jean-François Richard − we perform an EIS-based classical and Bayesian analysis of a panel probit model with unobserved individual heterogeneity and autocorrelated errors. We do not impose any orthogonality condition on the unobserved individual effects with respect to the observed regressors. In the LDV context, the classical EIS-based approach has been implemented for analyzing a binary logit panel data model in Richard and Zhang (2006) as a Monte Carlo simulation pilot study, and in Liesenfeld and Richard (2006b) as an application in estimating a model of union/non-union decision of young men. Here we adopt the procedure to the panel probit case. A Bayesian analysis of an LDV model under our flexible assumptions has, to our knowledge, thus far not been performed and represents a methodological contribution. We embed EIS within the Markov Chain Monte Carlo (MCMC) simulation method to perform posterior analysis. Specifically, we implement the Gibbs sampling scheme where we augment the data with latent variables. We sample the unobserved individual heterogeneity component as N individual Gibbs blocks drawing from a piece-wise linear approximation to the marginal posterior density constructed with a nonparametric form of EIS. The time

effects are simulated as another Gibbs block with a parametric EIS proposal density for an Acceptance-Rejection Metropolis-Hastings step.

We apply our method to the product innovation activity of a panel of German manufacturing firms in response to imports, foreign direct investment and other control variables. The same dataset was analyzed by Bertschek and Lechner (1998) and Greene (2004) for different types of estimators under more restrictive assumptions providing a useful benchmark for comparison with our results. Our findings confirm the positive effect of imports and FDI on firms' innovation activity found in previous literature. However, our coefficient estimates of these variables were smaller than the ones reported in the benchmark studies. This discrepancy can be explained by the exclusion of three far outliers from our estimation and also by our flexible model assumptions relative to previously utilized models.

The remaining two chapters present my work on proposing an analyzing new estimation methods in the realm of models based on moment restrictions. In particular, economic theory frequently stipulates *conditional* moment restrictions as a model basis for estimation and inference in various economic problems. However, since estimation of parameters in such models in general poses a difficult ill-posed inverse problem, these models are typically converted into *unconditional* moment restrictions that are much easier to handle. The conversion is usually performed by multiplying the vector of moment functions with an arbitrary matrix-valued function of instruments. This procedure is used under the presumption that the chosen instruments identifies the model parameters which may not be true even if the parameters are identified in the conditional model. Moreover, the conversion to unconditional moments results in a loss of efficiency with respect to the information contained in the conditional moments.

The methods typically employed for estimation of the resulting unconditional model have also been subject to criticism. While the optimally-weighted two-step GMM is first-order asymptotically efficient, its finite sample properties have been increasingly recognized as relatively poor. A number of alternative estimators, such as the Empirical Likelihood, have been suggested to overcome this problem. These alternative estimators have been shown to fall into broader families of estimators such as the Generalized Empirical Likelihood (GEL) estimators and the Generalized Minimum Contrast (GMC) estimators that share numer-

ous common properties. The GEL/GMC estimators circumvent the need for estimating a weight matrix in the two-step GMM procedure by directly minimizing a discrepancy measure between the estimated distribution and the empirical distribution. Specifically, the GMC family is derived on the basis of an information-theory-based concept of closeness between probability measures. A growing body of Monte Carlo evidence has revealed favorable finite-sample properties of the special cases of the GEL/GMC estimators compared to the GMM. Specifically, the Empirical Likelihood has been singled out as being higher-order efficient relative to other GEL/GMC estimators (Newey and Smith, 2004).

Most of the GEL/GMC estimators analyzed in previous literature are based on unconditional moment restrictions subjected to the criticism mentioned above. In addressing this problem, Kitamura, Tripathi, and Ahn (2004) (KTA) recently developed a Conditional Empirical Likelihood (CEL) estimator that makes efficient use of the information contained in conditional moment restrictions. Their one-step estimator achieves the semiparametric efficiency bound without explicitly estimating the optimal instruments. Similar analysis has been performed by other special cases of GEL/GMC: Antoine, Bonnal, and Renault (2006) for the case of Conditional Euclidean Likelihood and Smith (2003, 2006) for the Cressie-Read family of estimators.

Using the GMC information-theoretic framework, in Chapter 2 I extend this line of research by proposing a new class of estimators based directly on conditional moment restrictions that encompasses the entire GMC family. Moreover, I show that in constructing their special cases the previous literature use an arbitrary uniform weighting scheme over the space of exogenous variables. This leads to minimizing a discrepancy from a probability measure that is different, almost surely, from the one under which the data was distributed. The reason for this phenomenon is that the previously analyzed estimators were all based on simple local kernel smoothing of the *unconditional* moment restrictions model over the exogenous variables. In contrast, in deriving the new class of conditional GMC estimators I consider an information-theoretic dual locally weighted GMC optimization problem built directly on the *conditional* moment restrictions that minimizes a discrepancy from a probability measure according to which the data was distributed. As a result, the newly proposed class of estimators not only includes the previously analyzed conditional estimators as spe-

cial cases but seeks to replace them with locally weighted alternatives that improve on the former in terms of finite sample properties. Particular attention is devoted to the Locally Weighted Conditional Empirical Likelihood (LWCEL) based on the conjecture that its desirable higher-order efficiency found in the unconditional case will carry over to the conditional environment. I analyze the differences between the new LWCEL and KTA's CEL in detail and show in a Monte Carlo study that the LWCEL estimator exhibits better finite-sample properties than the CEL. Asymptotic properties of the LWCEL are considered as a special case of the ones derived for its semiparametric extension in the following Chapter.

All currently available GMC/GEL estimators analyzed in the literature are based on moment conditions containing finite-dimensional Euclidean parameters. Such models impose relatively strong restrictions in assuming that social phenomena occur in a certain specific way. Yet, economic theories seldom produce exact functional forms warranting purely parametric models, and misspecifications in functional forms may lead to inconsistent parameter estimates. By specifying the model partially, i.e. by including an unknown function as a part of the unknown parameters, the inconsistency problem can be alleviated. For this purpose, in Chapter 3 I propose a new Sieve-based Locally Weighted Conditional Empirical Likelihood (SLWCEL) estimator for models of conditional moment restrictions containing finite-dimensional unknown parameters and infinite dimensional unknown functions, extending the LWCEL analyzed in Chapter 2. I first derive consistency of the SLWCEL under a general metric. Then I show that the estimator converges to its population counterpart under the Fisher metric sufficiently fast to yield asymptotic normality of SLWCEL's parametric component.

The GMC/GEL-based SLWCEL is a one-step information-theoretic alternative to the two-step Sieve Minimum Distance (SMD) estimator analyzed by Ai and Chen (2003). The SMD estimator is the only current simultaneous estimation technique that can be used to estimate models of semiparametric conditional moment restrictions. The SMD's founding optimization problem of minimizing the distance between vectors of moment conditions is akin to the one used in the parametric GMM estimators. Hence, development of an alternative GMC/GEL-based estimator appears desirable in the light of the above-mentioned comparisons of parametric GMM - GMC/GEL estimators promulgated in the literature.

2.0 PANEL DATA PROBIT MODEL

Full title: **Classical and Bayesian Analysis of a Probit Panel Data Model with Unobserved Individual Heterogeneity and Autocorrelated Errors**

Co-authored with Roman Liesenfeld and Jean-François Richard

It has long been recognized that maximum likelihood analysis of limited dependent variable (LDV) models with panel data is feasible only under relatively restrictive assumptions (Butler and Moffitt, 1982). The difficulty that such models pose in general lies in the likelihood function containing multivariate integrals that are often analytically intractable. This obstacle is typically overcome with the use of simulation methods (see e.g. Geweke and Keane, 2001, and references therein) that replace integrals with computationally inexpensive Monte Carlo (MC) estimates. By the law of large numbers, such integral estimates can be made arbitrarily accurate by increasing the size of the simulated data.

Simulation-based estimation methods for LDV models generally take one of two approaches (Hyslop, 1999). The first approach, often called the Simulated Maximum Likelihood[1] (SML), involves obtaining an unbiased simulator[2] for the likelihood function and maximizing the resulting log simulated likelihood function instead of the actual likelihood function. The second approach utilizes simulation of an expression for the score of the likelihood. Two leading examples are the Method of Simulated Moments (MSM) estimator (McFadden, 1989) and the Method of Simulated Scores (MSS) estimator (Hajivassiliou and McFadden, 1998). Under the MSM estimator, the score of the likelihood is first expressed as a moment condition, the moment condition is then simulated and the estimator solves for the

[1] Gourieroux and Monfort (1996) provide the essential statistical background for the SML estimator.

[2] Here we refer to a method of drawing random numbers involving an appropriate density for the random draws.

root of the simulated condition. The MSS estimator solves for the root of the simulated score directly. Based on available MC evidence, Geweke and Keane (2001, p. 3505)report that "in most contexts the choice between SML and MSM is not important."[3] On the other hand, Hyslop (1999, p. 1268-1269) expresses preference for SML based on ease of implementation, numerical stability and computational burden. Notably, while SML is comparatively simple to implement, "MSM and MSS often require significant manipulation of the score function."

For a successful implementation of any of these estimators, it is essential to use a highly accurate smooth probability simulator. Among the currently available methods, the GHK simulator[4] (developed by Geweke, 1991; Börsch-Supan and Hajivassiliou, 1993; Keane, 1994) is the most popular one and it has been reported to perform very well in MC studies for simulating the multivariate normal choice probabilities (see Geweke and Keane, 2001, and references therein). However, the recently developed Efficient Importance Sampling technique (Richard and Zhang, 2000, 2006) has been found highly competitive with the GHK sampler. Zhang and Lee (2004) show in an MC study that while the performance of GHK-SML and EIS-SML is comparable for short panels ($T = 8$), for longer panels ($T > 50$) the GHK-SML estimates of the lagged dependent variable coefficient and the serial correlation coefficient are biased (upward and downward, respectively), while the EIS-SML estimates avoid this bias. The appealing theoretical justification for EIS is one of minimizing the MC sampling variance in construction of the SML whereas the GHK simulator lacks this property. Moreover, the EIS sequential implementation (Danielsson and Richard, 1993; Richard and Zhang, 2006) is well suited for evaluation of likelihood functions expressed as integrals with very high dimensions (>1,000).

In this paper, we perform EIS-SML classical and Bayesian analysis of a panel probit model with unobserved individual heterogeneity and autocorrelated errors. We do not impose any orthogonality condition on the unobserved individual effects with respect to the observed

[3] These authors note that one known exception is the case of panel data models with serially correlated errors - the type of models considered in this paper. This conclusion is based on a study by Lee (1997)that compared the perfomance of SML and MSM based on the GHK simulator and found GHK-SML serial correlation parameters severely biased relative to GHK-MSM. However, in this paper we use a different simulator, the EIS, which has been found to improve on the GHK simulator in terms of bias (Zhang and Lee, 2004).

[4] It is sometimes also called the Smooth Recursive Conditioning (SRC) simulator (Börsch-Supan and Hajivassiliou, 1993).

regressors. Our model thus falls outside of the class of what is called in the traditional econometric parlance "random effects" models (Wooldridge, 2001, p. 252).

In the LDV context[5], the classical EIS-SML approach has been implemented in Richard and Zhang (2006) as a binary logit model in a Monte Carlo simulation pilot study, and in Liesenfeld and Richard (2006b)analyzing the union/non-union decision of young men with the data set of Vella and Verbeek (1998). Here we adopt the EIS-SML procedure to the panel probit case. Two other studies that used the SML method for the panel probit model with the same assumptions as ours are Falcetti and Tudela (2006), and Hyslop (1999). However, these authors utilized the competing GHK simulator which is tantamount to using a different estimation technique in the construction of the simulated log likelihood function.

In the Bayesian part, we embed EIS within the Markov Chain Monte Carlo (MCMC) simulation method to perform posterior analysis. Specifically, we implement the Gibbs sampling scheme where we augment the data with latent variables. We sample the unobserved individual heterogeneity component as N individual Gibbs blocks drawing from a piece-wise linear approximation to the marginal posterior density constructed with a nonparametric form of EIS. The time effects are simulated as another Gibbs block with a parametric EIS proposal density for an Acceptance-Rejection Metropolis-Hastings step. The general approach to augmented Gibbs sampling has been implemented in Liesenfeld and Richard (2003, 2006a) in models of stochastic volatility for sampling the autocorrelated error component. However, Bayesian analysis of an LDV model with unobserved heterogeneity and autocorrelated errors has, to our knowledge, thus far not been performed and represents a methodological contribution of this paper. The use of nonparametric EIS represents another novel feature.

We apply our method to the product innovation activity of a panel of German manufacturing firms in response to imports, foreign direct investment and other control variables. The same dataset was analyzed by Bertschek and Lechner (1998) and Greene (2004) for different types of estimators under more restrictive assumptions providing a useful benchmark for comparison with our results.[6] Specifically, Bertschek and Lechner (1998) proposed sev-

[5]The EIS technique has been successfully implemented in other models, specifically stochastic volatility models (Liesenfeld and Richard, 2003, 2006a), dynamic parameter models involving counts (Jung and Liesenfeld, 2001), and stochastic autoregressive duration models (Bauwens and Hautsch, 0003).

[6]Similar data set was used in an interesting paper by Inkmann (2000) but with some regressors different from ours.

eral variants of a GMM estimator based on the period specific regression functions. Greene (2004) performed maximum likelihood analysis with GHK-SML and the Butler and Moffitt (1982) Hermite quadrature method. None of these authors considered a model with unobserved individual heterogeneity and autocorrelated errors as analyzed in this paper.

2.1 EMPIRICAL EXAMPLE

The goal of our empirical application is to investigate firms' innovative activity as a response to imports and foreign direct investment (FDI). This problem was originally considered in Bertschek (1995) who suggested that imports and inward FDI had a positive effect on the innovative activity of domestic firms. The rationale behind this argument is that imports and FDI represent a competitive threat to domestic firms. Competition on the domestic market is enhanced and the profitability of the domestic firms might be reduced. Consequently, these firms have to produce more efficiently. One possibility to react to this competitive threat is to increase innovative activity.

The analyzed dataset contains $N = 1270$ cross-section units observed over $T = 5$ time periods. The dependent variable y_{it} in the data takes the value one if a product innovation occurred within the last year and the value zero otherwise. The $K-$vector of control variables is denoted by $\underline{z}_{it}$ and the corresponding vector of parameters to be estimated by $\underline{\beta}$. The independent variables refer to the market structure, in particular the market size of the industry ($\ln(sales)$), the shares of imports and FDI in the supply on the domestic market (*import share* and FDI *share*), the *productivity* as a measure of the competitiveness of the industry as well as two variables indicating whether a firm belongs to the *raw materials* or to the *investment goods* industry. Also, including the *relative firm size* accounts for the innovation – firm size relation often discussed in the literature. All variables with exception of the firm size are measured at the industry level. Descriptive statistics and further discussion appear in Bertschek and Lechner (1998) and Greene (2004).

Two distinct sources of time dependence have been identified in the literature.[7] In the

[7] An illuminating discussion is provided in Falcetti and Tudela (2006, p. 454), drawing on Heckman (1981)

context of our empirical application, the first arises from the possibility that innovation occurring in the present period may alter the conditions for the occurrence of innovation in the next period. In this case past experience has a behavioral effect in the sense that otherwise identical company that did not experience the event would behave differently from the company that experienced the event. This phenomenon is known as *true state dependence* and is typically captured by including a lagged dependent variable among the regressors.

The second source of time dependence derives from the fact that companies may differ in their propensity to innovate. Two components are distinguished in this case. The first one relates to the existence of company-specific attributes that are time-invariant. This component is typically called *unobserved heterogeneity* and we allow for it by including a time-invariant company-specific error term τ_i. It may reflect institutional factors that are difficult to control for by direct inclusion among the regressors. The second component takes into account that economy-wide factors influencing all companies alike may be correlated over time. Improper treatment of the error structure may result in a conditional relationship between future and past experience that is termed *spurious state dependence* (Hyslop, 1999). We avoid this problem by assuming an $AR(1)$ structure for the latent error term λ_t.

2.2 EXISTING PANEL PROBIT MODEL SPECIFICATIONS

The panel probit model has been analyzed extensively under various assumptions in the literature. In this Section, in addition to the basic probit model, we briefly review two studies, Bertschek and Lechner (1998) and Greene (2004), which used the same dataset as in this paper and are therefore of particular relevance as benchmarks for discussion of our results. In doing so, we present only the least restrictive models of the ones analyzed by these authors.

and Börsch-Supan et al. (1992).

2.2.1 Model 1: Pooled Probit

This is the simplest probit estimator that treats the entire sample as if it were a large cross-section. Specifically, it postulates the latent variable probit model specification

$$y_{it}^{*} = \underline{\beta}_0' \underline{z}_{it} + \epsilon_{it} \tag{2.1}$$

with the observation rule

$$y_{it} = \mathbf{1}\left(y_{it}^{*} \geq 0\right), \qquad i : 1, ..., N \; ; \qquad t : 1, ...T \tag{2.2}$$

where $\mathbf{1}\left(\cdot\right)$ denotes the indicator function. The error terms ϵ_{it} are normally distributed with zero mean and unit variance.

2.2.2 Model 2: Panel Probit with Autocorrelated Errors

Bertschek and Lechner (1998) assume the latent variable probit model specification (2.1) with the observation rule (2.2). However, their error terms $\underline{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{iT})'$ are modeled as jointly normally distributed with mean zero and covariance matrix Σ. Also, $\underline{\epsilon}_i$ are independent of the explanatory variables which implies strict exogeneity of the latter. The error terms may be correlated over time for a given firm, but uncorrelated over firms. The diagonal elements of Σ are set to unity to facilitate identification of $\underline{\beta}$ and the off-diagonal elements are considered nuisance parameters. On the basis of the model (2.1) Bertschek and Lechner (1998) formulated the following set of moment conditions

$$\begin{aligned} E[W(Z, \beta_0)|X] &= 0 \\ W(z, \beta) &= [w_1(Z_1, \beta), ..., w_T(Z_T, \beta)]' \\ w_t(Z_t, \beta) &= Y_t - \Phi(\underline{\beta}' \underline{z}_{it}) \end{aligned} \tag{2.3}$$

where Φ denotes the CDF of a univariate normal distribution. The main advantage of using these moments is that their evaluation does not require multidimensional integration and they do not depend on the $T(T-1)/2$ off-diagonal elements of Σ. In line with the GMM literature, (2.3) implies

$$E\{A(X)W(Z, \beta_0)\} = 0$$

where $A(X)$ is a $P \times T$ matrix of instrumental variables. The efficient GMM estimator of β_0 is then defined as

$$\widehat{\beta}_N = \arg\min_{\beta} g_N'(\beta)\Omega^{-1}g_N(\beta) \tag{2.4}$$

where

$$g_N(\beta) = \frac{1}{N}\sum_{i=1}^{N} A(x_i)W(Z_i, \beta)$$

Bertschek and Lechner (1998) obtained a nonparametric estimate of the optimal weighting matrix Ω using a k-nearest neighbor (k-NN) approach.

2.2.3 Model 3: Random Parameters Model

Greene (2004) noted that the dataset used contains a considerable amount of between group variation (97.6% of the FDI variation and 92.9% of the imports share variation is accounted for by differences in the group means). Thus, the dataset was likely to contain significant degree of unobserved individual heterogeneity, while none of the models above accounted for it. Greene (2004) suggested two alternative formulations of the panel probit model: the Random Parameters Model and the Latent Class Model (discussed further below). The Random Parameters Model (or 'Hierarchical' or 'Multilevel' Model) is based on the latent variable probit model specification

$$y_{it}^* = \underline{\beta}_0' z_{it} + \epsilon_{it}$$

with the observation rule (2.2), $\epsilon_{it} \sim NID[0, 1]$, and

$$\beta_i = \mu + \Delta z_i + \Gamma w_i$$

where μ is $K \times 1$ vector of location parameters, Δ is $K \times L$ matrix of unknown location parameters, Γ is $K \times K$ lower triangular matrix of unknown variance parameters, z_i is $L \times 1$ vector of individual characteristics, w_i is $K \times 1$ vector of random latent individual effects. It holds that $E[w_i|X_i, z_i] = 0$ and $Var[w_i|X_i, z_i] = V$, a $K \times K$ diagonal matrix of known constants. Hence $E[\beta_i|X_i, z_i] = \mu + \Delta z_i$ and $Var[\beta_i|X_i, z_i] = \Gamma V \Gamma'$. Conditional on w_i,

observations of y_{it} are independent across time; timewise correlation would arise through correlation of elements of β_i. The joint conditional density on y_{it} is

$$f\left(y_i|X_i,\beta\right) = \prod_{t=1}^{T} \Phi[(2y_{it}-1)\underline{\beta}'\underline{z}_{it}] \tag{2.5}$$

The contribution of this observation to the log-likelihood function for the observed data is obtained by integrating the latent heterogeneity out of the distribution. Thus

$$\log L = \sum_{i=1}^{N} \log L_i = \sum_{i=1}^{N} \log \int_{\beta_i} \prod_{t=1}^{T} \Phi[(2y_{it}-1)\underline{\beta}'\underline{z}_{it}] g(\beta_i|\mu,\Delta,\Gamma,z_i)d\beta_i \tag{2.6}$$

Estimates of μ, Δ and Γ are obtained by maximizing the SML version of (2.6).

2.2.4 Model 4: Latent Class, Finite Mixture Model

This model arises if we assume a discrete distribution for β_i instead of the continuous one postulated in the previous Random Parameters Model. Alternatively, the Latent Class model can be viewed as arising from a discrete, unobserved sorting of firms into groups, each of which has its own set of characteristics. If the distribution of β_i has finite, discrete support over J points (classes) with probabilities $p(\beta_j|\mu,\Delta,\Gamma,z_i)$, $j = 1,...,J$, then the resulting formulation of the analog of L_i from (2.6) is

$$L_i = \sum_{j=1}^{J} p(\beta_j|\mu,\Delta,\Gamma,z_i) f\left(y_i|X_i,\beta_j\right)$$

The model can then be estimated using the EM algorithm (see Greene, 2004, for details).

2.3 ALTERNATIVE PANEL PROBIT MODEL

Our panel probit model differs from the ones described above by an explicit inclusion of variables for both individual unobserved heterogeneity and time effects accounting for spurious state dependence. Specifically, our standardized probit model specification assumes a latent variable regression for individual i and time period t

$$y_{it}^{*} = \underline{\beta}' \underline{z}_{it} + \tau_i + \lambda_t + \epsilon_{it}, \qquad i : 1, ..., N \; ; \qquad t : 1, ...T \tag{2.7}$$

under the observation rule (2.2), where $\underline{z}_{it}$ is a vector of explanatory variables and $\epsilon_{it} \sim N(0,1)$ is a stochastic error component uncorrelated with any other regressor. $\tau_i \sim N(0, \sigma_\tau^2)$ represents individual unobserved heterogeneity that can be arbitrarily correlated with other regressors. λ_t captures latent time effects and is assumed to follow a stationary autoregressive process

$$\lambda_t = \rho \lambda_{t-1} + \eta_t$$

where $\eta_t \sim N(0, \sigma_\eta^2)$ such that the mean of λ_t is zero and the variance σ_λ^2 is stationary. It is assumed that ϵ_{ti}, τ_i and η_t are mutually independent. The vector of parameters to be estimated is $\underline{\theta} = (\underline{\beta}', \sigma_\tau, \rho_1, ..., \rho_k, \sigma_\eta)'$. Denote $\underline{\lambda} = (\lambda_1, ..., \lambda_T)'$ and $\underline{\tau} = (\tau_1, ..., \tau_N)'$.

The likelihood function associated with $\underline{y} = (y_{11}, ..., y_{TN})'$ can be written as

$$L(\underline{\theta}; \underline{y}) = \int g(\underline{\tau}, \underline{\lambda}; \underline{\theta}, \underline{y}) p(\underline{\tau}, \underline{\lambda}; \underline{\theta}) d\underline{\tau} d\underline{\lambda} \tag{2.8}$$

with

$$g(\underline{\tau}, \underline{\lambda}; \underline{\theta}, \underline{y}) = \prod_{i=1}^{N} \prod_{t=1}^{T} [\Phi(v_{it})]^{y_{it}} [1 - \Phi(v_{it})]^{1-y_{it}}$$

where

$$\Phi(v_{ti}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{v_{ti}} \exp\left(-\frac{1}{2} t^2\right) dt$$

$$v_{it} = \underline{\beta}' \underline{z}_{it} + \tau_i + \lambda_t$$

$$p(\underline{\tau}, \underline{\lambda}; \underline{\theta}) = \sigma_\tau^{-N} (2\pi)^{-N/2} \exp\left[-\frac{1}{2\sigma_\tau^2} \sum_{i=1}^{N} \tau_i^2\right] (2\pi)^{-T/2} |\Sigma_\lambda|^{-1/2} \exp\left[-\frac{1}{2} \underline{\lambda}' \Sigma_\lambda^{-1} \underline{\lambda}\right] \tag{2.9}$$

and Σ_λ denotes the stationary variance-covariance matrix of $\underline{\lambda}$.

2.3.1 EIS Evaluation of the Likelihood

We factorize the global high-dimensional Efficient Importance Sampling (EIS) optimization problem associated with (2.8) into a sequence of low-dimensional subproblems according to an appropriate factorization of the integrand $\phi(\underline{\tau}, \underline{\lambda}; \underline{\theta}, \underline{y}) = g(\underline{\tau}, \underline{\lambda}; \underline{\theta}, \underline{y}) p(\underline{\tau}, \underline{\lambda}; \underline{\theta})$. Thus (2.8) becomes

$$\begin{aligned} L(\underline{\theta}; \underline{y}) &= \int \left[\prod_{i=1}^{N} \prod_{t=1}^{T} [\Phi(v_{it})]^{y_{it}} [1 - \Phi(v_{it})]^{1-y_{it}} \right] \\ &\quad \times \sigma_\tau^{-N} (2\pi)^{-N/2} \exp\left[-\frac{1}{2\sigma_\tau^2} \sum_{i=1}^{N} \tau_i^2 \right] (2\pi)^{-T/2} |\Sigma_\lambda|^{-1/2} \exp\left[-\frac{1}{2} \underline{\lambda}' \Sigma_\lambda^{-1} \underline{\lambda} \right] d\underline{\tau} d\underline{\lambda} \\ &= \int (2\pi)^{-T/2} |\Sigma_\lambda|^{-1/2} \exp\left[-\frac{1}{2} \underline{\lambda}' \Sigma_\lambda^{-1} \underline{\lambda} \right] \sigma_\tau^{-N} (2\pi)^{-N/2} \\ &\quad \times \prod_{i=1}^{N} \left\{ \exp\left[-\frac{1}{2\sigma_\tau^2} \tau_i^2 \right] \prod_{t=1}^{T} [\Phi(v_{it})]^{y_{it}} [1 - \Phi(v_{it})]^{1-y_{it}} \right\} d\underline{\tau} d\underline{\lambda} \\ &= \int \phi_0(\underline{\lambda}; \underline{\theta}) \prod_{i=1}^{N} \phi_i(\tau_i, \underline{\lambda}; \underline{\theta}, \underline{y}) d\underline{\tau} d\underline{\lambda} \end{aligned} \tag{2.10}$$

where

$$\phi_0(\underline{\lambda}; \underline{\theta}) = (2\pi)^{-T/2} |\Sigma_\lambda|^{-1/2} \exp\left[-\frac{1}{2} \underline{\lambda}' \Sigma_\lambda^{-1} \underline{\lambda} \right] \sigma_\tau^{-N} (2\pi)^{-N/2}$$

$$\phi_i(\tau_i, \underline{\lambda}; \underline{\theta}, \underline{y}) = \exp\left[-\frac{1}{2\sigma_\tau^2} \tau_i^2 \right] \prod_{t=1}^{T} [\Phi(v_{it})]^{y_{it}} [1 - \Phi(v_{it})]^{1-y_{it}} \tag{2.11}$$

Since ϕ_i introduces interdependencies between τ_i and λ_t, the efficient sampler can be constructed as a sequence of sampling densities with an unconditional density $m_0(\underline{\lambda}; \underline{\alpha}_0)$ for $\underline{\lambda}$ and a sequence of conditional densities $m_i(\tau_i | \underline{\lambda}; \underline{\alpha}_i)$ for $\tau_i | \underline{\lambda}$. The resulting factorization is given by

$$m(\underline{\tau}, \underline{\lambda} | \underline{\alpha}) = m_0(\underline{\lambda}; \underline{\alpha}_0) \prod_{i=1}^{N} m_i(\tau_i | \underline{\lambda}; \underline{\alpha}_i)$$

For any given value of $\underline{\alpha}$, the likelihood (2.10) can be rewritten as

$$L(\underline{\theta}; \underline{y}) = \int \frac{\phi_0(\underline{\lambda}; \underline{\theta})}{m_0(\underline{\lambda}; \underline{\alpha}_0)} \prod_{i=1}^{N} \frac{\phi_i(\tau_i, \underline{\lambda}; \underline{\theta}, \underline{y})}{m_i(\tau_i | \underline{\lambda}; \underline{\alpha}_i)} m(\underline{\tau}, \underline{\lambda} | \underline{\alpha}) d\underline{\tau} d\underline{\lambda} \tag{2.12}$$

The corresponding EIS estimate is given by

$$\widetilde{L}_{S;m}(\underline{\theta};\underline{y}) = \frac{1}{S}\sum_{r=1}^{S}\frac{\phi_0(\widetilde{\underline{\lambda}}_r(\underline{\alpha}_0);\underline{\theta})}{m_0(\widetilde{\underline{\lambda}}_r(\underline{\alpha}_0);\underline{\alpha}_0)}\prod_{i=1}^{N}\frac{\phi_i(\widetilde{\tau}_{ir}(\underline{\alpha}_i),\widetilde{\underline{\lambda}}_r(\underline{\alpha}_0);\underline{\theta},\underline{y})}{m_i(\widetilde{\tau}_{ir}(\underline{\alpha}_i)|\widetilde{\underline{\lambda}}_r(\underline{\alpha}_0);\underline{\alpha}_i)} \tag{2.13}$$

where $\left\{\left[\widetilde{\tau}_{1r}(\underline{\alpha}_1),...,\widetilde{\tau}_{Nr}(\underline{\alpha}_N),\widetilde{\underline{\lambda}}_r(\underline{\alpha}_0)\right];\ r=1,...,S\right\}$ are iid draws from the auxiliary importance sampling density $m(\underline{\tau},\underline{\lambda}|\underline{\alpha})$.

A density kernel $k_i(\tau_i;\underline{\lambda};\underline{\alpha}_i)$ for $m_i(\tau_i|\underline{\lambda};\underline{\alpha}_i)$ is given by

$$m_i(\tau_i|\underline{\lambda};\underline{\alpha}_i) = \frac{k_i(\tau_i;\underline{\lambda};\underline{\alpha}_i)}{\chi_i(\underline{\lambda};\underline{\alpha}_i)}$$

with

$$\chi_i(\underline{\lambda};\underline{\alpha}_i) = \int k_i(\tau_i;\underline{\lambda},\underline{\alpha}_i)d\tau_i$$

The likelihood (2.12) can now be rewritten as

$$L(\underline{\theta};\underline{y}) = \int\frac{\phi_0(\underline{\lambda};\underline{\theta})\prod_{i=1}^{N}\chi_i(\underline{\lambda};\underline{\alpha}_i)}{m_0(\underline{\lambda};\underline{\alpha}_0)}\prod_{i=1}^{N}\frac{\phi_i(\tau_i,\underline{\lambda};\underline{\theta},\underline{y})}{k_i(\tau_i;\underline{\lambda};\underline{\alpha}_i)}m(\underline{\tau},\underline{\lambda}|\underline{\alpha})d\underline{\tau}d\underline{\lambda}$$

where γ_i is a proportionality constant.

The EIS optimization problem requires solving a sequence of $N+1$ weighted LS problems of the form

$$\widehat{\underline{\alpha}}_i = \arg\min_{\underline{\alpha}_i}\sum_{r=1}^{S}\left\{\ln\phi_i(\widetilde{\tau}_i,\widetilde{\underline{\lambda}};\underline{\theta},\underline{y}) - q_i - \ln k_i(\widetilde{\tau}_i,\widetilde{\underline{\lambda}};\underline{\alpha}_i)\right\}^2 g_i(\widetilde{\tau}_i,\widetilde{\underline{\lambda}};\underline{\theta},\underline{y}) \tag{2.14}$$

for $i=1,...,N$ and

$$\widehat{\underline{\alpha}}_0\arg\min_{\underline{\alpha}_0}\sum_{r=1}^{S}\left\{\ln\left[\phi_0(\widetilde{\underline{\lambda}}_r;\underline{\theta})\prod_{i=1}^{N}\chi_i(\widetilde{\underline{\lambda}}_r;\widehat{\underline{\alpha}}_i) - q_0 - \ln m_0(\widetilde{\underline{\lambda}}_r;\underline{\alpha}_0)\right]\right\}^2$$

where $\widetilde{\underline{\tau}},\widetilde{\underline{\lambda}}$ are draws from $m(\underline{\tau},\underline{\lambda}|\underline{\alpha})$. Based on these draws, the EIS estimate of the likelihood (2.13) is calculated as

$$\widetilde{L}_{r,m}(\underline{\theta};\underline{y}) = \frac{\phi_0(\widetilde{\underline{\lambda}}_r;\underline{\theta})\prod_{i=1}^{N}\chi_i\left(\widetilde{\underline{\lambda}}_r,\underline{\alpha}_i\right)}{m_0(\widetilde{\underline{\lambda}}_r;\underline{\alpha}_0)}\prod_{i=1}^{N}\frac{\phi_i(\widetilde{\tau}_{ir},\widetilde{\underline{\lambda}}_r;\underline{\theta},\underline{y})}{k_i(\widetilde{\tau}_{ir}|\widetilde{\underline{\lambda}}_r;\underline{\alpha}_i)} \tag{2.15}$$

For further details on implementation, see Appendix 2.1.

2.3.2 Bayesian MCMC Approach Based on EIS

Bayesian MCMC simulation methods such as Gibbs sampling rely upon sampling from conditional posterior distributions in order to construct a Markov chain whose equilibrium distribution is the joint posterior of the parameters given the data. For the panel probit model, the joint posterior distribution of parameters can be augmented with the vectors of latent variables $\underline{\tau}$ and $\underline{\lambda}$. The complete joint posterior $f(\underline{\theta}, \underline{\tau}, \underline{\lambda}|Z)$ can then be drawn from using Gibbs sampling. The main difficulty with such an MCMC approach is that of efficiently sampling from τ_i and $\underline{\lambda}$ since the corresponding multivariate posterior distributions are high-dimensional and have no closed-form solution. To overcome this problem, Liesenfeld and Richard (2006a) proposed combining the EIS sampler with the Acceptance-Rejection Metropolis-Hastings (AR-MH) algorithm of Tierney (1994) in simulating the autocorrelated error component in stochastic volatility models. We adopt the approach to the panel probit model by simulating $\tau_i|\underline{\theta}, Z$ and $\underline{\lambda}|\underline{\theta}, Z$ as Gibbs blocks: We sample the unobserved individual heterogeneity component $\tau_i|\underline{\theta}, Z$ as one Gibbs block drawing from a piece-wise linear approximation to the marginal posterior density constructed with a nonparametric form of EIS. The time effects $\underline{\lambda}|\underline{\theta}, Z$ are simulated as another Gibbs block with a parametric EIS proposal density for an AR-MH step. The basis of this procedure is that the EIS densities for $\tau_i|\underline{\theta}, Z$ and $\underline{\lambda}|\underline{\theta}, Z$ provide very close approximations to $f(\tau_i|\underline{\theta}, Z)$ and $f(\underline{\lambda}|\underline{\theta}, Z)$, respectively. The piece-wise linear approximation to $f(\tau_i|\underline{\theta}, Z)$ is dominated by $f(\tau_i|\underline{\theta}, Z)$ everywhere and can be made arbitrarily precise by increasing the size of the simulated grid. For $f(\underline{\lambda}|\underline{\theta}, Z)$ given the model assumptions, one can expect that the EIS parametric density provides an efficient proposal density for the target posterior $f(\underline{\lambda}|\underline{\theta}, Z)$ in the AR-MH step. This conjecture has been validated by AR-MH acceptance rates close to 100% in our empirical application.

Liesenfeld and Richard (2006a) list three attractive features that hold for the EIS-AR-MH approach in general: 1) only minor modifications of the code for the classical SML analysis are necessary in order to obtain a corresponding code for the Bayesian analysis (and vice versa), 2) it allows for a direct comparison between Bayesian and classical estimation results and for a corresponding analysis of the impact of the prior density on the inference

process, and 3) its basic structure does not depend upon a specific model.

For a given vector of parameters $(\underline{\theta}, \Upsilon)$ the augmented likelihood $L(\underline{\theta}, \Upsilon; Z)$ is defined in (2.8). Let $\underline{\theta}$ without the subvector θ_j be denoted by $\underline{\theta}_{/\theta_j}$. For each Gibbs block of a generic parameter θ_j the Bayesian optimal updating of prior beliefs, $\pi(\theta_j)$, with new information (data Z) takes the form

$$f(\theta_j|\underline{\theta}_{/\theta_j}, \Upsilon, Z) \propto L(\underline{\theta}, \Upsilon; Z)\pi(\theta_j) \tag{2.16}$$

The individual Gibbs blocks used are $\underline{\beta}$, σ_τ, σ_η, $\underline{\rho}$, $\underline{\lambda}$, and $\underline{\tau}$, given data and the remaining augmented parameters. Throughout the analysis we make use of non-informative priors. Details of sampling from the posterior distributions are described in Appendix 2.2.

2.4 EMPIRICAL RESULTS

In this section, we first reproduce the pooled probit estimates and the results obtained by Bertschek and Lechner (1998) and Greene (2004) as a benchmark for comparison with our results. Although these authors also report estimates of models other than shown below, we only select the ones with the least restrictive assumptions on the underlying probit models.

Table 2.1 presents the basic case of Pooled Estimator of *Model 1* in (2.1) estimated in Stata using the command 'probit'. Table 2.1 also reports the Bertschek and Lechner (1998) GMM parameter estimates of *Model 2* with a k-NN estimate of Ω in (2.4) and the Greene (2004) random parameter model prior means estimates of *Model 3*. As discussed in Greene (2004), there are some substantial differences compared to the other two models. Especially noteworthy are the greater impacts of the two central parameters of imports and FDI share on innovations as implied by the random parameters model. Nonetheless, these effects are positive in all cases as predicted.

Table 2.2 lists the Greene (2004) latent class estimates of *Model 4*. According to Greene (2004), working down from the number of classes $J = 5$ the estimates stabilized at the reported $J = 3$. Despite a large amount of variation across the three classes, the original conclusion that FDI and imports positively affect the probability of product innovation continued to be supported.

Table 2.3 presents our classical EIS-SML estimates and Bayesian posterior means of parameters in the model (2.7) with unobserved heterogeneity and autocorrelated errors. Posterior marginal densities of the Bayesian analysis are given in Figure 2.1 and autocorrelation functions of the parameter draws in Figure 2.3.

We excluded from estimation three distant outliers with relative firm size larger than 0.1 and productivity larger than 0.8 (see Figure 2.2) as these observations were inducing numerical instabilities into our EIS-SML estimator. Thus our sample size was $N = 1267$ and $T = 5$. The EIS-SML asymptotic (statistical) standard errors were obtained as the square root of the diagonal of the negative of the inverse of the Hessian evaluated at the estimated parameter values. The EIS-SML estimates are all within one standard deviation from the EIS-MCMC posterior means. One exception is the unobserved heterogeneity parameter σ_τ. Its EIS-MCMC value $\widehat{\sigma}_\tau = 1.021$ lies close to the value 1.1707 of an analogous parameter reported by Greene (2004, p.35) for the random effects model, but its EIS-SML value is about half that size. This can be explained by a potentially high skewness of its sampling density for companies whose response variable was constant (1 or 0) throughout the sample period. Both estimates of $\widehat{\sigma}_\eta$ indicate that the role of time effects in this dataset is very small relative to individual unobserved effects. Large standard errors on $\widehat{\rho}$ and its posterior distribution imply that this parameter could not be empirically identified, which further confirms the small significance of the time effects.

Most of our coefficient estimates fit into a convex combination of Greene (2004)'s Class 1 – Class 3. The estimates of the two key parameters of *FDI* and *import share* are positive, further validating the original hypothesis. However, both our estimates of the *FDI* coefficient are smaller relative to previous results. The *import share* coefficient estimates are also very close to the lower bound of Greene (2004)'s Class1 – Class 3 estimates. We attribute this finding to our flexible model assumptions whereby the influence the unobserved effects on *product innovation* was previously unaccounted for and channeled through *FDI* and *import share* in the model. Also, the exclusion of far outliers on this variables from our estimates may have played a role in this respect. The three excluded observations with large relative size have also disproportionately large values of *import share* and *FDI*; the means of the three outliers are 0.402 and 0.208 contrasting with means of the rest of the sample of 0.252

Table 2.1: Panel Probit - Models 1-3

	Pooled Probit[a]		Model 1[b]		Model 2[c]	
Variable	Estimate	Std.Err.	Estimate	Std.Err.	Estimate	Std.Err.
Constant	−1.960**	0.230	−1.74**	0.37	−3.134	0.191
log sales	0.177**	0.022	0.15**	0.034	0.306	–
Rel size	1.072**	0.142	0.95**	0.20	3.735	0.184
Imports	1.133**	0.151	1.14**	0.24	1.582	0.126
FDI	2.853**	0.402	2.59**	0.59	3.111	0.320
Prod.	−2.341**	0.715	−1.91**	0.82	−5.786	0.755
Raw Mtl	−0.279**	0.081	−0.28**	0.12	−0.346	0.077
Inv good	0.188**	0.039	0.21**	0.063	0.238	0.453

[a] Estimated in Stata by the simple command 'probit'.

[b] Bertschek and Lechner (1998), WNP-joint uniform estimates with $k = 880$, Table 9, standard errors from Table 10. Reprinted from *Journal of Econometrics*, Vol. 87(2), Bertschek, I. and M. Lechner, "Convenient estimators for the panel probit model," 329-371, Copyright (1998), with permission from Elsevier.

[c] Greene (2004), $\hat{\mu}$ in Table 5. Reprinted from *Empirical Economics*, Vol. 29(1), Greene, W., "Convenient estimators for the panel probit model: Further results," 21-47, Copyright (2004), with kind permission of Springer Science and Business Media.

* Indicates significant at the 95% level

** Indicates significant at the 99% level

Table 2.2: Panel Probit - Model 4

	Class 1		Class 2		Class 3	
Variable	Estimate	Std.Err.	Estimate	Std.Err.	Estimate	Std.Err.
Constant	-2.32^{**}	0.768	-2.71^{**}	0.766	-8.97^{**}	2.50
log sales	0.323^{**}	0.075	0.233^{**}	0.0675	0.571^{**}	0.197
Rel size	4.38^{**}	0.882	0.720^{**}	0.253	1.42^{*}	0.616
Imports	0.936^{**}	0.491	2.26^{**}	0.503	3.12^{*}	1.35
FDI	2.20	2.54	2.80^{**}	0.926	8.37^{**}	2.27
Prod.	-5.86^{**}	1.69	-7.70^{**}	1.16	-0.910	1.26
Raw Mtl	-0.110	0.172	-0.599^{**}	0.295	-0.856^{*}	0.424
Inv good	0.131	0.143	0.413^{**}	0.132	0.469^{*}	0.225

Greene (2004), Table 7. Reprinted from *Empirical Economics*, Vol. 29(1), Greene, W., "Convenient estimators for the panel probit model: Further results," 21-47, Copyright (2004), with kind permission of Springer Science and Business Media.

* Indicates significant at the 95% level

** Indicates significant at the 99% level

and 0.045 for *import share* and *FDI*, respectively. The outliers' means thus correspond to approximately to the 82^{nd} percentile and 98^{th} percentile, respectively, of the remaining observations of these variables.

2.5 CONCLUSION

In this paper, we performed classical simulated maximum likelihood (SML) and Bayesian analysis of a panel probit model with unobserved individual heterogeneity and autocorrelated errors. The SML analysis was facilitated with the Efficient Importance Sampling (EIS) method that was found competitive with the GHK simulator in previous studies and was newly adopted to the panel probit case in this paper. In the Bayesian part, we embedded EIS within an Markov Chain Monte Carlo (MCMC) simulation method to perform posterior analysis augmented with both the time and cross-section latent variables. Thus, the posterior for the unobserved individual heterogeneity was sampled from as one Gibbs block, using a nonparametric version of EIS to form a piece-wise linear approximation to the posterior as a proposal density. The posterior for the vector of latent time effects was treated as another Gibbs block, using a parametric EIS approximation as the proposal density for an AR-MH step. This approach represents a methodological contribution to the limited dependent variable panel literature.

We applied our method to the product innovation activity of a panel of German manufacturing firms in response to imports, foreign direct investment and other control variables. Our findings confirm the positive effect of imports and FDI on firms' innovation activity found in previous literature. However, our coefficient estimates of these variables were smaller than the ones reported by Bertschek and Lechner (1998) and Greene (2004) who analyzed the same dataset under more restrictive model assumptions. This discrepancy can be explained by the exclusion of three far outliers from our estimation and also by our weak model assumptions relative to these authors.

The work presented in this paper can be extended in several directions. First, the parametric EIS used in the classical evaluation of the likelihood function can be replaced by

Table 2.3: Panel Probit - EIS-SML and EIS-MCMC

	EIS-SML[e]			EIS-MCMC[f]	
Variable	Estimate	Std.Err.	MC Err.	Posterior mean	Std.Err.
Constant	-1.612^{**}	0.215	0.054	-1.427^{**}	0.347
log sales	0.155^{**}	0.022	0.006	0.137^{**}	0.035
Rel size	0.613^{**}	0.134	0.030	0.795^{**}	0.197
Imports	0.947^{**}	0.176	0.061	0.753^{**}	0.231
FDI	2.057^{**}	0.465	0.128	2.010^{**}	0.577
Prod.	-3.035^{**}	1.592	0.460	-2.787	2.015
Raw Mtl	-0.108	0.308	0.018	-0.108	0.166
Inv good	0.141^{**}	0.046	0.015	0.147^{**}	0.059
σ_τ	0.471^{**}	0.015	0.001	1.021^{**}	0.030
σ_η	0.036^{*}	0.010	0.012	0.041^{g}	–
ρ	0.002	0.567	0.001	0.002	0.571

[e] EIS-SML estimates are the averages of 10 estimation rounds starting with different CRNs. Each round is based on an MC sample size $S = 600$. An average of 6-7 EIS iterations were needed for full parameter convergence. A grid search optimization procedure in Fortran 90 took approximately 9 hours, with relative function tolerance of 10^{-4} on a 1.7 GHz opteron unix machine.

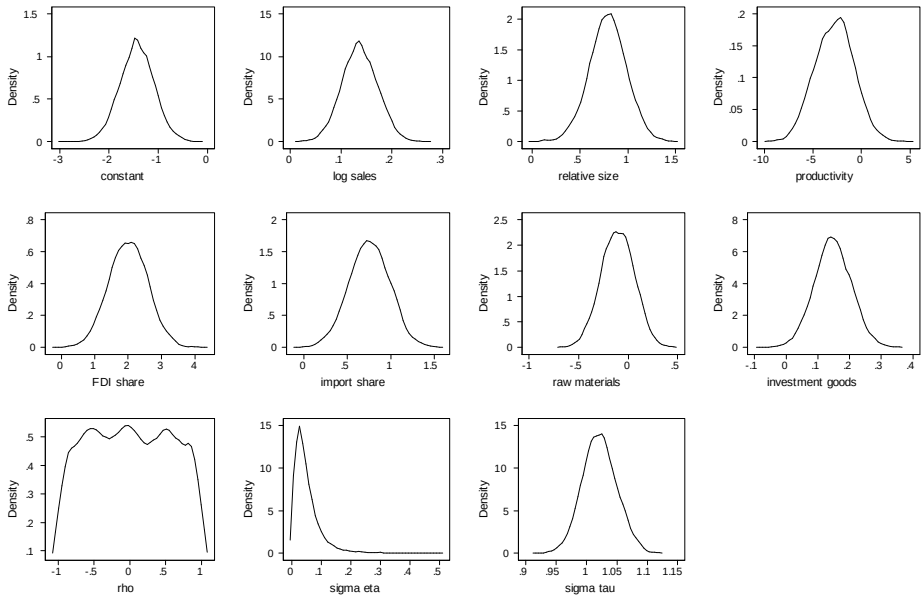
[f] Posterior moments are based on $12,000$ Gibbs iterations (discarding the first $2,000$ draws). One Gibbs iteration took approximately 28 seconds on a 1.7 GHz opteron unix machine. The EIS simulation smoother is based on an MC sample of 400. On average, it took less than 6 EIS iterations for full convergence of the EIS parameters in sampling from the posteriors of the latent variables μ and λ. The AR and MH acceptance rates for λ were 99.00% and 99.85%, respectively.

[g] Due to the skewness of the marginal posterior distribution (see Figure 1), the median is reported. The mean is 0.07842, interquartile range $[0.022, 0.072]$, and the 95% confidence interval is $[0.006, 0.255]$.

[*] Indicates significant at the 95% level

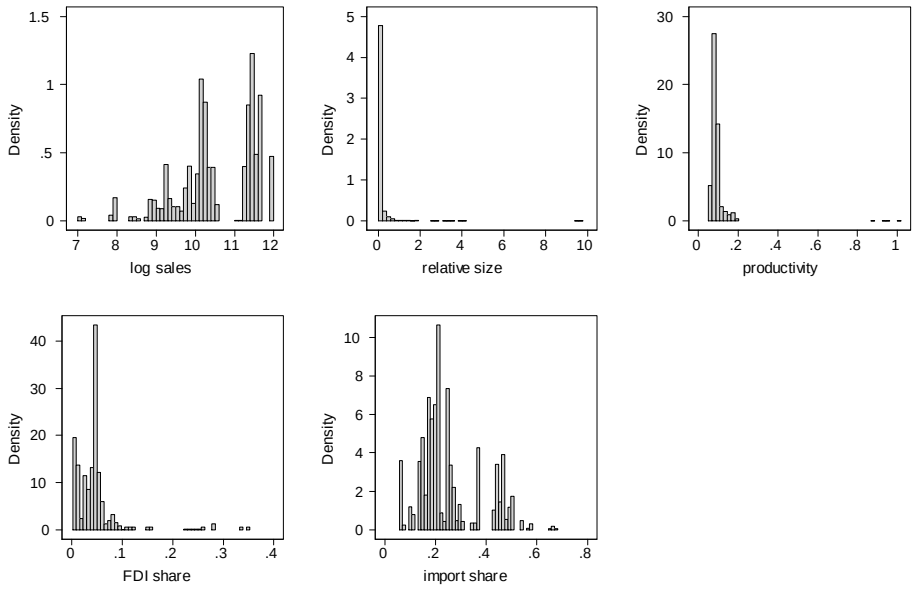
[**] Indicates significant at the 99% level

Figure 2.1: Marginal Posterior Densities



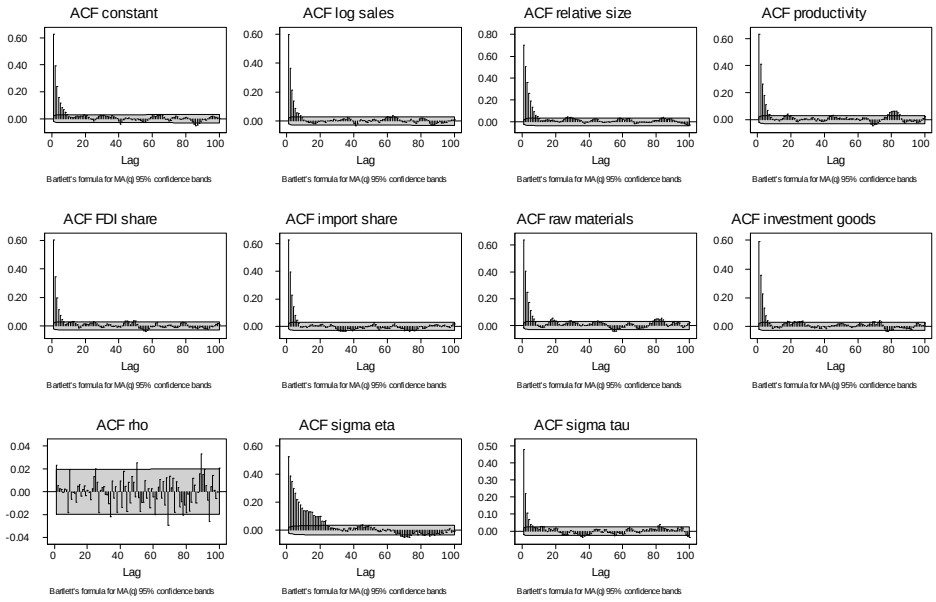
the nonparametric EIS version used in sampling from the posterior of τ_i. Implementation of the nonparametric EIS for approximating the density kernels of the unobserved firm heterogeneity component is currently subject to our research. We anticipate further efficiency improvements in the SML evaluation relative to the present parametric EIS. Second, despite the theoretical appeal of EIS, the current Monte Carlo evidence comparing its performance to other samplers is rather sparse. An MC study comparing both the parametric and nonparametric EIS to, for example, GHK in the SML, MSM, and MSS environments would undoubtedly be of interest to applied researchers using simulation estimators. Furthermore, EIS as a procedure for fast and accurate numerical evaluation of multivariate integrals can be imbedded in more complicated structural models beyond its current reduced form use.

Figure 2.2: Descriptive Histograms for the Data



In line with Bertschek and Lechner (1998) and Greene (2004) we have normalized the relative size by the factor of 30.

Figure 2.3: Autocorrelation Functions of the Parameters



The autocorrelation functions are based on 12,000 parameter draws.

APPENDIX

Appendix 2.1: EIS Likelihood Evaluation

We consider the density kernel k_i for $\tau_i|\underline{\lambda}$ as given by

$$k_i(\tau_i;\underline{\lambda},\underline{\alpha}_i) = \exp\left\{-\frac{1}{2}\left(\underline{b}_i'\underline{v}_i + \underline{v}_i'C_i\underline{v}_i\right) - \frac{\tau_i^2}{2\sigma_\tau^2}\right\} \tag{2.17}$$

where

$$\begin{aligned}
\underline{b}_i &= (b_{1i},...,b_{Ti})' \\
C_i &= diag(\underline{c}_i) \\
\underline{c}_i &= (c_{1i},...,c_{Ti})' \\
\underline{v}_i &= \underline{\lambda} + \tau_i\underline{\iota} + Z_i\underline{\beta} \\
\underline{\iota} &= (1,...,1)' \\
Z_i &= (\underline{z}_{1i},...\underline{z}_{Ti})'
\end{aligned}$$

and the auxiliary parameters are $\underline{\alpha}_i = (b_i', c_i')'$.

Let $\underline{l}_i = \underline{\lambda} + Z_i\underline{\beta}$ which implies $\underline{v}_i = \underline{l}_i + \tau_i\underline{\iota}$. Then

$$k_i(\tau_i;\underline{\lambda},\underline{\alpha}_i) = \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{\sigma_\tau^2} + \underline{\iota}'C_i\underline{\iota}\right)\tau_i^2 + \left(\underline{b}_i'\underline{\iota} + 2\underline{\iota}'C_i\underline{l}_i\right)\tau_i + \underline{b}_i'\underline{l}_i + \underline{l}_i'C_i\underline{l}_i\right]\right\} \tag{2.18}$$

Matching (2.18) with a Gaussian kernel we obtain the conditional mean of $\tau_i|\underline{\lambda}$ as

$$\mu_i\left(\underline{\alpha}_i\right) = -\sigma_i^2\left(\frac{1}{2}\underline{b}_i'\underline{\iota} + \underline{\iota}'C_i\underline{l}_i\right) \tag{2.19}$$

and variance of $\tau_i|\underline{\lambda}$ as

$$\begin{aligned}
\sigma_i^2\left(\underline{\alpha}_i\right) &= \left(\frac{1}{\sigma_\tau^2} + \underline{\iota}'C_i\underline{\iota}\right)^{-1} \\
&= \frac{\sigma_\tau^2}{(1-\sigma_\tau^2\underline{\iota}'C_i\underline{\iota})}
\end{aligned} \tag{2.20}$$

In what follows we will suppress dependence of μ_i and σ_i^2 on $\underline{\alpha}_i$ for notational convenience. Integrating k_i with respect to τ_i leads to the following form of the integrating constant

$$\chi_i\left(\underline{\lambda},\underline{\alpha}_i\right) = \sqrt{2\pi}\sigma_i\exp\left\{-\frac{1}{2}\left(\underline{b}_i'\underline{l}_i + \underline{l}_i'C_i\underline{l}_i\right) + \frac{1}{2}\frac{\mu_i^2}{\sigma_i^2}\right\} \tag{2.21}$$

which itself is a Gaussian density kernel for $\underline{\lambda}$.

Let

$$m_i(\tau_i|\underline{\lambda};\underline{\alpha}_i) = \frac{k_i(\tau_i;\underline{\lambda};\underline{\alpha}_i)}{\chi_i(\underline{\lambda};\underline{\alpha}_i)}$$

The EIS regression (without weights) introduced in (2.14) is derived for each i from (2.11) and (2.17) as

$$\begin{aligned}\ln \phi_i(\widetilde{\tau}_i, \widetilde{\underline{\lambda}}; \underline{\theta}, \underline{y}) &= q_i + \ln k_i(\widetilde{\tau}_i; \widetilde{\underline{\lambda}}, \underline{\alpha}_i) + \xi_{ir} \\ -\frac{\widetilde{\tau}_i^2}{2\sigma_\tau^2} + \sum_{t=1}^{T} &\left[(1-y_{ti}) \ln\left[1-\Phi(\widetilde{v}_{tir})\right] + y_{ti} \ln \Phi(\widetilde{v}_{tir})\right] \\ &= q_i - \frac{1}{2}\left(\underline{b}_i' \underline{\widetilde{v}}_{ir} + \underline{\widetilde{v}}_i' C_i \underline{\widetilde{v}}_{ir}\right) - \frac{\widetilde{\tau}_i^2}{2\sigma_\tau^2} + \xi_{ir}\end{aligned}$$

$$\begin{aligned}\sum_{t=1}^{T} \left[(1-y_{ti}) \ln\left[1-\Phi(\widetilde{v}_{tir})\right] + y_{ti} \ln \Phi(\widetilde{v}_{tir})\right] &= q_i - \frac{1}{2}\left(\underline{b}_i' \underline{\widetilde{v}}_{ir} + \underline{\widetilde{v}}_{ir}' C_i \underline{\widetilde{v}}_{ir}\right) + \xi_{ir} \\ &= q_i + (-b_{1i}/2)\, \widetilde{v}_{1ir} + ... + (-b_{Ti}/2)\, \widetilde{v}_{Tir} \\ &+ (-c_{1i}/2)\, \widetilde{v}_{1ir}^2 + ... + (-c_{Ti}/2)\, \widetilde{v}_{Tir}^2 + \xi_{ir}\end{aligned} \tag{2.22}$$

with weights

$$g_i(\widetilde{\tau}_{ir}, \widetilde{\underline{\lambda}}; \underline{\theta}, \underline{y}) = \exp\left[-\frac{\widetilde{\tau}_i^2}{2\sigma_\tau^2}\right] \prod_{t=1}^{T} \left[\Phi(\widetilde{v}_{itr})\right]^{y_{it}} \left[1-\Phi(\widetilde{v}_{itr})\right]^{1-y_{it}}$$

where ξ_{ir} denotes the regression error term and $\{\widetilde{v}_{tir} : r = 1, ..., S\}$ are the simulated draws v_{ti}.

Using (2.21), the function to be approximated by the Gaussian sampler m_0 is given by

$$\begin{aligned}\phi_0(\underline{\lambda}; \underline{\theta}) \prod_{i=1}^{N} \chi_i\left(\underline{\lambda}, \underline{\alpha}_i\right) &= (2\pi)^{-T/2} \left|\Sigma_\lambda\right|^{-1/2} \exp\left[-\frac{1}{2}\underline{\lambda}' \Sigma_\lambda^{-1} \underline{\lambda}\right] \sigma_\tau^{-N} (2\pi)^{-N/2} \\ &\times \prod_{i=1}^{N} \sqrt{2\pi}\sigma_i \exp\left\{-\frac{1}{2}\left[\left(\underline{b}_i' \underline{l}_i + \underline{l}_i' C_i \underline{l}_i\right) - \frac{\mu_i^2}{\sigma_i^2}\right]\right\}\end{aligned} \tag{2.23}$$

Consider for the moment the very last term $\frac{\mu_i^2}{\sigma_i^2}$ of (2.23)

$$\frac{\mu_i^2}{\sigma_i^2} = 2\sigma_i^2 \underline{\lambda}' \underline{c}_i \left(\underline{c}_i' Z_i \underline{\beta} + \frac{1}{2}\underline{\iota}' \underline{b}_i\right) + \sigma_i^2 \underline{\lambda}' \underline{c}_i \underline{c}_i' \underline{\lambda} + \sigma_i^2 \left[\left(\underline{c}_i' Z_i \underline{\beta}\right)^2 + \left(\frac{1}{2}\underline{b}_i' \underline{\iota}\right)^2 + \underline{b}_i' \underline{\iota} \underline{c}_i' Z_i \underline{\beta}\right] \tag{2.24}$$

Substituting (2.24) into (2.23) yields

$$\begin{aligned}\phi_0(\underline{\lambda}; \underline{\theta}) \prod_{i=1}^{N} \chi_i\left(\underline{\lambda}, \underline{\alpha}_i\right) &= (2\pi)^{-T/2} \left|\Sigma_\lambda\right|^{-1/2} \sigma_\tau^{-N} (2\pi)^{-N/2} (2\pi)^{1/2} \left[\prod_{i=1}^{N} \sigma_i\right] \\ &\times \exp\left\{-\frac{1}{2}\left[\underline{\lambda}' \Sigma_\lambda^{-1} \underline{\lambda} + \sum_{i=1}^{N}\left(\underline{b}_i' \underline{l}_i + \underline{l}_i' C_i \underline{l}_i - \frac{\mu_i^2}{\sigma_i^2}\right)\right]\right\} \\ = \psi \exp&\left\{-\frac{1}{2}\left(\underline{\lambda}' \sum_{i=1}^{N}\left(-\underline{a}_i + \underline{b}_i + 2C_i Z_i \underline{\beta}\right) + \underline{\lambda}'\left[\Sigma_\lambda^{-1} + \sum_{i=1}^{N}\left[C_i - B_i\right]\right]\underline{\lambda} + \underline{r}\right)\right\}\end{aligned} \tag{2.25}$$

where

$$\begin{aligned}\psi &= (2\pi)^{-(T+N-1)/2}\left|\Sigma_\lambda\right|^{-1/2}\sigma_\tau^{-N}\prod_{i=1}^{N}\sigma_i \\ \underline{r} &= \sum_{i=1}^{N}\left[\underline{b}_i'Z_i\underline{\beta}+\underline{\beta}'Z_i'C_iZ_i\underline{\beta}-\sigma_i^2\left[\left(\underline{c}_i'Z_i\underline{\beta}\right)^2+\left(\frac{1}{2}\underline{b}_i'\underline{\iota}\right)^2+\underline{b}_i'\underline{\iota}\underline{c}_i'Z_i\underline{\beta}\right]\right]\end{aligned}$$

Matching a multivariate Gaussian kernel with (2.25) yields the variance-covariance matrix of $\underline{\lambda}$ on m_0

$$\Sigma_0\left(\underline{\alpha}_0\right)=\left[\Sigma_\lambda^{-1}+\sum_{i=1}^{N}\left[C_i-\sigma_i^2\underline{c}_i\underline{c}_i'\right]\right]^{-1} \tag{2.26}$$

and the mean of $\underline{\lambda}$ on m_0

$$\underline{\mu}_0\left(\underline{\alpha}_0\right)=\Sigma_0\sum_{i=1}^{N}\left(\sigma_i^2\underline{c}_i\left(\underline{c}_i'Z_i\underline{\beta}+\frac{1}{2}\underline{\iota}'\underline{b}_i\right)-\frac{1}{2}\underline{b}_i-C_iZ_i\underline{\beta}\right) \tag{2.27}$$

Matching the last term of (2.25) with a multivariate Gaussian kernel we obtain the integrating constant of $\underline{\lambda}$ on m_0

$$\chi_0=(2\pi)^{T/2}\left|\Sigma_0\left(\underline{\alpha}_0\right)\right|^{1/2}\psi\exp\left\{-\frac{1}{2}\left(\underline{r}-\underline{\mu}_0'\Sigma_0^{-1}\underline{\mu}_0\right)\right\}$$

The EIS estimate of the likelihood (2.13) is calculated from (2.15) as

$$\begin{aligned}\widetilde{L}_{r,m}(\underline{\theta};\underline{y}) &= (2\pi)^{-(N-1)/2}\left|\Sigma_0\left(\underline{\alpha}_0\right)\right|^{1/2}\left|\Sigma_\lambda\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\underline{r}-\underline{\mu}_0'\Sigma_0^{-1}\underline{\mu}_0\right)\right\} \\ &\quad\times\sigma_\tau^{-N}\left[\prod_{i=1}^{N}\sigma_i\right]\prod_{i=1}^{N}\frac{\phi_i(\widetilde{\tau}_{ir},\widetilde{\underline{\lambda}}_r;\underline{\theta},\underline{y})}{k_i(\widetilde{\tau}_{ir}|\widetilde{\underline{\lambda}}_r;\underline{\alpha}_i)}\end{aligned}$$

and the log-likelihood as

$$\begin{aligned}\ln\widetilde{L}_{S;m}(\underline{\theta};\underline{y}) &= \frac{1}{S}\sum_{r=1}^{S}\ln\widetilde{L}_{r,m}(\underline{\theta};\underline{y}) \\ &= -\frac{N-1}{2}\ln(2\pi)+\frac{1}{2}\left(\ln\left|\left(\Sigma_0\left(\underline{\alpha}_0\right)\right|-\ln\left|\Sigma_\lambda\right|\right)-\frac{1}{2}\left(\underline{r}-\underline{\mu}_0'\Sigma_0^{-1}\underline{\mu}_0\right)-N\ln\sigma_\tau \\ &\quad+\sum_{i=1}^{N}\ln\sigma_i+\ln\left[\frac{1}{S}\sum_{r=1}^{S}\exp\left\{\sum_{i=1}^{N}\left(\ln\phi_i(\widetilde{\tau}_{ir},\widetilde{\underline{\lambda}}_r;\underline{\theta},\underline{y})-\ln k_i(\widetilde{\tau}_{ir}|\widetilde{\underline{\lambda}}_r;\underline{\alpha}_i)\right)\right\}\right]\end{aligned} \tag{2.28}$$

Algorithm

Based on these derivations, the computation of an efficient MC estimate of the likelihood for the panel probit model requires the following steps:

Step (1): Use the natural sampling density p in (2.9) to draw S independent realizations of the latent process $(\widetilde{\underline{\tau}}_r,\widetilde{\underline{\lambda}}_r)$.

Step (2): Use these random draws to solve the sequence of N weighted (unweighted for the first iteration of the importance sampling construction) LS problems defined in equation (2.22).

Step (3): Use the sampling density from m_0 with moments given in (2.26) and (2.27) to draw S trajectories $\left\{\widetilde{\underline{\lambda}}_r(\widehat{\alpha}_0) : r = 1, ..., S\right\}$. Conditional on these trajectories, draw from the conditional densities $\{m_i\}$ characterized by the moments (2.19) and (2.20) the vectors $\{\widetilde{\underline{\tau}}_r(\widehat{\underline{\alpha}}_1, ..., \widehat{\underline{\alpha}}_N); r = 1, ..., S\}$. Throughout the text, these draws are denoted by a shorthand notation $\widetilde{\underline{\tau}}_r$ and $\widetilde{\underline{\lambda}}_r$.

Step (4): Maximize the simulated log-likelihood (2.28), evaluated at $\widetilde{\underline{\tau}}_r$ and $\widetilde{\underline{\lambda}}_r$ in each step, with respect to the parameters $\underline{\theta}$.

Appendix 2.2: Sampling from Posterior Densities

Sampling from $f(\underline{\beta}|\underline{\theta}_{/\underline{\beta}}, \Upsilon, Z)$

Here we adopt the methodology elaborated in (Albert and Chib, 1993). In our panel application,

$$\begin{aligned} Y_i^* &= Z_i\underline{\beta} + \underline{\lambda} + \tau_i\underline{\iota} + \underline{\varepsilon}_i \\ Y_{/\Upsilon,i}^* &= Y_i^* - \underline{\lambda} - \tau_i\underline{\iota} + \underline{\varepsilon}_i \\ Y_{/\Upsilon,i}^* &= Z_i\underline{\beta} + \underline{\varepsilon}_i \end{aligned}$$

Assigning a noninformative prior $\pi(\underline{\beta})$ to $\underline{\beta}$ results in

$$f(\underline{\beta}|\underline{\theta}_{/\underline{\beta}}, \Upsilon, Z) = N(\widehat{\underline{\beta}}, \widehat{\Sigma}_{\underline{\beta}}) \tag{2.29}$$

where $\widehat{\underline{\beta}} = (Z'Z)^{-1} Z'Y_{/\Upsilon}^*$, the dependent variable is a $(NT \times k)$ matrix $Y_{/\Upsilon}^* = (Y_{/\Upsilon,1}^{*\prime}, ..., Y_{/\Upsilon,N}^{*\prime})'$ and $\widehat{\Sigma}_{\underline{\beta}} = (Z'Z)^{-1}$. The random variables Y_{it}^* are independent with

$$\begin{aligned} f(Y_{it}^*|\underline{\theta}, \Upsilon, Z) &= N(\mu_{it}^*, 1) \\ \mu_{it}^* &= Z_{it}\underline{\beta} + \lambda_t + \tau_i \end{aligned} \tag{2.30}$$

truncated at the left by 0 if $Y_{it} = 1$ and truncated at the right by 0 if $Y_{it} = 0$. Given a previous value of $\underline{\beta}$, τ_i and λ_t, one cycle the Gibbs algorithm would produce Y_{it}^* and $\underline{\beta}$ from the distributions (2.30) and (2.29); see Train (2003, p. 210) for simulation algorithm. The starting value $\underline{\beta}^{(0)}$ may be taken to be the ML estimate.

Sampling from $f(\tau_i|\underline{\lambda}, \underline{\theta}, Z)$

From (2.10),

$$f(\tau_i|\underline{\lambda}, \underline{\theta}, Z) \propto \sigma_\tau^{-1} \exp\left[-\frac{1}{2\sigma_\tau^2}\tau_i^2\right] \prod_{t=1}^{T} [\Phi(v_{it})]^{y_{it}} [1 - \Phi(v_{it})]^{1-y_{it}} \tag{2.31}$$

The posterior $f(\tau_i|\underline{\lambda}, \underline{\theta}, Z)$ is a convolution of a Gaussian density and a product of standard normal CDFs. As such, it can be asymmetric with the direction of skewness depending on the particular realization of the vector of dependent variables $\underline{y}_i$. Therefore, for our simulator we use a piece-wise linear approximation to $f(\tau_i|\underline{\lambda}, \underline{\theta}, Z)$ which is a form of nonparametric EIS capable of accurately sampling from any distribution irrespective of its shape. The procedure works as follows. First, we obtain an empirical distribution function of $f(\tau_i|\underline{\lambda}, \underline{\theta}, Z)$ evaluated over an equispaced grid of τ_i around the importance region and then we invert S draws from $U[0, 1]$ through this edf to obtain a new grid whose values are concentrated in the importance region. We update the edf over this grid and iterate this process until the change of the maxima of the edf parameters (intercept and

slope of individual segments) converges within a tolerance level around zero. Then we invert one draw from $U[0,1]$ for the given τ_i via the final edf to obtain the new value of the τ_i in the Gibbs block. Aside from shape adaptability, another advantage of this nonparametric form of EIS is that the degree of accuracy of this procedure can be made arbitrarily precise by increasing the size of the mesh, at the expense of computational cost.

Sampling from $f(\underline{\lambda}|\underline{\tau},\underline{\theta},Z)$

Given a relatively small $T=5$, we perform a one-shot EIS to draw from this posterior. Let

$$g(\lambda_t) \equiv \prod_{i=1}^{N} \left[\Phi(v_{it})\right]^{y_{it}} \left[1-\Phi(v_{it})\right]^{1-y_{it}}$$

and note that

$$f(\underline{\lambda}|\underline{\tau},\underline{\theta},Z) \propto p(\underline{\lambda}|\underline{\theta}) \prod_{t=1}^{T} g(\lambda_t)$$

where

$$p(\underline{\lambda}|\underline{\theta}) = (2\pi)^{-T/2} \left|\Sigma_\lambda\right|^{-1/2} \exp\left[-\frac{1}{2}\underline{\lambda}'\Sigma_\lambda^{-1}\underline{\lambda}\right] \tag{2.32}$$

We approximate $p(\underline{\lambda}|\underline{\theta}) \prod_{t=1}^{T} g(\lambda_t)$ with a Gaussian kernel $k(\underline{\lambda}|\underline{\tau},\underline{\theta},Z,\underline{\gamma})$ in $\underline{\lambda}$ such that

$$k(\underline{\lambda}|\underline{\tau},\underline{\theta},Z,\underline{\gamma}) = p(\underline{\lambda}|\underline{\theta}) \prod_{t=1}^{T} k_t(\lambda_t;\gamma_t) \tag{2.33}$$

Due to independence of $k_t(\lambda_t;\gamma_t)$ over time, we can perform the EIS regressions of $\ln g(\lambda_t)$ on $\ln k_t(\lambda_t;\gamma_t)$ for each t individually using

$$\ln k_t(\lambda_t;\gamma_t) = -\frac{1}{2}\left[\gamma_{0,t} + \gamma_{1,t}\lambda_t + \gamma_{2,t}\lambda_t^2\right] \tag{2.34}$$

and then recombine $k_t(\lambda_t;\gamma_t)$ with $p(\underline{\lambda}|\underline{\theta})$ into a joint multivariate Gaussian kernel. Let $\underline{\gamma}_2 = \left(\gamma_{2,1},...,\gamma_{2,T}\right)'$, $\Gamma_2 = diag\{\underline{\gamma}_2\}$, and $\underline{\gamma}_1 = \left(\gamma_{1,1},...,\gamma_{1,T}\right)'$. From (2.32), (2.33) and (2.34) we obtain

$$k(\underline{\lambda}|\underline{\tau},\underline{\theta},Z,\underline{\gamma}) = (2\pi)^{-T/2} \left|\Sigma_\lambda\right|^{-1/2} \exp\left[-\frac{1}{2}\underline{\lambda}'\Sigma_\lambda^{-1}\underline{\lambda}\right] \exp\left[-\frac{1}{2}\left(\underline{\lambda}'\Gamma_2\underline{\lambda}+\underline{\gamma}_1'\underline{\lambda}+c\right)\right] \tag{2.35}$$

where $c = \sum_{t=1}^{T}\gamma_{0,t}$. Thus (2.35) is a multivariate Gaussian kernel of

$$\begin{aligned} M(\underline{\lambda}|\underline{\tau},\underline{\theta},Z,\gamma) &\equiv N\left(\Sigma_m,\underline{\mu}_m\right) \\ &= \chi\left(\underline{\tau},\underline{\theta},Z,\underline{\gamma}\right)^{-1} k(\underline{\lambda}|\underline{\tau},\underline{\theta},Z,\underline{\gamma}) \end{aligned}$$

where

$$\begin{aligned} \Sigma_m &= \left[\Sigma_\lambda^{-1}+\Gamma_2\right]^{-1} \\ \underline{\mu}_m &= -\frac{1}{2}\Sigma_m\underline{\gamma}_1 \end{aligned}$$

and the integrating constant

$$\chi\left(\underline{\tau},\underline{\theta},Z,\underline{\gamma}\right)=|\Sigma_m|^{1/2}|\Sigma_\lambda|^{-1/2}\exp\left[\frac{1}{2}\left(\underline{\mu}_m'\Sigma_m^{-1}\underline{\mu}_m-c\right)\right]$$

AR-MH Algorithm

Given K draws $\{\underline{\lambda}_1,...,\underline{\lambda}_K\}$ from the EIS-MCMC algorithm, potential new candidate draws are sampled from $m(\underline{\lambda}|\underline{\tau},\underline{\theta},Z,\widehat{\underline{\gamma}})$ until acceptance of a candidate $\widetilde{\underline{\lambda}}$ in the AR step with probability

$$P(\underline{\lambda})=\min\left(\frac{f(\underline{\lambda}|\underline{\tau},\underline{\theta},Z)}{M(\underline{\lambda}|\underline{\tau},\underline{\theta},Z,\widehat{\underline{\gamma}})},1\right)$$

In the MH-step $\widetilde{\underline{\lambda}}$ is accepted as the $K+1-$th draw $\underline{\lambda}_{K+1}$ from the EIS-MCMC algorithm with probability $\alpha(\underline{\lambda}_K,\widetilde{\underline{\lambda}})$, otherwise $\underline{\lambda}_{K+1}$ is set to equal $\underline{\lambda}_K$. It holds that

$$\alpha(\underline{\lambda}_K,\widetilde{\underline{\lambda}})=\min\left(\frac{f(\widetilde{\underline{\lambda}}|\underline{\tau},\underline{\theta},Z)\min\left[f(\underline{\lambda}|\underline{\tau},\underline{\theta},Z),M(\underline{\lambda}|\underline{\tau},\underline{\theta},Z,\widehat{\underline{\gamma}})\right]}{f(\underline{\lambda}|\underline{\tau},\underline{\theta},Z)\min\left[f(\widetilde{\underline{\lambda}}|\underline{\tau},\underline{\theta},Z),M(\widetilde{\underline{\lambda}}|\underline{\tau},\underline{\theta},Z,\widehat{\underline{\gamma}})\right]},1\right)$$

The AR-MH step for τ_i is repeated 10 times before the parameters are updated in the Gibbs sequence.

Sampling from $f(\sigma_\tau^2|\underline{\theta}_{/\sigma_\tau},\Upsilon,Z)$

We follow the same philosophy of simulated data augmentation as applied in (Albert and Chib, 1993) to draws from $f(\underline{\beta}|\underline{\theta}_{/\underline{\beta}},\Upsilon,Z)$. Since

$$\tau_i\sim N(0,\sigma_\tau^2)$$

the likelihood of the sample $\underline{\tau}$, treated as a function of σ_τ^2, is

$$L(\underline{\tau}|\sigma_\tau^2)=\left(2\pi\sigma_\tau^2\right)^{-N/2}\exp\left(-\frac{1}{2}\frac{S_\tau}{\sigma_\tau^2}\right)$$

where

$$S_\tau=\sum_{i=1}^{N}\tau_i^2$$

A commonly used prior for the variance of Gaussian random variables is the inverted gamma-2 density $IG(s_0,v_0)$ with kernel

$$k_\tau(\sigma_\tau^2)=\sigma_\tau^{-(v_0+2)}\exp\left(-\frac{s_0}{2\sigma_\tau^2}\right)$$

(see Train, 2003, ch. 12). The posterior is then

$$\begin{aligned}f(\sigma_\tau^2|\underline{\theta}_{/\sigma_\tau},\Upsilon,Z) &\propto \sigma_\tau^{-N}\exp\left(-\frac{1}{2}\frac{S_\tau}{\sigma_\tau^2}\right)\sigma_\tau^{-(v_0+2)}\exp\left(-\frac{s_0}{2\sigma_\tau^2}\right)\\ &= \sigma_\tau^{-(v_0+2+N)}\exp\left(-\frac{1}{2}\frac{S_\tau+s_0}{\sigma_\tau^2}\right)\\ &= \sigma_\tau^{-(v_1+2)}\exp\left(-\frac{1}{2}\frac{s_1}{\sigma_\tau^2}\right)\end{aligned}$$

which is a kernel of $IG(s_1, v_1)$ with $s_1 = s_0 + S_\tau$ and $v_1 = v_0 + N$.

Following Bauwens et al. (1999, p. 114) we specify a non-informative prior $\pi(\sigma_\tau^2)$ as the limit of the $IG(s_0, v_0)$ kernel

$$k_\tau(\sigma_\tau^2) = \sigma_\tau^{-(v_0+2)} \exp\left(-\frac{s_0}{2\sigma_\tau^2}\right)$$

where $s_0 \to 0$ and $v_0 = 0$. Thus

$$\begin{aligned} f(\sigma_\tau^2|\underline{\theta}_{/\sigma_\tau}, \Upsilon, Z) &= IG(s_1, v_1) \\ s_1 &= S_\tau \\ v_1 &= N \end{aligned}$$

To draw from this posterior, draw $z \sim U[0,1]$, compute $y_1 = Ga^{-1}\left(\frac{v_1}{2}, 1, z\right)$, $y_2 = \frac{2}{s_1} y_1$ and $\sigma_\tau^2 = y_2^{-1}$.

Sampling from $f(\underline{\rho}|\underline{\theta}_{/\underline{\rho}}, \Upsilon, Z)$

The time random effects λ_t are assumed to follow a stationary autoregressive process of order p

$$A(L)\lambda_t = \sum_{i=0}^{p} (a_i L^i)\lambda_t = \eta_t$$

with $\eta_t \sim N(0, \sigma_\eta^2)$. For AR(1) process, the likelihood function is given by

$$\begin{aligned} L(a; y) &\propto f(\lambda_1, ..., \lambda_t|\sigma_\eta^2, \rho) \\ &= \prod_{t=1}^{T} p(\lambda_t|\lambda_{t-1}, \cdot) \end{aligned}$$

where $f(\lambda_1, ..., \lambda_t|\sigma_\eta^2, \rho)$ is the joint density of $\{\lambda_t\}_{t=1}^T$, and

$$p(\lambda_t|\lambda_{t-1}, \cdot) \propto \begin{cases} \exp\left(-\frac{(1-\rho^2)}{2\sigma_\eta^2}\lambda_1^2\right), & t = 1 \\ \exp\left(-\frac{1}{2\sigma_\eta^2}(\lambda_t - \rho\lambda_{t-1})^2\right), & t = 2, ..., T \end{cases}$$

The joint density is given by

$$\begin{aligned} f(\lambda_1, ..., \lambda_t|\sigma_\eta^2, \rho) &\propto \frac{1}{\sqrt{2\pi\frac{\sigma_\eta^2}{(1-\rho^2)}}} \exp\left(-\frac{(1-\rho^2)}{2\sigma_\eta^2}\lambda_1^2\right) \\ &\quad \times \prod_{t=1}^{T} \left\{ \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp\left(-\frac{1}{2\sigma_\eta^2}(\lambda_t - \rho\lambda_{t-1})^2\right) \right\} \\ &\propto \alpha_\rho \exp\left(-\frac{1}{2}\left[\rho^2\frac{1}{\sigma_\eta^2}\left(\sum_{t=2}^{T}\lambda_{t-1}^2 - \lambda_1^2\right) - \rho\frac{2}{\sigma_\eta^2}\sum_{t=2}^{T}\lambda_t\lambda_{t-1}\right]\right) \\ &\quad \times \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_\eta^2}\sum_{t=1}^{T}\lambda_t^2\right]\right) \end{aligned} \tag{2.36}$$

where

$$\alpha_\rho = \frac{\sqrt{(1-\rho^2)}}{2\pi\sigma_\eta^2}$$

Matching (2.36) with a Gaussian kernel yields

$$\begin{aligned}
\sigma_\rho^2 &= \sigma_\eta^2 \left(\sum_{t=2}^{T} \lambda_t^2 \right)^{-1} \\
\mu_\rho &= \frac{\sigma_\rho^2}{\sigma_\eta^2} \sum_{t=2}^{T} \lambda_t \lambda_{t-1} \\
\gamma_\rho &= -\frac{1}{2\sigma_\eta^2} \sum_{t=1}^{T} \lambda_t^2 - \frac{\mu_\rho^2}{\sigma_\rho^2}
\end{aligned}$$

Hence, draw ρ from $\frac{1}{\exp(\gamma_\rho)} N(\mu_\rho, \sigma_\rho^2)$ truncated to $|\rho| < 1$.

Sampling from $f(\sigma_\eta^2 | \underline{\theta}_{/\sigma_\eta}, \Upsilon, Z)$

For a given $\underline{\lambda}$ and ρ the likelihood function can be formulated as

$$L(\sigma_\eta; a, y) \propto \sigma_\eta^{-T} \left[\exp -\frac{S_\lambda}{2\sigma_\eta^2} \right]$$

where, for AR(1),

$$S_\lambda = (1-\rho^2)\lambda_1^2 \sum_{t=2}^{T} (\lambda_t - \rho\lambda_{t-1})^2$$

Similarly to the case of $f(\sigma_\tau^2 | \underline{\theta}_{/\sigma_\tau}, \Upsilon, Z)$, we postulate the prior on σ_η^2 as the inverted gamma-2 density $IG(s_0, v_0)$ with kernel

$$k_\lambda(\sigma_\lambda^2) = \sigma_\lambda^{-(v_0+2)} \exp\left(-\frac{s_0}{2\sigma_\lambda^2} \right)$$

The posterior becomes

$$\begin{aligned}
f(\sigma_\eta^2 | \underline{\theta}_{/\sigma_\eta}, \Upsilon, Z) &\propto \sigma_\eta^{-T} \left[\exp -\frac{S_\lambda}{2\sigma_\eta^2} \right] \sigma_\tau^{-(v_0+2)} \exp\left(-\frac{s_0}{2\sigma_\lambda^2} \right) \\
&= \sigma_\eta^{-(2+T+v_0)} \exp\left(-\frac{1}{2} \frac{S_\lambda + s_0}{\sigma_\eta^2} \right) \\
&= \sigma_\eta^{-(v_1+2)} \exp\left(-\frac{1}{2} \frac{s_1}{\sigma_\eta^2} \right)
\end{aligned}$$

which is a kernel of $IG(s_1, v_1)$ with $s_1 = s_0 + S_\lambda$ and $v_1 = v_0 + T$. We again specify a non-informative prior $\pi(\sigma_\tau^2)$ as the limit of the $IG(s_0, v_0)$ kernel with $s_0 \to 0$ and $v_0 = 0$. Thus

$$\begin{aligned}
f(\sigma_\eta^2 | \underline{\theta}_{/\sigma_\eta}, \Upsilon, Z) &= IG(s_1, v_1) \\
s_1 &= S_\lambda \\
v_1 &= T
\end{aligned}$$

To draw from this posterior, draw $z \sim U[0,1]$, compute $y_1 = Ga^{-1}\left(\frac{v_1}{2}, 1, z\right)$, $y_2 = \frac{2}{s_1} y_1$ and $\sigma_\eta^2 = y_2^{-1}$.

3.0 LOCALLY WEIGHTED GENERALIZED MINIMUM CONTRAST ESTIMATION UNDER CONDITIONAL MOMENT RESTRICTIONS

Moment restrictions frequently provide the basis for estimation and inference in economic problems. A general framework for analyzing economic data (Y, X) is to postulate conditional moment restrictions of the form

$$E\left[g\left(Z, \theta_0\right)|X\right] = 0 \tag{3.1}$$

Typically, faced with the model (3.1) for estimation of θ_0, researchers would pick an arbitrary matrix-valued function $a(X)$ and estimate $E\left[a(X)g\left(Z, \theta_0\right)\right] = 0$ which is an *unconditional* moment model implied by (3.1) with an estimator such as the Generalized Method of Moments (GMM) (see e.g. Kitamura, 2006, p 26 for a discussion). This procedure is used under the presumption that the chosen instrument $a(X)$ identifies θ, which may not be true even if θ is identified in the conditional model (3.1) (Domínguez and Lobato, 2004). Moreover, the conversion to unconditional moments results in a loss of efficiency with respect to the information contained in (3.1). Chamberlain (1987) showed that such loss can be avoided by using the optimal IV estimator $a^*(X) = D'(X)V^{-1}(X)$ where $D(X) = E\left[\nabla_\theta g\left(Z, \theta_0\right)|X\right]$ and $V(X) = E\left[g\left(Z, \theta_0\right) g\left(Z, \theta_0\right)'|X\right]$. In practice, $a^*(X)$ can be estimated with a two-step procedure (Robinson, 1987; Newey, 1993). First an inefficient preliminary estimator $\widetilde{\theta}$ for θ_0 is obtained and the unknown functions $D(X)$ and $V(X)$ are estimated via a nonparametric regression of $\nabla_\theta g(Z, \widetilde{\theta})$ and $g(Z, \widetilde{\theta})g(Z, \widetilde{\theta})'$ on X. Second, the estimate of $a^*(X)$ is constructed with the estimates of $D(X)$ and $V(X)$ from the first step. However, as noted by Domínguez and Lobato (2004), the resulting moment condition $E\left[a^*(X)g\left(Z, \theta_0\right)\right] = 0$ may fail to identify θ while θ is identified under the original model (3.1). Moreover, satisfactory

implementation of the nonparametric regression may require large samples thereby affecting the finite-sample performance of the feasible estimator of $a^*(X)$.

The methods typically employed for estimation of $E\left[a(X)g\left(Z,\theta_0\right)\right]=0$ have also been subject to criticism. While the optimally-weighted two-step GMM (Hansen, 1982) is first-order asymptotically efficient, its finite sample properties have been reported as relatively poor. For example, a simulation study by Altonji and Segal (1996) documented a substantial small-sample bias of GMM when used to estimate covariance models. Other Monte Carlo experiments have shown that tests based on GMM often have true levels that differ greatly from their nominal levels when asymptotic critical values are used (Hall and Horowitz, 1996). Indeed, it has been widely recognized that the first-order asymptotic distribution of the GMM estimator provides a poor approximation to its finite-sample distribution (Ramalho, 2005).

A number of alternative estimators have been suggested to overcome this problem: Empirical Likelihood (EL) (Owen, 1988; Qin and Lawless, 1994; Imbens, 1997), the Euclidean Likelihood (EuL) corresponding to the Continuous Updating Estimator (CUE) (Hansen et al., 1996) the Exponential Tilting Estimator (ET) (Kitamura and Stutzer, 1997; Imbens et al., 1998), and variations on these such as the Exponentially Tilted Empirical Likelihood (ETEL) (Schennach, 2006). The EL, EuL and ET share some common properties and can be derived from a common model basis for estimation. Thus, they and can be viewed as members of broader classes of estimators such as the Generalized Empirical Likelihood (GEL) estimators (Smith, 1997; Newey and Smith, 2004) and the Generalized Minimum Contrast (GMC) estimators (Bickel et al., 1998). Recently, Kitamura (2006) showed that for unconditional moment restriction models, the GEL class is essentially equivalent to the GMC class even if the GEL are derived somewhat differently from the GMC. Both GEL and GMC lead to the same saddle-point optimization problem yielding the same form the individual estimators.

The GEL/GMC estimators circumvent the need for estimating a weight matrix in the two-step GMM procedure by directly minimizing an information-theory-based concept of closeness between the estimated distribution and the empirical distribution. A growing body of Monte Carlo evidence has revealed favorable finite-sample properties of the GEL/GMC estimators compared to GMM (see e.g. Ramalho, 2005, and references therein). Recently,

Newey and Smith (2004) showed analytically that while GMM and GEL share the same first-order asymptotic properties, their higher-order properties are different. Specifically, while the asymptotic bias of GMM often grows with the number of moment restrictions, the relatively smaller bias of EL does not. Moreover, after EL is bias corrected (using probabilities obtained from EL) it is higher-order efficient relative to other bias-corrected estimators.[1]

It is worth emphasizing that the GMM and GEL/GMC estimators mentioned so far are all based on *unconditional* moment restrictions

$$E\left[g\left(X,\theta_0\right)\right]=0 \tag{3.2}$$

burdened by the potential pitfalls described above. In addressing this problem, Kitamura, Tripathi, and Ahn (2004) (henceforth KTA) recently developed a Conditional Empirical Likelihood (CEL) estimator that makes efficient use of the information contained in (3.1). Their one-step estimator achieves the semiparametric efficiency bound without explicitly estimating the optimal instruments. Similar analysis has been performed by Antoine, Bonnal, and Renault (2006) (henceforth ABR) for the case of Conditional Euclidean Likelihood[2] and Smith (2003, 2006) for the Cressie-Read family of estimators.

In this Chapter we extend this line of research by proposing a new class of estimators based directly on conditional moment restrictions that encompasses the entire GMC family. Moreover, using the GMC information-theoretic framework we show that in constructing the estimators for the *conditional* moment restrictions (3.1) the previous literature implicitly use an arbitrary uniform weighting scheme. This leads to minimizing a discrepancy from a probability measure that is different from the one under which the data was distributed. The reason for this phenomenon is that the previously analyzed estimators for (3.1) are based on local kernel smoothing of the unconditional statistical model (3.2). In contrast, we consider an information-theoretic dual locally weighted GMC optimization problem built directly on (3.1) that minimizes a discrepancy from a probability measure according to which the

[1] Accordingly, the initial focus of this paper lies in EL as opposed to any other member of the GEL family of estimators.

[2] ABR show that the Euclidean empirical likelihood estimator coincides with the continuously updated GMM (CUE-GMM) as first proposed by Hansen et al. (1996).

data was distributed. Consequently, our newly proposed class of estimators includes locally weighted alternatives to the estimators analyzed in previous literature, in particular the Locally Weighted Conditional Empirical Likelihood (LWCEL) . We analyze the differences between the new LWCEL and KTA's CEL in detail. In a Monte Carlo study we show that the LWCEL estimator exhibits better finite-sample properties than found in the previous literature.

EXISTING METHODS

Information-theoretic Approaches to Estimation

In this Section we will now develop some intuition useful for subsequent analysis by briefly introducing the heuristic background behind GMM estimation and information-theoretic alternatives such as empirical likelihood. In general terms, suppose that theory is represented by the prediction $E_Q\left[g\left(X,\theta_0\right)\right]=0$. GMM-type estimators are defined by setting the sample moments as close as possible to the zero vector of population moments fixed by the probability measure Q.

In contrast, the information-theoretic approach focuses on a change of measure $dQ/d\Pi$ which enables $\theta\neq\theta_0$ to satisfy the transformed condition $E_\Pi\left[g\left(X,\theta\right)\right]=0$. The estimator of θ_0 then sets the probability measure Π as close as possible to Q. Such approach thus uses closeness of probability measures, rather than moments, to estimate θ_0.

More specifically, define by $\mathcal{P}(\theta)$ the set of probability measures Π that satisfy a given condition, such as $E_\Pi\left[g\left(X,\theta\right)\right]=0$. In order to find the most suitable Π for each $\theta\in\Theta$, the information-theoretic approach suggests the use of the convex optimization problem

$$\min_{\Pi\in\mathcal{P}(\theta)} D\left(\Pi,Q\right) \quad \text{s.t.} \quad E_\Pi\left[g\left(Z,\theta\right)\right]=0 \tag{3.3}$$

where $D\left(\Pi,Q\right)$ is a measure of divergence between Π and Q,

$$D\left(\Pi,Q\right)=\int\phi\left(\frac{d\Pi}{dQ}\right)dQ \tag{3.4}$$

(Csiszar, 1967). For a finite sample distributed according to Q, the resulting estimator of θ_0 minimizes the finite-sample counterpart of (3.3) over Θ. In practice, this involves "re-weighting" the sample data to fit the given restriction.

The information-theoretic approach has a long history in mathematical statistics. Its theoretical basis includes maximum entropy principle (Jaynes, 1957) and the principle of minimum discrimination information (Kullback and Leibler, 1951). These principles are related to Bayesian methods in that they make explicit use of prior information (Kullback, 1997).

Unconditional Moment Restrictions

A substantial body of literature has been devoted to estimation under the *unconditional* moment restriction (3.2). In contrast to the conditional case (3.1), under the unconditional framework all data is treated as exogenous which results in significant simplifications in subsequent analysis. Most notably, Qin and Lawless (1994), Hansen et al. (1996), Kitamura and Stutzer (1997), Imbens et al. (1998), Newey and Smith (2004), and Schennach (2006) belong to this category. In a comprehensive manuscript, Kitamura (2006) elaborates on the use of duality theory from convex analysis in construction of a general class of unconditional GMC estimators. This elegant framework enables one to derive a computationally friendly saddle-point GMC estimator from a dual optimization problem directly related to a primal unfeasible optimization problem that is based on an information-theoretic population specification. This approach, which we build on herein, is tantamount to a generic version of the Lagrange multiplier derivation of GEL estimators utilized in earlier literature.

Conditional Moment Restrictions

Estimation techniques based directly on the *conditional* moment restrictions (3.1) have so far been analyzed for *special cases* of the finite-sample conditional counterpart of the divergence measure (3.4): the Conditional Empirical Likelihood (CEL) with

$$\phi\left(\frac{\pi(x_{ij})}{q(x_{ij})}\right) = -\log\left(\frac{\pi(x_{ij})}{q(x_{ij})}\right)$$

by KTA, the Conditional Euclidean Likelihood with

$$\phi(x) = \frac{1}{2}\left[\left(\frac{\pi(x_{ij})}{q(x_{ij})}\right)^2 - 1\right]$$

by ABR, and the Cressie-Read parametric family with

$$\phi\left(\frac{\pi(x_{ij})}{q(x_{ij})}\right) = \frac{2}{\gamma(\gamma+1)}\left[\left(\frac{\pi(x_{ij})}{q(x_{ij})}\right)^{-\gamma} - 1\right]$$

where $\gamma \in \mathbb{R}$ by Smith (2006). These estimators are all based on local kernel smoothing of the unconditional model (3.2).

CONDITIONAL MOMENT RESTRICTIONS: ALTERNATIVE ESTIMATION METHOD

In this Chapter, we derive a new class of estimators for a *generic functional form* of ϕ requiring only that ϕ be convex on its domain. Based on such generic ϕ we specify an information-theoretic dual locally weighted conditional GMC optimization problem that minimizes a discrepancy from a probability measure according to which the data was distributed. Consequently, we propose a new class of estimators that include locally weighted alternatives to the estimators analyzed in previous literature, in particular the Locally Weighted Conditional Empirical Likelihood (LWCEL).

The theoretical foundations of our new class of estimators extend the dual GMC approach of Kitamura (2006) to account specifically for the conditional moment restrictions. In contrast to a single GMC optimization problem utilized in Kitamura (2006) suitable for the unconditional moments (3.2), though, we consider a continuum of GMC optimization problems - one at each X. The resulting estimator then minimizes the expected value of the primal or dual GMC value functions, the expectation being taken with respect to the marginal distribution of the exogenous variables X.

Stochastic Environment

Suppose that the observations $\{(x_i, y_i) : i = 1, ..., n\}$ are drawn independently from the joint distribution $Q(x, y)$ with support $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is a compact subset of $\mathbb{R}^{d_X}$ and $\mathcal{Y}$ is a subset of $\mathbb{R}^{d_Y}$. Suppose that the unknown distribution $Q(x, y)$ satisfies the conditional moment restrictions given by (3.1), where $g : Z \times \Theta \to \mathbb{R}^{d_g}$ is a known mapping, up to an unknown vector of parameters $\theta_0 \in \Theta$, and $Z \equiv (Y', X_z')' \in Y \times X_Z \equiv Z \subseteq \mathbb{R}^{d_Z}$ where $X_Z \subseteq X$. The restriction (3.1) can then be reformulated as

$$\int g(Z, \theta_0)\, dQ(y|x) = 0$$

where $Q(y|x)$ is the "true" conditional distribution of Y given X.

Information-theoretic Model of the Conditional GMC Problem

In addition to conditioning, our general approach to specifying the GMC optimization problem differs from Kitamura (2006) by another important aspect: instead of Q and Π we will formulate the population specification as one involving the derivatives all probability measures taken with respect to the Lebesgue measure using the concept of Radon-Nikodym derivative (Royden, 1987). This will enable us to account explicitly for the differences between marginal and conditional densities and hence derive the conditional version of the GMC estimator. In Kitamura (2006)'s unconditional case, such distinction was unnecessary and therefore not specified. Hence, denote by $\pi(y|x)$, $q(y|x)$, $\pi(x, y)$, $q(x, y)$, $\pi(x)$, $q(x)$ the Radon-Nikodym derivatives of the probability measures $\Pi(y|x)$, $Q(y|x)$, $\Pi(x, y)$, $Q(x, y)$, $\Pi(x)$, $Q(x)$ with respect to the Lebesgue measure $m(\cdot)$, respectively.

Let $\mathbf{M}_Y$ denote the set of all probability densities on $\mathbb{R}^{d_Y}$ and let

$$\boldsymbol{\pi}(X; \theta) \equiv \left\{ \pi(y|x) \in \mathbf{M}_Y : \int \pi(y|x) g(Z, \theta)\, dm(y|x) = 0;\ X \in \mathcal{X} \right\}$$

Define the set of all probability densities that are compatible with the conditional moment restriction (3.1) by

$$\boldsymbol{\pi}(X) \equiv \cup_{\theta \in \Theta} \boldsymbol{\pi}(X; \theta)$$

The set $\boldsymbol{\pi}(X)$ indexed by X represents a statistical model that is correctly specified if $q(y|x) \in \boldsymbol{\pi}(X)$.

Consider the measure of conditional divergence[3]

$$\begin{aligned} D\left(\Pi(y|x), Q(y|x)\right) &= \int_{\mathcal{Y}} \phi\left(\frac{d\Pi(y|x)}{dQ(y|x)}\right) dQ(y|x) \\ &= \int q(y|x)\phi\left(\frac{\pi(y|x)}{q(y|x)}\right) dm(y|x) \end{aligned} \tag{3.5}$$

where ϕ is a convex function and $\Pi(y|x)$ is absolutely continuous with respect to $Q(y|x)$ (for other cases let $D \equiv \infty$). Note that $D\left(\cdot, Q(y|x)\right)$ attains its minimum at $Q(y|x)$. For a given $X \in \mathcal{X}$, at the population level, the GMC optimization problem is specified as

$$\inf_{\theta \in \Theta} \rho\left(\theta, Q(x,y)\right) \equiv \inf_{\pi(y|x) \in \boldsymbol{\pi}(X)} D\left(\pi(y|x), q(y|x)\right) \tag{3.6}$$

which, using (3.5), corresponds to the the constrained optimization problem

$$\begin{aligned} v(X;\theta) &= \inf_{\pi(y|x) \in \mathbf{M}_Y} D\left(\Pi(y|x), Q(y|x)\right) \\ &= \inf_{\pi(y|x) \in \mathbf{M}_Y} \int q(y|x)\phi\left(\frac{\pi(y|x)}{q(y|x)}\right) dm(y|x) \end{aligned}$$

subject to

$$\begin{aligned} \int \pi(y|x) g\left(Z, \theta\right) dm(y|x) &= 0 \\ \int \pi(y|x) dm(y|x) &= 1 \end{aligned}$$

for a given $\theta \in \Theta$ where $v(X;\theta)$ is the value function. In convex analysis, such problem is called the *primal* problem (Borwein and Lewis, 2006). Our formulation corresponds to the conditional version of the primal convex optimization problem (3.3).

The estimator of θ_0 should minimize the value function in the primal problem. Since our value function $v(X;\theta)$ is defined for each X, we specify the estimator as one that minimizes the expected value function where the expectation is taken over X with respect to the

[3]This conditional measure of divergence is a natural extension of the conditional discrepancy measure formulated by Shannon (1948) for the special case of conditional entropy with $\phi(x) = x\log(x)$.

probability measure $Q(x)$ according to which the exogenous X were distributed. Hence, as the basis for finite-sample estimation, θ_0 solves

$$\begin{aligned}\theta_0 &= \arg\min_{\theta\in\Theta} E_{Q(x)}\left[v(X;\theta)\right] \\ &= \arg\min_{\theta\in\Theta} \int v(X;\theta)dQ(X) \\ &= \arg\min_{\theta\in\Theta} \int q(X)v(X;\theta)dm(x)\end{aligned} \tag{3.7}$$

The marginal distribution of X is independent of the parameter θ and hence the former can be estimated directly from the data. The same holds for the "choice" marginal distribution $\Pi(x)$ in the optimization problem and hence $\Pi(x) = Q(x)$. Multiplying the argument inside $\phi(\cdot)$ by $\frac{d\Pi(x)}{dQ(x)} = \frac{\pi(x)}{q(x)} = 1$ we obtain

$$\begin{aligned}E_{Q(x)}\left[v(X;\theta)\right] &= \int \left[\inf_{\pi(y|x)\in\mathbf{M}_Y} \int q(y|x)\phi\left(\frac{\pi(y|x)}{q(y|x)}\right)dm(y|x)\right]dQ(x) \\ &= \int q(x)\left[\inf_{\pi(y|x)\in\mathbf{M}_Y} \int q(y|x)\phi\left(\frac{\pi(y|x)}{q(y|x)}\right)dm(y|x)\right]dm(x) \\ &= \inf_{\{\pi(y|x)\in\mathbf{M}_Y : X\in\mathcal{X}\}} \int\int q(x)q(y|x)\phi\left(\frac{\pi(y|x)}{q(y|x)}\frac{\pi(x)}{q(x)}\right)dm(y|x)dm(x) \\ &= \inf_{\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} \int q(x,y)\phi\left(\frac{\pi(x,y)}{q(x,y)}\right)dm(x,y) \\ &= \inf_{\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} D\left(\Pi(x,y), Q(x,y)\right)\end{aligned} \tag{3.8}$$

and hence θ_0 minimizes the divergence between the two *joint* distributions $\Pi(x,y)$ and $Q(x,y)$.

Since the primal problem involves a numerically unfeasible optimization over function, it is beneficial to convert it into its dual form that facilitates feasible finite-dimensional optimization. There are numerous results in convex analysis that specify the conditions for existence of the dual form (see e.g. Luenberger, 1969; Borwein and Lewis, 2006). For a given $X \in \mathcal{X}$, the primal problem (3.6) corresponds to the *dual* problem

$$v^*(X;\theta) = \max_{\lambda(X)\in\mathbb{R}^{d_g},\mu(X)\in\mathbb{R}} \left[\mu - \int q(y|x)\phi^*\left(\mu(X) + \lambda(X)'g\left(Z,\theta\right)\right)dm(y|x)\right] \tag{3.9}$$

where $\phi^*(\cdot)$ is the convex conjugate (or Legendre transformation) of $\phi(\cdot)$. This is a finite-dimensional unconstrained convex maximization problem. By Fenchel duality,

$$v(X;\theta) = v^*(X;\theta)$$

Analogously to (3.6), θ_0 solves the minimization problem

$$\begin{aligned}
\inf_{\theta\in\Theta} E_{Q(x)}\left[v^*(X;\theta)\right] &= \inf_{\theta\in\Theta}\int q(x)\max_{\lambda\in\mathbb{R}^{d_g},\mu\in\mathbb{R}}\left[\mu(X) - \int q(y|x)\phi^*\left(\mu(X)+\lambda(X)'g\left(Z,\theta\right)\right)dm(y|x)\right]dm(x) \\
&= \inf_{\theta\in\Theta}\max_{\lambda\in\mathbb{R}^{d_g},\mu\in\mathbb{R}}\left[\int q(x)\mu(X)dm(x) - \int\int q(x)q(y|x)\phi^*\left(\mu(X)+\lambda(X)'g\left(Z,\theta\right)\right)dm(y|x)dm(x)\right] \\
&= \inf_{\theta\in\Theta}\max_{\lambda\in\mathbb{R}^{d_g},\mu\in\mathbb{R}}\left[\int q(x)\mu(X)dm(x) - \int q(x,y)\phi^*\left(\mu(X)+\lambda(X)'g\left(Z,\theta\right)\right)dm(x,y)\right]
\end{aligned}$$

Given a sample $\{(x_i,y_i) : i = 1,...,n\}$ from $Q(x,y)$, the population criteria described above provide a basis for statistical inference wherein we replace the unknown probability measures $Q(x,y)$ and $Q(y|x)$ with their empirical counterparts $Q(x_i,y_j)$ and $Q(y_j|x_i)$, respectively. However, in contrast to the unconditional case where it suffices to set $q(x_i) = 1/n$, the densities $q(x,y)$ and $q(y|x)$ now need to be estimated nonparametrically as probability mass functions $q(x_i,y_j)$ and $q(y_j|x_i)$ with support on the data. Numerous methods have been suggested in the literature to obtain such estimates with various desirable properties using e.g. kernels, series or nearest neighbors to name just a few (see e.g. Pagan and Ullah, 1999, and references therein).

A sample version of the GMC problem (3.6) is

$$\text{minimize } \widehat{v}(\theta) \equiv \left\{\sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i,y_j)\phi\left(\frac{\pi(y_j|x_i)}{q(y_j|x_i)}\right) : \sum_{j=1}^{n}\pi(y_j|x_i)g\left(z_j,\theta\right) = 0, \sum_{j=1}^{n}\pi(y_j|x_i) = 1\right\} \tag{3.10}$$

This leads to the Locally Weighted Conditional GMC estimator for θ

$$\widehat{\theta} = \arg\min_{\theta\in\Theta}\widehat{v}(\theta) \tag{3.11}$$

This estimator corresponds to the *conditional* locally weighted forms of the "Minimum Discrepancy Statistic" of Corcoran (1998) and the "Minimum Distance Estimator" of Newey and Smith (2004).

The primal optimization problem (3.10) corresponds to a computationally convenient dual problem (Borwein and Lewis, 2006)

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \widehat{v}^*(\theta) \tag{3.12}$$

where

$$\widehat{v}^*(\theta) \equiv \max_{\lambda\in\mathbb{R}^{d_g},\mu\in\mathbb{R}} \left[\sum_{i=1}^{n} q(x_i)\mu(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i,y_j)\phi^*\left(\mu(x_i) + \lambda(x_i)'g\left(z_j,\theta\right)\right)\right]$$

For a sample $\{(y_i, x_i) : i = 1, ..., n\}$ estimation of $q(y|x)$ and $q(x, y)$ amounts to the use of localization methods (Tibshirani and Hastie, 1987). In the stream of literature most relevant to this paper, localization schemes have been used in the conditional moment context in LeBlanc and Crowley (1995), Zhang and Gijbels (2003), KTA for CEL, ABR for the EuL, and Smith (2003, 2005) for GEL. Information on $Q(y|x)$ is inferred from the nearby observations if we assume that $Q(y|x)$ is continuous with respect to X. In other words, in a neighborhood around x_i we approximate $Q(y|x)$ by $Q(y|x) \approx Q(y|x_i)$. This implies that all the z_j with x_j lying in this neighborhood can be roughly viewed as observations from $Q(y|x_i)$. Note that, unlike in the unconditional moment case (3.2) where $q(x_i) = 1/n$, now the $q(x_i, y_j)$ and $q(y_j|x_i)$ are not derived directly from observed data, since only one realization of the random vector y_j was actually observed at x_i. Rather, these probability masses are inferred from neighboring observations. The data-determined $q(x_i, y_j)$ and $q(y_j|x_i)$ are then used as a benchmark in the value function of the GMC optimization problem in derivations of $\widehat{\theta}$.

Locally Weighted Conditional Empirical Likelihood

Various choices for the discrepancy measure $\phi(\cdot)$ lead to various special cases of the Dual Locally Weighted Conditional GMC estimator. Setting $\phi(x) = -\log(x)$ corresponds to Locally Weighted Conditional Empirical Likelihood (LWCEL). The unfeasible GMC estimator of (3.7) becomes

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \widehat{v}(\theta) \equiv \left\{ -\sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i,y_j)\log\left(\frac{\pi(y_j|x_i)}{q(y_j|x_i)}\right) : \sum_{j=1}^{n}\pi(y_j|x_i)g\left(z_j,\theta\right) = 0, \sum_{j=1}^{n}\pi(y_j|x_i) = 1 \right\} \tag{3.13}$$

The convex conjugate of $\phi(x) = -\log(x)$ is $\phi^*(y) = -1 - \log(-y)$. Using this expression in the feasible dual formulation (3.12) we obtain

$$\widehat{\theta}_{LWCEL} = \arg\min_{\theta\in\Theta} \widehat{v}^*(\theta) \equiv \max_{\lambda\in\mathbb{R}^{d_g},\mu\in\mathbb{R}} \left[\sum_{i=1}^{n} q(x_i)\mu(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i,y_j) \log\left(-\mu(x_i) - \lambda(x_i)' g\left(z_j,\theta\right)\right) \right]$$

From (3.8), it is worth noting that on the population level, the LWCEL minimizes the discrepancy measure

$$\begin{aligned} D\left(\Pi(x,y), Q(x,y)\right) &= \int \log\left(\frac{dQ(x,y)}{d\Pi(x,y)}\right) dQ(x,y) \\ &= K(Q(x,y), \Pi(x,y)) \end{aligned}$$

where $K(Q(x,y), \Pi(x,y))$ is the Kullback-Leibler (KL) divergence between the *joint* probability measures $Q(x,y)$ and $\Pi(x,y)$ with $Q(x,y)$ being the true probability measure according to which the data are distributed. The $\widehat{\theta}_{LWCEL}$ then solves the minimization problem

$$\inf_{\theta\in\Theta} \inf_{\boldsymbol{\pi}(x,y):\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} K(Q_n(x,y), \Pi(x,y))$$

where $Q_n(x,y)$ is the empirical measure and $\Pi(x,y)$ represents the moment conditions model.

Note that this estimator contains two important modifications in comparison to the Conditional Empirical Likelihood (CEL) analyzed by KTA specified in our notation as

$$\widehat{\theta}_{CEL} = \arg\min_{\theta\in\Theta} \max_{\lambda\in\mathbb{R}^{d_g}} \left[\sum_{i=1}^{n}\sum_{j=1}^{n} q(y_j|x_i) \log\left(1 + \lambda(x_i)' g\left(z_j,\theta\right)\right) \right]$$

First, the weight of the logarithmic function in $\widehat{\theta}_{CEL}$ is $q(y_j|x_i)$ as opposed to $q(x_i,y_j)$ in $\widehat{\theta}_{LWCEL}$. This is a consequence by taking simple summation of the local discrepancies at x_i in derivation of $\widehat{\theta}_{CEL}$ as opposed to a weighted sum that would capture the relative importance of each local discrepancy in the global objective function. Thus, in the population version of the GMC optimization problem with $E_{m(X)}\left[v(X;\theta)\right]$ the $\widehat{\theta}_{CEL}$ minimizes $D\left(\Pi(y|x), U(X)Q(y|x)\right)$ as opposed to $D\left(\Pi(x,y), Q(x,y)\right)$ for $\widehat{\theta}_{LWCEL}$, where $U(x)$ is the uniform probability measure over X. However, $Q(x,y) \neq U(x)q(y|x)$, almost surely. Second, $\widehat{\theta}_{CEL}$ sets $\mu(x_i) = 1$ which is an artefact of using a specific kernel estimation method where individual weights sum up to 1. In general, however, $\mu(x_i) \neq 1$ a.s.

A closer look on the structure of the optimization problem behind $\widehat{\theta}_{LWCEL}$ reveals interesting comparisons with the form of empirical likelihood established in the literature for unconditional moment restrictions. Taking first-order conditions of the GMC Lagrangian

$$\begin{aligned} L(\theta,\lambda,\mu,\pi) &= \sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i,y_j)\ln\left(\frac{\pi(y_j|x_i)}{q(y_j|x_i)}\right) - \sum_{i=1}^{n}\lambda(x_i)'\sum_{j=1}^{n}\pi(y_j|x_i)g\left(z_j,\theta\right) \\ &\quad - \sum_{i=1}^{n}\mu(x_i)\left(\sum_{j=1}^{n}\pi(y_j|x_i)-1\right) \end{aligned} \tag{3.14}$$

corresponding to the GMC objective function (3.13) yields

$$\frac{\widehat{q}(x_i,y_j)}{\widehat{\pi}(y_j|x_i)} = \widehat{\lambda}(x_i)'g\left(z_j,\widehat{\theta}\right) + \widehat{\mu}_i \quad , \ \forall i,j \tag{3.15}$$

$$\sum_{j=1}^{n}\widehat{\pi}(y_j|x_i)g\left(z_j,\widehat{\theta}\right) = 0 \quad , \ \forall i \tag{3.16}$$

$$\sum_{j=1}^{n}\widehat{\pi}(y_j|x_i) = 1 \tag{3.17}$$

Summing (3.15) over j and using (3.16) yields, for each i,

$$\begin{aligned} \sigma(x_i) &\equiv \sum_{j=1}^{n}\widehat{q}(x_i,y_j) \\ &= \widehat{\lambda}(x_i)'\sum_{j=1}^{n}\widehat{\pi}(y_j|x_i)g\left(z_j,\widehat{\theta}\right) + \widehat{\mu}(x_i)\sum_{j=1}^{n}\widehat{\pi}_{ij} \\ &= \widehat{\mu}(x_i) \end{aligned} \tag{3.18}$$

Substituting (3.18) into (3.15) gives, for each i and j,

$$\widehat{\pi}(y_j|x_i) = \frac{\widehat{q}(x_i,y_j)}{\sigma(x_i)+\widehat{\lambda}(x_i)'g\left(z_j,\widehat{\theta}\right)} \tag{3.19}$$

Substituting (3.19) into the Lagrangian (3.14), and using (3.16) and (3.17), yields

$$L(\theta,\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\ln\left(\frac{\widehat{q}(x_i)}{\sigma(x_i)+\widehat{\lambda}(x_i)'g\left(z_j,\widehat{\theta}\right)}\right) \tag{3.20}$$

Then the Locally Weighted Conditional Empirical Likelihood estimator with the new weighting scheme is defined as

$$\widehat{\theta}_{LWCEL} = \arg\max_{\theta\in\Theta} L(\theta,\lambda_i) \tag{3.21}$$

where $\widehat{\lambda}_i$ solves[4]

$$\sum_{j=1}^{n} \frac{\widehat{q}(x_i, y_j) g\left(z_j, \widehat{\theta}\right)}{\sigma_i + \widehat{\lambda}_i' g\left(z_j, \widehat{\theta}\right)} = 0$$

obtained from (3.16) and (3.19). As discussed above, in general $\sigma_i \neq 1$. The presence of σ_i is the hallmark of LWCEL compared to the previous literature where, invariably, $\sigma_i = 1$.

The $\widehat{\theta}_{LWCEL}$ estimator defined in (3.21) is a special case of a corresponding estimator derived under semiparametric conditional moment restrictions in the next Chapter. For this reason, we will perform the asymptotic analysis pertaining to both estimators in the next chapter. The MD estimator analyzed by Smith (2003, 2005) as well as the CEL estimator elaborated in KTA achieve the semiparametric efficiency lower bound (see Chamberlain, 1987). The weighting introduced for $\widehat{\theta}_{LWCEL}$ in this paper postulates more flexible weights that improve on the fixed-bandwidth kernel weights in finite samples in terms of MSE. We conclude that our new forms of the MD and CEL estimators exhibit first-order asymptotic equivalence in terms of consistency and asymptotic normality with the ones formulated in the previous literature, and hence also achieve the first-order asymptotic semiparametric efficiency lower bound. However, our $\widehat{\theta}_{LWCEL}$ improves on its previously analyzed forms in terms of finite sample performance.

Other GMC Class Members

Other choices of $\phi(x)$ lead to various other estimators but these are not the subject of focus of this Dissertation. Therefore, we will only briefly touch upon their derivation from the GMC class without performing the asymptotic analysis verifying their validity.

Let $\phi(x) = x\log(x)$ implying $\phi^*(y) = e^{y-1}$. From (3.12), the Locally Weighted Conditional Exponential Tilting (LWCET) estimator is obtained as

$$\widehat{\theta}_{LWCET} = \arg\min_{\theta\in\Theta} \widehat{v}^*(\theta) \equiv \max_{\lambda\in\mathbb{R}^{d_g}, \mu\in\mathbb{R}} \left[\sum_{i=1}^{n} q(x_i)\mu(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} q(x_i, y_j)\phi^*\left(\mu(x_i) + \lambda(x_i)' g\left(z_j, \theta\right)\right) \right]$$

[4] In line with KTA we adopt the notation $\widehat{\lambda}_i$ as shorthand for $\widehat{\lambda}(x_i, \widehat{\theta})$. In the same spirit, we denote $\sigma(x_i)$ with σ_i in the sequel. When necessary, we explicitly write the full form to ensure that our arguments are unambiguous.

In the primal GMC problem, using (3.7) and (3.8), the exponential tilting estimator minimizes

$$\begin{aligned} E_{Q(X)}\left[v(X;\theta)\right] &= \inf_{\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} \int q(x,y)\phi\left(\frac{\pi(x,y)}{q(x,y)}\right) dm(x,y) \\ &= \inf_{\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} \int \frac{\pi(x,y)}{q(x,y)} q(x,y) \log\left(\frac{\pi(x,y)}{q(x,y)}\right) dm(x,y) \\ &= \inf_{\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} \int \pi(x,y) \log\left(\frac{\pi(x,y)}{q(x,y)}\right) dm(x,y) \\ &= \inf_{\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} \int \log\left(\frac{\Pi(x,y)}{Q(x,y)}\right) d\Pi(x,y) \\ &= \inf_{\pi(x,y)\in\{\mathbf{M}_Y : X\in\mathcal{X}\}} \int K(\Pi(x,y), Q(x,y)) \end{aligned}$$

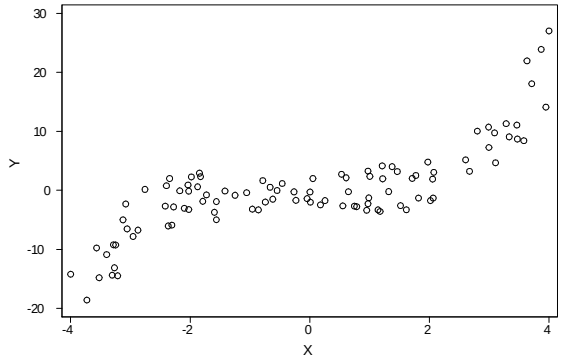
which is the KL divergence with the roles of $Q(x,y)$ and $\Pi(x,y)$ reversed relative to empirical likelihood.

A particularly convenient parametrization of $\phi(x)$, the Cressie-Read (CR) form $\phi(x) = \frac{2}{\gamma(\gamma+1)}(x^{-\gamma} - 1)$ has been used extensively in the literature on the unconditional case (3.2). The conjugate is given by $\phi^*(y) = -\frac{2}{\gamma}\left(-\frac{\gamma+1}{2}y\right)^{\frac{\gamma}{\gamma+1}} + \frac{2}{\gamma(\gamma+1)}$. Parameter values $\gamma = -2, -1, 0$ and 1 yield Euclidean likelihood, exponential tilting, empirical likelihood and Pearson's χ^2, respectively. The conditional case (3.1) has been analyzed by Smith (2006). Nonetheless, an analogous difference as described for $\widehat{\theta}_{LWCEL}$ vs. $\widehat{\theta}_{CEL}$ also holds for the locally weighted CR family of estimators introduced here as opposed to the ones considered in Smith (2006).

SIMULATION

To evaluate the finite sample performance of the estimator $\widehat{\theta}_{LWCEL}$ defined in (3.21) against KTA's $\widehat{\theta}_{CEL}$ we have conducted a small scale pilot Monte Carlo (MC) simulation study aimed at maximum simplicity of the simulation design. We set $Z = X$ and $Y = \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + e$ with heteroskedastic $e = 0.5u|X|$, $u = U(-5,5)$. A random sample $N = 100$ of $X \tilde{} N(0,2)$ was truncated at -1 and 1 and spread over the interval $[-4,4]$ to avoid far outliers. The true parameter values were set at $\beta_1 = -0.2$, $\beta_2 = 0.1$, $\beta_3 = 0.3$. A typical data draw looks as illustrated in 3.1

Figure 3.1: Sample Simulated Data



In order to deal with possible negative arguments in the log function, we followed the approach suggested by Owen (2001) cited in Kitamura (2006) (p. 51): for a small number $\delta = 0.2$ we used the objective function

$$\log_* y = \begin{cases} \log(y) & \text{if } y > \delta \\ \log(\delta) - 1.5 + 2y/\delta - \delta^2/2\delta^2 & \text{if } y \leq \delta \end{cases}$$

Indeed, the proportion of $y \leq \delta$ in the overall sample was 6.6×10^{-3} and 4.7×10^{-3} for $\widehat{\theta}_{LWCEL}$ and $\widehat{\theta}_{CEL}$, respectively. The Nadaraya-Watson kernel estimator (Pagan and Ullah, 1999, p.86) with the Gaussian kernel, employing the Silverman's rule of thumb for the bandwidth determination (Silverman, 1986, p.45), was used to calculate $q(x_i, y_j)$ the case of $\widehat{\theta}_{CEL}$. Thus each i-th local conditional empirical likelihood of $\widehat{\theta}_{CEL}$ was normalized with its corresponding $\sum_{j=1}^{N} q(x_i, y_j)$ in the denominator of the Nadaraya-Watson kernel estimator. In contrast, the denominator of the Nadaraya-Watson kernel estimator was replaced with $n^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} q(x_i, y_j)$ for the case of $\widehat{\theta}_{LWCEL}$. This is equivalent (up to a constant of proportionality) to weighting each i-th local conditional empirical likelihood of $\widehat{\theta}_{LWCEL}$ with σ_i. We compared bias, variance and mean-square error over 100 MC iterations on the three estimated coefficients β_1, β_2 and β_3. The results are presented in Table 3.1.

Table 3.1: Simulation Results

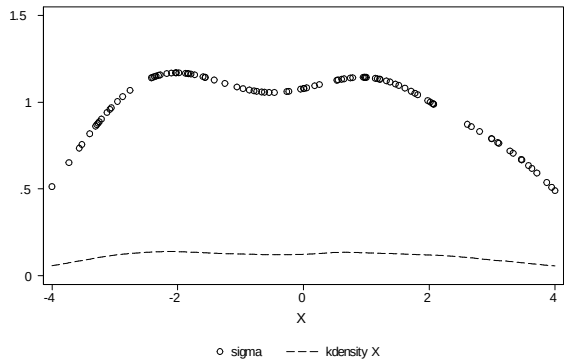
Criterion	*Estimate*	*CEL*	*LWCEL*
Bias	$\widehat{\beta}_1$	-9.100×10^{-2}	-8.619×10^{-2}
	$\widehat{\beta}_2$	1.436×10^{-2}	1.471×10^{-2}
	$\widehat{\beta}_3$	1.050×10^{-2}	9.416×10^{-3}
Variance	$\widehat{\beta}_1$	8.297×10^{-3}	6.189×10^{-3}
	$\widehat{\beta}_2$	2.474×10^{-3}	2.351×10^{-3}
	$\widehat{\beta}_3$	4.202×10^{-4}	3.916×10^{-4}
MSE	$\widehat{\beta}_1$	1.652×10^{-2}	1.362×10^{-2}
	$\widehat{\beta}_2$	2.681×10^{-3}	2.568×10^{-3}
	$\widehat{\beta}_3$	5.304×10^{-4}	4.802×10^{-4}

Both estimators performed relatively well under the simulation scenario which can be attributed to the relatively well-behaved nature of the data. Nonetheless, the $\widehat{\theta}_{LWCEL}$ improved on the $\widehat{\theta}_{CEL}$ in all cases, barring one bias term. The values of σ_i were also retained as an interesting byproduct of the $\widehat{\theta}_{LWCEL}$ estimation procedure, weighting individual local conditional empirical log likelihoods. Naturally, their magnitude follows the density of the data juxtaposed against σ_i in Figure 3.2:

CONCLUSION

In this Chapter we proposed a new class of estimators based directly on conditional moment restrictions that encompasses the entire GMC family. Moreover, using the GMC information-theoretic framework we showed that in constructing the estimators for the conditional moment restrictions previous literature implicitly use an arbitrary uniform weighting scheme. This lead to minimizing a discrepancy from a probability measure that is different from

Figure 3.2: Plot of sigma against x



the one under which the data was distributed. The reason for this phenomenon is that the previously analyzed estimators were based on local kernel smoothing of the unconditional statistical model. In contrast, we considered an information-theoretic dual locally weighted GMC optimization problem built directly on the conditional restrictions that minimizes a discrepancy from a probability measure according to which the data was distributed. Consequently, our newly proposed class of estimators includes locally weighted alternatives to the estimators analyzed in previous literature, in particular the Locally Weighted Conditional Empirical Likelihood (LWCEL). We analyzed the differences between the new LWCEL and the CEL of Kitamura, Tripathi, and Ahn (2004) in detail. In a Monte Carlo study we showed that the LWCEL estimator exhibits better finite-sample properties than found in the previous literature.

4.0 SIEVE-BASED EMPIRICAL LIKELIHOOD

Full title: **Sieve-based Empirical Likelihood under Semiparametric Conditional Moment Restrictions**

A general framework for analyzing economic data (Y, X) is to postulate conditional moment restrictions of the form

$$E\left[g\left(Z, \alpha_0\right)|X\right] = 0 \tag{4.1}$$

where $Z \equiv (Y', X_z')'$, Y is a vector of endogenous variables, X is a vector of conditioning variables (instruments), X_z is a subset of X, $g(\cdot)$ is a vector of functions known up a parameter α, and $F_{Y|X}$ is assumed unknown. The parameters of interest $\alpha_0 \equiv (\theta_0', h_0')'$ contain a vector of finite dimensional unknown parameters θ_0 and a vector of infinite dimensional unknown functions $h_0(\cdot) \equiv (h_{01}(\cdot), ..., h_{0q}(\cdot))'$. The inclusion of h_0 renders the condition (4.1) semiparametric, encompassing many important economic models. It includes for example the partially linear regression $g\left(Z, \alpha_0\right) = Y - X_1'\theta_0 - h_0(X_2)$ analyzed by Robinson (1988) and the index regression $g\left(Z, \alpha_0\right) = Y - h_0(X'\theta_0)$ studied by Powell, Stock, and Stoker (1989) and Ichimura (1993).

Recently, Kitamura, Tripathi, and Ahn (2004) (henceforth KTA) analyzed the Conditional Empirical Likelihood (CEL)[1] based on a parametric counterpart (without h_0) of (4.1)

$$E\left[g\left(Z, \theta_0\right)|X\right] = 0 \tag{4.2}$$

[1] A note on terminology: CEL is called "smoothed" and "sieve" empirical likelihood in KTA and Zhang and Gijbels (2003), respectively. Other types of smoothing have been introduced by Otsu (2003a) on moment restrictions in the quantile regression setting and hence KTA's original method is referred to as "conditional" empirical likelihood to avoid confusion. The CEL terminology was also adopted in Kitamura (2006).

The CEL estimator was shown to exhibit finite-sample properties superior to the Generalized Method of Moments. A conjecture that a similar type of result will also hold in the semiparametric scenario provided the intuitive basis for our analysis.

In this chapter we extend the LWCEL estimator analyzed in the previous chapter to the semiparametric environment defined by (4.1) proposing a new Sieve-based Locally Weighted Conditional Empirical Likelihood (SLWCEL) estimator. The SLWCEL can be viewed as a one-step information-theoretic alternative to the two-step Sieve Minimum Distance (SMD) estimator analyzed by Ai and Chen (2003) (henceforth AC). The SMD is based on a similar estimating principle as the GMM by first estimating a weighting matrix and then setting the weighted distance between vectors of moments close to zero. We approximate h with a sieve and estimate θ_0 and h_0 simultaneously with LWCEL. We establish consistency of the resulting one-step SLWCEL and asymptotic normality for its parametric component of θ.

A semiparametric extension of (4.2) to model (4.1) is unquestionably desirable because economic theories seldom produce exact functional forms, and misspecifications in functional forms may lead to inconsistent parameter estimates. By specifying the model partially (i.e. including h_0 as part of the unknown parameters), the inconsistency problem can be alleviated. In general, semiparametric literature related to the model (4.1) has been growing rapidly (see e.g. Powell, 1994; Pagan and Ullah, 1999, for reviews). Most of the available results are derived using a plug-in procedure: first h_0 is estimated nonparametrically by $\widehat{h}$ and then θ_0 is estimated using a parametric method (e.g. GMM or GEL) with h_0 replaced by $\widehat{h}$. However, such plug-in estimators are not capable of handling models where the unknown functions h_0 depend on the endogenous variables Y, because in such models θ_0 affects h_0 as well. Thus, in models where h_0 depends on an endogenous regressor, h_0 and θ_0 need to be estimated simultaneously. There are very few results concerning simultaneous estimators. Earlier applications include a semiparametric censored regression estimator (Duncan, 1986) and a semi-nonparametric maximum likelihood estimator (Gallant and Nychka, 1987).

However, a general estimation method for the model (4.1) that permits dependence of h_0 on Y and θ_0 was not well analyzed until a recent work by AC. These authors proposed a Sieve Minimum Distance (SMD) estimator of α_0 under (4.1), based on identification and consistency conditions derived by Newey and Powell (2003). Subsequent applications of the

SMD estimator include Chen and Ludvigson (2006) in a habit-based asset pricing model (with unknown functional form of the habit) testing various hypotheses on stock return data, Blundell et al. (2006) in a dynamic optimization model describing the allocation of total non-durable consumption expenditure, and Ai et al. (2006) investigating co-movement of commodity prices.

The first analysis that ventured into the realm of GEL-type estimators subject to conditional moment restrictions containing unknown functions is due to Otsu (2003b).[2] His shrinkage-type estimator is based on a penalized empirical log-likelihood ratio (PELR) which utilizes a penalty function $J(h)$ confining the minimization problem to a parameter space specified by the researcher. Usually, $J(h)$ is used to control some physical plausibility of h such as roughness of h. Otsu (2003b)'s penalized likelihood method differs from sieve analysis and hence his treatment of asymptotics differs from ours.[3]

Otsu (2003b) suggests (in Remark 2.2) that it is also possible to use a deterministic sieve approximations, instead of the penalty function approach, resulting in a deterministic sieve empirical likelihood estimator (DSELE) that would also be, under suitable conditions, [first-order] asymptotically equivalent to the SMD of AC. Similar conjecture has been raised in Nishiyama et al. (2005) who noted the lack of theoretical justification for such procedure. Chen (2005, footnote 39) made the same type of conjecture in relation to the conditional parametric Euclidean empirical likelihood estimator of Antoine, Bonnal, and Renault (2006) (henceforth ABR). However, despite calls for a theoretical justification of such procedures, no previous paper has performed the necessary theoretical analysis. Yet, in analogy to the parametric literature described above, developing a one-step simultaneous GEL-type sieve alternative to the two-step simultaneous SMD in the semiparametric case can lead to a similar type of improvement in terms of bias and higher-order efficiency and is therefore of great theoretical and practical interest.

All of the simultaneous estimators mentioned above are based on the method of sieves

[2] Up to date, the author has not been able to obtain a full copy of this paper. Only a google-cached html version containing parts of the paper's text is currently publicly available.

[3] In the seminal paper by Shen (1997), penalized likelihood and the method of sieves are treated as two separate concepts. To achieve asymptotic normality, Otsu (2003b) extends Theorem 2 of Shen (1997), whereas we extend Theorem 1 of Shen (1997) which is a separate result derived under different conditions from the former.

(Grenander, 1981; Chen, 2005) where h_0 is estimated over a compact subspace that is dense in the full parameter space as sample size increases. This feature of sieves conveniently simplifies the infinite-dimensional model h_0 to its finite-dimensional counterpart suitable for estimation. Here we also adhere to the sieve methodology. However, the currently available relevant general theory papers dealing with sieve M-estimation (Wong and Severini, 1991; Shen and Wong, 1994; Shen, 1997; Chen and Shen, 1998) consider only one set of exogenous variables without endogenous regressors and hence we can not apply these results directly in our case. Therefore, in the asymptotic analysis we combine them with several results of AC and our own new results necessitated by the specific nature of SLWCEL under (4.1). In particular, among other issues we derive an extension of Shen (1997) theorem on asymptotic normality of general simultaneous sieve estimators for the case of endogenous regressors under strong conditions and then apply it to the SLWCEL case under weak primitive conditions.

SIEVE-BASED CONDITIONAL EMPIRICAL LIKELIHOOD

In this chapter we will use series estimation (see e.g. Newey, 1997) as a particular form of linear sieves in both approximating h and determining the weights w_{ij}. Series estimators are known to contain functional bases that are superior in terms of MSE criteria to fixed-bandwidth kernel estimators, especially in the presence of spatial inhomogeneities in the data (see e.g. Ramsey, 1999). Silverman (1984) showed that series estimators with spline basis functions behave approximately like the variable-bandwidth kernel estimator which improves on fixed-bandwidth kernels in terms of MSE by the virtue of local adaptation. Another advantage of working with the LWCEL estimator based on series approximation is that truncation arguments in regions with small data density are not required in contrast to kernel weights.

The environment setup parallels the one of Newey and Powell (2003) and AC. Suppose that the observations $\{(Y_i, X_i) : i = 1, ..., n\}$ are drawn independently from the distribution of (Y, X) with support $\mathcal{Y} \times \mathcal{X}$, where $\mathcal{Y}$ is a subset of $\mathbb{R}^{d_Y}$ and $\mathcal{X}$ is a compact subset of $\mathbb{R}^{d_X}$. Suppose that the unknown distribution of (Y, X) satisfies the semiparametric conditional

moment restrictions given by (4.1), where $g : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}^{d_g}$ is a known mapping, up to an unknown vector of parameters, $\alpha_0 \equiv (\theta_0', h_0')' \in \mathcal{A} \equiv \Theta \times \mathcal{H}$, and $Z \equiv (Y', X_z')' \in \mathcal{Y} \times \mathcal{X}_Z \equiv \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ where $\mathcal{X}_Z \subseteq \mathcal{X}$. We assume that $\Theta \subseteq \mathbb{R}^{d_\theta}$ is compact with non-empty interior and that $\mathcal{H} \equiv \mathcal{H}^1 \times ... \times \mathcal{H}^{d_h}$ is a space of continuous functions. Since $\mathcal{H}$ is infinite-dimensional, in constructing a feasible estimator we follow the sieve literature (Grenander, 1981; Chen, 2005) by replacing $\mathcal{H}$ with a sieve space $\mathcal{H}_n \equiv \mathcal{H}_n^1 \times ... \times \mathcal{H}_n^{d_h}$ which is a computable and finite-dimensional compact parameter space that becomes dense in $\mathcal{H}$ as n increases.

Next, we introduce the series estimator used in the analysis (Newey, 1997), AC. For each $l = 1, ..., d_g$, and for a given α, let $\{p_{0j}(X),\ j = 1, 2, ...k_n\}$ denote a sequence of known basis functions (power series, splines, wavelets, etc.) and let $p^{k_n}(X) \equiv (p_{01}(X), ..., p_{0k_n}(X))'$. Let further $p^{k_n}(X)$ be a tensor-product linear sieve basis, which is a product of univariate sieves over d_X (for details see AC). Let $P = (p^{k_n}(x_1), ..., p^{k_n}(x_n))'$ be an $(n \times k_n)$ matrix. Consider the model (4.1) and denote the conditional mean function

$$\begin{aligned} m(X, \alpha) &\equiv E\left[g\left(Z, \alpha\right)|X\right] \\ &= \int g\left(Z, \alpha\right) dF_{Y|X} \end{aligned} \tag{4.3}$$

Let $\widehat{m}(X, \alpha) \equiv (\widehat{m}_1(X, \alpha), ..., \widehat{m}_{d_\rho}(X, \alpha))'$. A consistent nonparametric linear sieve estimator of $m_l(X, \alpha)$ is given by

$$\widehat{m}_l(X, \alpha) = p^{k_n}(X)'\widehat{\kappa}_l$$

where h in $\alpha = (\theta', h')'$ is restricted to the sieve space $\mathcal{H}_n$ and $\widehat{\kappa}_l$ is an OLS estimate obtained by regressing $g_l\left(Y, X_z, \alpha\right)$ on $p^{k_n}(X)$,

$$\begin{aligned} \widehat{\kappa}_l &= \left(P'P\right)^{-1} P' g_l\left(Z, \alpha\right) \\ &= \sum_{j=1}^{n} p^{k_n}(x_j)' \left(P'P\right)^{-1} g_l\left(z_j, \alpha\right) \end{aligned} \tag{4.4}$$

and hence

$$\begin{aligned}\widehat{m}_l(x_i,\alpha) &= \widehat{E}_{Z|X}\left[g_l\left(Z,\alpha\right)|X=x_i\right] \\ &= p^{k_n}(x_i)'\widehat{\kappa}_l \\ &= \sum_{j=1}^{n} p^{k_n}(x_j)'\left(P'P\right)^{-1}p^{k_n}(x_i)g_l\left(z_j,\alpha\right) \\ &= \sum_{j=1}^{n} w_{ij}g_l\left(z_j,\alpha\right)\end{aligned}$$

after substituting from (4.4), $l=\{1,...,d_g\}$. In the vector form

$$\widehat{m}(x_i,\alpha)=\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)$$

The weights are given by

$$w_{ij}=p^{k_n}(x_j)'\left(P'P\right)^{-1}p^{k_n}(x_i) \tag{4.5}$$

and

$$\begin{aligned}\sigma_i &= \sum_{j=1}^{n} w_{ij} \\ &= \sum_{j=1}^{n} p^{k_n}(x_j)'\left(P'P\right)^{-1}p^{k_n}(x_i) \\ &= \mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i)\end{aligned}$$

where $\mathbf{i}$ is a $(n\times 1)-$vector of ones.

We now turn to the derivation of LWCEL under (4.1). The Lagrangian[4] for the local semiparametric EL estimator is

$$\max_{\pi_{ij}}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\ln\pi_{ij}\quad\text{s.t.}\quad \pi_{ij}\geq 0,\ \sum_{j=1}^{n}\pi_{ij}=1,\ \sum_{j=1}^{n} g\left(z_j,\alpha_n\right)\pi_{ij}=0,\ \text{for } i,j=1,...,n$$

where α_n is α restricted to the sieve space $\mathcal{A}_n$. Then,

$$\widehat{\pi}_{ij}=\frac{w_{ij}}{\sigma_i+\lambda_i' g\left(z_j,\alpha_n\right)} \tag{4.6}$$

[4] As discussed above, omission of q_{ij} from the denominator of $\ln\left(\pi_{ij}/q_{ij}\right)$ is inconsequential in the case of LWCEL.

and for each $\alpha_n \in \mathcal{A}_n$, λ_i solves

$$\sum_{j=1}^{n} \frac{w_{ij} g\left(z_j, \alpha_n\right)}{\sigma_i + \lambda_i' g\left(z_j, \alpha_n\right)} = 0 \tag{4.7}$$

The Sieve-based Locally Weighted Conditional Empirical Likelihood (SLWCEL) evaluated at α_n is defined as

$$L_{SLWCEL}(\alpha_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \ln \left\{ \frac{w_{ij}}{\sigma_i + \lambda_i' g\left(z_j, \alpha_n\right)} \right\}$$

where λ_i solves (4.7). The estimator of α_0 is defined as

$$\widehat{\alpha}_n = \arg \max_{\alpha_n \in \mathcal{A}_n} L_{SLWCEL}(\alpha_n) \tag{4.8}$$

Solving (4.8) is equivalent to solving

$$\widehat{\alpha}_n = \arg \max_{\alpha_n \in \mathcal{A}_n} G_n(\alpha_n) \tag{4.9}$$

where

$$G_n(\alpha_n) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \ln \left\{ \sigma_i + \lambda_i' g\left(z_j, \alpha_n\right) \right\} \tag{4.10}$$

Implementing our estimator is straightforward. One advantage of the sieve approach is that once $h \in \mathcal{H}$ is replaced by $h_n \in \mathcal{H}_n$, the estimation problem effectively becomes a parametric one. Commonly used statistical and econometric packages can then be used to compute the estimate. From (4.7) it follows that

$$\lambda_i = \arg \max_{\rho \in \mathbb{R}^{d_g}} \sum_{j=1}^{n} w_{ij} \ln \left\{ \sigma_i + \rho' g\left(z_j, \alpha_n\right) \right\} \tag{4.11}$$

This is a well-behaved optimization problem since the objective function is globally concave and can be solved by a Newton-Raphson numerical procedure. The outer loop (4.9) can be carried out using a numerical optimization procedure. For a relevant discussion of computational issues, see for example Kitamura (2006, section 8.1).

CONSISTENCY

In this section we present some asymptotic results for the smoothed empirical likelihood estimator as defined in (4.8). The general approach follows closely the one developed in KTA. The following definitions, adopted from AC, are introduced:

Definition 1. *A real-valued measurable function $g(Z,\alpha)$ is Hölder continuous in $\alpha \in \mathcal{A}$ if there exist a constant $\overline{\kappa} \in (0,1]$ and a measurable function $c_2(Z)$ with $E\left[c_2(Z)^2|X\right]$ bounded, such that $|g(Z,\alpha_1) - g(Z,\alpha_2)| \leq c_2(Z)\left\|\alpha_1 - \alpha_2\right\|^{\overline{\kappa}}$ for all $Z \in \mathcal{Z}$, $\alpha_1, \alpha_2 \in \mathcal{A}$.*

The Hölder space of smooth functions $\Lambda^{\overline{\gamma}}(\mathcal{X})$ of order $\overline{\gamma} > 0$ and the corresponding Hölder ball $\Lambda_c^{\overline{\gamma}}(\mathcal{X}) \equiv \{g \in \Lambda^{\overline{\gamma}}(\mathcal{X}) : \|g\|_{\Lambda^{\overline{\gamma}}} \leq c < \infty\}$ with radius c are defined in AC, p. 1800.

Definition 2. *A real-valued measurable function $g(Z,\alpha)$ satisfies an envelope condition over $\alpha \in \mathcal{A}$ if there exists a measurable function $c_1(Z)$ with $E\left\{c_1(Z)^4\right\} < \infty$ such that $|g(Z,\alpha)| \leq c_1(Z)$ for all $Z \in \mathcal{Z}$ and $\alpha \in \mathcal{A}$.*

The following Assumptions are made to facilitate the analysis:

Assumption 4.1. *For each $\alpha \neq \alpha_0$ there exists a set $\mathcal{X}_\alpha$ such that $\Pr\left\{x \in \mathcal{X}_\alpha\right\}$, and $E\left[g\left(z,\alpha\right)|x\right] \neq 0$ for every $x \in \mathcal{X}_\alpha$.*

Assumption 4.2. *(i) The data $\{(Y_i, X_i)_{i=1}^n\}$ are i.i.d.; (ii) $\mathcal{X}$ is compact with nonempty interior; (iii) the density of X is bounded and bounded away from zero.*

Assumption 4.3. *(i) The smallest and the largest eigenvalues of $E\left[p^{k_n}(X) \times p^{k_n}(X)'\right]$ are bounded and bounded away from zero for all k_n; (ii) for any $g\left(\cdot\right)$ with $E\left[g(X)^2\right] < \infty$, there exists $p^{k_n}(X)'\kappa$ such that $E\left[\left\{g(X) - p^{k_n}(X)'\kappa\right\}^2\right] = o(1)$.*

Assumption 4.4. *(i) There is a metric $\|\cdot\|$ such that $\mathcal{A} \equiv \Theta \times \mathcal{H}$ is compact under $\|\cdot\|$; (ii) for any $\alpha \in \mathcal{A}$, there exists $\Pi_n\alpha \in \mathcal{A}_n \equiv \Theta \times \mathcal{H}_n$ such that $\|\Pi_n\alpha - \alpha\| = o(1)$.*

Assumption 4.5. *(i) $E\left[\left|g\left(Z,\alpha_0\right)\right|^2|X\right]$ is bounded; (ii) $g\left(Z,\alpha\right)$ is Hölder continuous in $\alpha \in \mathcal{A}$.*

Let $k_{1n} \equiv \dim(\mathcal{H}_n)$ denote the number of unknown sieve parameters in $h_n \in \mathcal{H}_n$.

Assumption 4.6. *$k_{1n} \to \infty$, $k_n \to \infty$, $k_n/n \to 0$ and $d_g k_n \geq d_\theta + k_{1n}$.*

Assumption 4.7. $E\|x\|^{1+\varrho} < \infty$ *for some* $\varrho < \infty$.

Assumption 4.8. $E\{\sup_{\alpha\in\mathcal{A}} \|g(Z,\alpha)\|^m\} < \infty$ *for some* $m \geq 8$.

Assumption 4.1 is Assumption 3.1 in KTA that guarantees identification of θ_0. Assumptions 4.2–4.6 are essentially the same conditions imposed in Newey and Powell (2003) and AC. Assumption 4.2 rules out time series observations. Assumptions 4.3–4.6 are typical conditions imposed for series (or linear sieve) estimation of conditional mean functions. Assumption 4.4(i) restricts the parameter space as well as the choice of the metric $\|\cdot\|$. It is a commonly imposed condition in the semiparametric econometrics literature, and is satisfied when the infinite-dimensional parameter space $\mathcal{H}$ consists of bounded and smooth functions (see Gallant and Nychka, 1987). Assumption 4.4(ii) is the definition of a sieve space. Assumption 4.5 is typically imposed on the residual function in the literature on parametric nonlinear estimation. Assumption 4.6 restricts the growth rate of the number of basis functions in the series approximation. Assumption 4.7 is Assumption 3.4(ii) in KTA, used in Lemma A.1. Assumption 4.8 is Assumption 3.2 in KTA used in Lemma A8.

The following Theorem provides a consistency result:

Theorem 4.1. *Let the Assumptions 4.1–4.7 hold. Then* $\|\widehat{\alpha}_n - \alpha_0\| = o_p(1)$.

The proof is derived in the Appendix. The proof proceeds along the lines of KTA. However, the fact that the sieve parameter space $\mathcal{H}_n$ grows dense in an infinite-dimensional space $\mathcal{H}$ now needs to be addressed. The inclusion of σ_i in the LWCEL objective function compared to KTA's CEL also complicates matters. We achieve some simplifications arising from not having to make use of truncation arguments for kernels. Since we are not dealing with kernels, unlike KTA we can not use Lemma B.1 of Ai (1997) to determine uniform convergence rates. For this purpose, we specialize Lemma A.1(A) of AC, derived for the combined space $\mathcal{X}\times\mathcal{A}$, to the space $\mathcal{X}$ only, with $g(z_j,\alpha)$ evaluated at a given fixed α. Since we do not have to account for growth restrictions on the parameter space in this Lemma, we are able to obtain faster convergence rate $\widetilde{\delta}_{1n}$ than AC.

CONVERGENCE RATES

Theorem 4.1 established consistency of $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{h}_n)$ under a general metric $\|\cdot\|$ constrained only by Assumption 4.4(i). In order to ascertain asymptotic normality of $\widehat{\theta}_n$, one typically needs that $\widehat{\alpha}_n$ converge to α_0 at a rate faster than $n^{-1/4}$ (see e.g. Newey, 1994). As noted by Newey and Powell (2003), for model (4.1) where the unknown h_0 can depend on endogenous variables Y, it is generally difficult to obtain fast convergence rate under $\|\cdot\|$. Nonetheless, as demonstrated by AC, in simultaneous estimation of $(\widehat{\theta}_n, \widehat{h}_n)$ it is sufficient to show fast convergence rate of $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{h}_n)$ for only a special case of $\|\cdot\|$ to derive asymptotic normality of $\widehat{\theta}_n$. Naturally, we will also follow this approach. However, since the objective function of the problem analyzed in AC is different from ours, our metric also differs. While AC used a quadratic form type metric, we perform the analysis under the Fisher metric $\|\cdot\|_F$ which is the natural choice for a likelihood-based scenario.

Some additional notation is necessary to introduce the Fisher metric. The properties of $\mathcal{A}$ and the notation for pathwise derivatives established in this paragraph borrows from AC. Suppose the parameter space $\mathcal{A}$ is connected in the sense that for any two points α_1, $\alpha_2 \in \mathcal{A}$ there exists a continuous path $\{\alpha(t) : t \in [0,1]\}$ in $\mathcal{A}$ such that $\alpha(0) = \alpha_1$ and $\alpha(1) = \alpha_2$. Also, suppose that $\mathcal{A}$ is convex at the true value α_0 in the sense that, for any $\alpha \in \mathcal{A}$, $(1-t)\alpha_0 + t\alpha \in$ for small $t > 0$. Furthermore, suppose that for almost all Z, $g(Z, (1-t)\alpha_0 + t\alpha)$ is continuously differentiable at $t = 0$. Denote the first pathwise derivative at the direction $[\alpha - \alpha_0]$ evaluated at α_0 by

$$\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha - \alpha_0] \equiv \left.\frac{dg(Z,(1-t)\alpha_0 + t\alpha)}{dt}\right|_{t=0} \quad \text{a.s. } Z$$

and for any α_1, $\alpha_2 \in \mathcal{A}$ denote

$$\begin{aligned} \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] &\equiv \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_0] - \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_2 - \alpha_0] \\ \frac{dm(X,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] &\equiv E\left\{ \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] \middle| X \right\} \end{aligned} \tag{4.12}$$

Furthermore, let

$$\varphi(X, Z, \alpha) \equiv \ln\{\sigma_x + \lambda'(X,\alpha) g(Z,\alpha)\} \tag{4.13}$$

$$\psi(X,\alpha) \equiv E[\varphi(X,Z,\alpha)|X] \tag{4.14}$$

where σ_x stands for σ_i evaluated at a generic $X = x$. For any α_1, $\alpha_2 \in \mathcal{A}$ the Fisher norm $\|\cdot\|_F$ (see e.g. Wong and Severini, 1991, p. 607) is defined[5] as

$$\|\alpha_1 - \alpha_2\|_F = \sqrt{E\left\{E\left[\left(\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\right)' \frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\middle| X\right]\right\}} \quad (4.15)$$

Let $\overline{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A} - \{\alpha_0\}$ under the metric $\|\cdot\|_F$. Then $\left(\overline{\mathbf{V}}, \|\cdot\|_F\right)$ is a Hilbert space with the inner product

$$\langle v_1, v_2\rangle_F = \|v_1 - v_2\|_F^2$$

We will now show that our metric $\|\alpha_1 - \alpha_2\|_F$ is equivalent to a *conditional version* of the metric used in AC. Let

$$\begin{aligned} s(X,Z,\alpha) &\equiv \lambda'(\alpha, X) g(Z,\alpha) \\ \varpi(X,Z,\alpha) &\equiv \frac{d\varphi(X,Z,\alpha_0)}{ds(X,Z,\alpha)} \\ &= \frac{1}{\sigma_x + s(X,Z,\alpha)} \end{aligned}$$

where $s(X,Z,\alpha)$ and $\varpi(X,Z,\alpha)$ is scalars. Note that from the conditional moment restriction (4.1), under the expectation taken over Z conditional on X

$$\lambda(X,\alpha_0) = 0 \quad (4.16)$$

which means that the constraints on $F_{Y|X}$ imposed by (4.1) are satisfied with equality and the Lagrange multiplier $\lambda(X,\alpha_0)$ takes on the value 0. This is also apparent from Lemma A.8. We have

$$\begin{aligned} & E\left[\left(\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\right)' \frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\middle| X\right] \\ = \; & E\left[\varpi(X,Z,\alpha_0)^2 \left(\frac{ds(X,Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\right)' \frac{ds(X,Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\middle| X\right] \\ = \; & E\left[\begin{array}{c} \varpi(X,Z,\alpha_0)^2 \left(\lambda'(X,\alpha_0)\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] + g(Z,\alpha_0)\frac{d\lambda'(X,\alpha_0)}{d\alpha}\right)' \\ \times \left(\lambda'(X,\alpha_0)\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] + g(Z,\alpha_0)\frac{d\lambda'(X,\alpha_0)}{d\alpha}\right) \end{array}\middle| X\right] \\ = \; & A_1 + A_2 + A_3 + A_4 \end{aligned} \quad (4.17)$$

[5] We use the inner product notation for the pathwise derivatives to explicitly account for the special case when $\alpha \equiv \theta \in \mathbb{R}^{d_\theta}$.

where

$$
\begin{aligned}
A_1 &= E\left[\varpi(X,Z,\alpha_0)^2\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'\lambda(X,\alpha_0)\lambda'(X,\alpha_0)\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \\
A_2 &= E\left[\varpi(X,Z,\alpha_0)^2\left(\frac{d\lambda(X,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'g(Z,\alpha_0)\lambda'(X,\alpha_0)\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \\
A_3 &= E\left[\varpi(X,Z,\alpha_0)^2\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'\lambda'(X,\alpha_0)g(Z,\alpha_0)\frac{d\lambda'(X,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \\
A_4 &= E\left[\varpi(X,Z,\alpha_0)^2\left(\frac{d\lambda(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'g(Z,\alpha_0)g'(Z,\alpha_0)\frac{d\lambda'(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \quad (4.18)
\end{aligned}
$$

Using (4.16) yields $A_1 = A_2 = A_3 = 0$. By the definition of $\lambda(X,\alpha)$ in (4.11), $\lambda(X,\alpha)$ is a function of $g(Z,\alpha)$ which is a function of α. Moreover, $\lambda(X,\alpha)$ is a function of α *only* via $g(Z,\alpha)$. Hence, under the expectation taken over Z conditional on X

$$\frac{d\lambda(X,\alpha)}{d\alpha}[\alpha_1-\alpha_2] = \frac{d\lambda(X,\alpha)}{dg(Z,\alpha)}\frac{dg(Z,\alpha)}{d\alpha}[\alpha_1-\alpha_2] \quad (4.19)$$

In particular, under the expectation over Z conditional on X, $\lambda(X,\alpha)$ is defined implicitly as a function of $g(Z,\alpha)$ by the relation

$$F(\lambda,g) = E\left[\frac{g(Z,\alpha)}{\sigma_x+\lambda'(X,\alpha)g(Z,\alpha)}\middle| X\right] = 0$$

By the Implicit Function Theorem

$$
\begin{aligned}
\frac{d\lambda(X,\alpha)}{dg(Z,\alpha)} &= \frac{\partial F(\lambda,g)/\partial g(Z,\alpha)}{\partial F(\lambda,g)/\partial\lambda(X,\alpha)} \\
&= E\left[\frac{(\sigma_x+\lambda'(X,\alpha)g(Z,\alpha)-\lambda'(X,\alpha)g(Z,\alpha))/(\sigma_x+\lambda'(X,\alpha)g(Z,\alpha))^2}{-g(Z,\alpha)g'(Z,\alpha)/(\sigma_x+\lambda'(X,\alpha)g(Z,\alpha))^2}\middle| X\right] \\
&= -\sigma_x\left\{E\left[g(Z,\alpha)g'(Z,\alpha)\middle| X\right]\right\}^{-1} \\
&= -\sigma_x\Sigma(X,\alpha)^{-1} \quad (4.20)
\end{aligned}
$$

Substituting (4.20) into (4.19) we obtain

$$\frac{d\lambda(\alpha,X,Z)}{d\alpha}[\alpha_1-\alpha_2] = -\sigma_x\Sigma(X,\alpha)^{-1}\frac{dg(Z,\alpha)}{d\alpha}[\alpha_1-\alpha_2] \quad (4.21)$$

Substituting (4.21) into (4.18) yields

$$A_4 = \sigma_x^2 E\left[\varpi(X,Z,\alpha_0)^2\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)'W_0(X,Z)^{-1}\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right]$$

where

$$W_0(X,Z)^{-1} \equiv \Sigma(X,\alpha_0)^{-1} g(Z,\alpha_0) g'(Z,\alpha_0) \Sigma(X,\alpha_0)^{-1}$$

Using (4.16) in $\varpi(X,Z,\alpha_0)$ results in

$$A_4 = E\left[\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)' W_0(X,Z)^{-1} \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \tag{4.22}$$

Substituting (4.22) into (4.18) and (4.15) yields

$$\|\alpha_1-\alpha_2\|_F = \sqrt{E\left\{E\left[\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)' W_0(X,Z)^{-1} \frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right]\right\}} \tag{4.23}$$

The expression (4.23) can be viewed as a conditional version of the metric used in AC. In particular, if $\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]$ and $g(Z,\alpha_0)$ are independent conditional on X, then (4.23) reduces to $\sqrt{E\left\{\left(\frac{dm(X,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)' \Sigma(X,\alpha_0)^{-1} \frac{dm(X,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right\}}$ which is the metric used in AC with the efficient weighting matrix.

Note that by (4.16)

$$\begin{aligned} E\left[\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] &= \lambda'(X,\alpha_0) E\left[\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] \\ &\quad + \frac{d\lambda'(X,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2] E\left[g(Z,\alpha_0)\middle| X\right] \\ &= 0 \end{aligned}$$

and hence

$$E\left[\left(\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)' \frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right] = Var\left(\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right)$$

implying

$$\begin{aligned} \|\alpha_1-\alpha_2\|_F &= \sqrt{E\left\{Var\left(\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\middle| X\right)\right\}} \\ \langle v,v\rangle_F &= E\left\{Var\left(\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[v]\middle| X\right)\right\} \end{aligned}$$

We will now introduce the conditions under which the desired convergence rates are derived.

Assumption 5.1. *(i) $\mathcal{A}$ is convex in α_0, and $g(Z,\alpha)$ is pathwise differentiable at α_0; (ii) for some $c_1, c_2 > 0$,*

$$\begin{aligned} c_1 E\left\{m(X,\alpha_n)' W_0(X)^{-1} m(X,\alpha_n)\right\} &\leq \|\alpha_n - \alpha_0\|_F^2 \\ &\leq c_2 E\left\{m(X,\alpha_n)' W_0(X)^{-1} m(X,\alpha_n)\right\} \end{aligned}$$

holds for all $\alpha_n \in \mathcal{A}_n$ with $\|\alpha_n - \alpha_0\| = o(1)$.

Assumption 5.2. *For any $\widetilde{g}(\cdot)$ in $\Lambda_c^{\overline{\gamma}}(\mathcal{X})$ with $\overline{\gamma} > d_x/2$, there exists $p^{k_n}(\cdot)'\kappa \in \Lambda_c^{\overline{\gamma}}(\mathcal{X})$ such that $\sup_{X\in\mathcal{X}} \left|\widetilde{g}(X) - p^{k_n}(X)'\kappa\right| = O(k_n^{-\overline{\gamma}/d_x})$, and $k_n^{-\overline{\gamma}/d_x} = o(n^{-1/4})$.*

Assumption 5.3. *(i) Each element of $g(Z,\alpha)$ satisfies an envelope condition in $\alpha_n \in \mathcal{A}_n$; (ii) each element of $m(X,\alpha) \in \Lambda_c^{\overline{\gamma}}(\mathcal{X})$ with $\overline{\gamma} > d_x/2$, for all $\alpha_n \in \mathcal{A}_n$.*

In line with AC, let $\xi_{0n} \equiv \sup_{X\in\mathcal{X}} \left\|p^{k_n}(X)\right\|_E$, which is nondecreasing in k_n. Denote $N(\delta, \mathcal{A}_n, \|\cdot\|)$ as the minimal number of radius δ covering balls of $\mathcal{A}_n$ under the $\|\cdot\|$ metric.

Assumption 5.4. $k_{1n} \times \ln n \times \xi_{0n}^2 \times n^{-1/2} = o(1)$.

Assumption 5.5. $\ln\left[N(\varepsilon^{1/\kappa}, \mathcal{A}_n, \|\cdot\|)\right] \leq const. \times k_{1n} \times \ln(k_{1n}/\varepsilon)$.

Assumption 5.6. $\Sigma_0(X) \equiv Var\left[g(Z,\alpha_0)|X\right]$ *is positive definite for all* $X \in \mathcal{X}$.

The following result gives the convergence rate of the SLWCEL estimator under the Fisher metric. The proof is provided in the Appendix.

Theorem 5.1. *Under Assumptions 4.1 - 5.6, we have* $\left\|\widehat{\alpha}_n - \alpha_0\right\|_F = o_p(n^{-1/4})$.

ASYMPTOTIC NORMALITY

To derive the asymptotic distribution of $\widehat{\theta}_n$, it suffices to derive the asymptotic distribution of $f\left(\widehat{\alpha}_n\right) \equiv \tau'\widehat{\theta}_n$ for any fixed non-zero $\tau \in R^{d_\theta}$. The difference $f\left(\widehat{\alpha}_n\right) - f\left(\alpha_0\right)$ is linked to the pathwise directional derivatives of the sample criterion function via the inner product involving a Riesz representor v^*. Application of a Central Limit Theorem for triangular arrays of functions indexed by a finite-dimensional parameter then shows the desired result.

In this Section we introduce the necessary notation, compute the Riesz representor v^* and state the Theorem of $\sqrt{n}$-normality of $\widehat{\theta}_n$.

Since $f(\alpha) \equiv \tau'\theta$ is a linear functional on $\overline{\mathbf{V}}$, it is bounded (i.e. continuous) if and only if

$$\sup_{0 \neq \alpha - \alpha_0 \in \overline{\mathbf{V}}} \frac{|f(\alpha) - f(\alpha_0)|}{\|\alpha - \alpha_0\|_F} < \infty$$

The Riesz Representation Theorem states that there exists a representor $v^* \in \overline{\mathbf{V}}$ satisfying

$$\|v^*\|_F \equiv \sup_{0 \neq \alpha - \alpha_0 \in \overline{\mathbf{V}}} \frac{|f(\alpha) - f(\alpha_0)|}{\|\alpha - \alpha_0\|_F} \tag{4.24}$$

and

$$f(\alpha) = f(\alpha_0) + \langle v^*, \alpha - \alpha_0 \rangle_F$$

Hence,

$$f(\widehat{\alpha}_n) - f(\alpha_0) = \langle v^*, \widehat{\alpha}_n - \alpha_0 \rangle_F$$

Let

$$\frac{dg(Z, \alpha_0)}{d\alpha}[\alpha - \alpha_0] \equiv \frac{dg(Z, \alpha_0)}{d\theta'}(\theta - \theta_0) + \frac{dg(Z, \alpha_0)}{dh}[h - h_0] \tag{4.25}$$

For any $h \in \overline{\mathcal{H}}$, there exists $w_j(\cdot) \in \overline{\mathcal{W}}$ for $j = 1, ..., d_\theta$ such that

$$h - h_0 = -(w_1, ..., w_{d_\theta})(\theta - \theta_0) = -w(\theta - \theta_0)$$

Define

$$\begin{aligned} \frac{dg(Z, \alpha_0)}{dh}[w] &\equiv \left(\frac{dg(Z, \alpha_0)}{dh}[w_1], ..., \frac{dg(Z, \alpha_0)}{dh}[w_{d_\theta}] \right) \\ D_w(Z) &\equiv \frac{dg(Z, \alpha_0)}{d\theta'} - \frac{dg(Z, \alpha_0)}{dh}[w] \end{aligned} \tag{4.26}$$

where $D_w(Z)$ is a $d_g \times d_\theta-$matrix valued function. Definitions (4.25) and (4.26) imply

$$\frac{dg(Z, \alpha_0)}{dh}[h - h_0] = -\frac{dg(Z, \alpha_0)}{dh}[w](\theta - \theta_0)$$

and hence

$$\begin{aligned} D_w(Z)(\theta - \theta_0) &= \frac{dg(Z, \alpha_0)}{d\theta'}(\theta - \theta_0) - \frac{dg(Z, \alpha_0)}{dh}[w](\theta - \theta_0) \\ &= \frac{dg(Z, \alpha_0)}{d\theta'}(\theta - \theta_0) + \frac{dg(Z, \alpha_0)}{dh}[h - h_0] \\ &= \frac{dg(Z, \alpha_0)}{d\alpha}[\alpha - \alpha_0] \end{aligned} \tag{4.27}$$

By definition of $\|\cdot\|_F$ this implies

$$\begin{aligned}\|\alpha-\alpha_0\|_F^2 &= E\left\{E\left[\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha-\alpha_0]\right)' W_0(Z,X)^{-1}\left(\frac{dg(Z,\alpha_0)}{d\alpha}[\alpha-\alpha_0]\right)\middle| X\right]\right\} \\ &= E\left\{E\left[(\theta-\theta_0)' D_w(Z)' W_0(Z,X)^{-1} D_w(Z)(\theta-\theta_0)\middle| X\right]\right\} \qquad (4.28)\end{aligned}$$

Let $w^* = \left(w_1^*, ..., w_{d_\theta}^*\right)$ be the solution to

$$\inf_{w_j\in\overline{\mathcal{W}}, j=1,...,d_\theta} E\left\{E\left[(\theta-\theta_0)' D_w(Z)' W_0(Z,X)^{-1} D_w(Z)(\theta-\theta_0)\middle| X\right]\right\} \qquad (4.29)$$

where "inf" is in positive semidefinite matrix sense. Using the definitions of w^*, $f(\alpha)$, (4.24) and (4.28)

$$\begin{aligned}\|v^*\|_F^2 &\equiv \sup_{0\neq\alpha-\alpha_0\in\overline{\mathbf{V}}} \frac{|f(\alpha)-f(\alpha_0)|^2}{\|\alpha-\alpha_0\|_F^2} \\ &= \frac{(\theta-\theta_0)'\tau\tau'(\theta-\theta_0)}{(\theta-\theta_0)' E\left\{E\left[D_w(Z)' W_0(Z,X)^{-1} D_w(Z)\middle| X\right]\right\}(\theta-\theta_0)} \\ &= \tau'\left[E\left\{E\left[D_w(Z)' W_0(Z,X)^{-1} D_w(Z)\middle| X\right]\right\}\right]^{-1}\tau \qquad (4.30)\end{aligned}$$

where $v^* \equiv (v_\theta^*, v_h^*) \in \overline{\mathbf{V}}$. By the definition of w^*, $v_h^* = -w^* \times v_\theta^*$. From this and (4.27) we have

$$\frac{dg(Z,\alpha_0)}{d\alpha}[v^*] = D_{w^*}(Z) v_\theta^* \qquad (4.31)$$

Let

$$v_\theta^* = \left[E\left\{E\left[D_w(Z)' W_0(Z,X)^{-1} D_w(Z)\middle| X\right]\right\}\right]^{-1}\tau \qquad (4.32)$$

Substituting (4.32) into the definition of $\|\cdot\|_F^2$ in (4.15) via the expression for (4.31) yields

$$\begin{aligned}\|v^*\|_F^2 &= E\left\{E\left[\left(\frac{dg(Z,\alpha_0)}{d\alpha}[v^*]\right)' W_0(Z,X)^{-1}\left(\frac{dg(Z,\alpha_0)}{d\alpha}[v^*]\right)\middle| X\right]\right\} \\ &= E\left\{E\left[(D_{w^*}(Z)v_\theta^*)' W_0(Z,X)^{-1}(D_{w^*}(Z)v_\theta^*)\middle| X\right]\right\} \\ &= v_\theta^{*\prime} E\left\{E\left[D_{w^*}(Z)' W_0(Z,X)^{-1} D_{w^*}(Z)\middle| X\right]\right\} v_\theta^* \\ &= \tau'\left[E\left\{E\left[D_w(Z)' W_0(Z,X)^{-1} D_w(Z)\middle| X\right]\right\}\right]^{-1} \\ &\quad \times E\left\{E\left[D_{w^*}(Z)' W_0(Z,X)^{-1} D_{w^*}(Z)\middle| X\right]\right\} \\ &\quad \times\left[E\left\{E\left[D_w(Z)' W_0(Z,X)^{-1} D_w(Z)\middle| X\right]\right\}\right]^{-1}\tau \\ &= \tau'\left[E\left\{E\left[D_w(Z)' W_0(Z,X)^{-1} D_w(Z)\middle| X\right]\right\}\right]^{-1}\tau\end{aligned}$$

which matches (4.30) and thus validates (4.32) shown unique by the Riesz Representation Theorem.

The following additional conditions correspond to Assumptions 4.1-4.3 in AC and are sufficient for the $\sqrt{n}$-normality of $\widehat{\theta}_n$:

Assumption 6.1. *(i)* $E\left\{E\left[D_w(Z)'W_0(Z,X)^{-1}D_w(Z)\middle|X\right]\right\}$ *is positive definite; (ii)* $\theta_0 \in int(\Theta)$; *(iii)* $\Sigma_0(X) \equiv Var[g(Z,\alpha_0)|X]$ *is positive definite for all* $X \in \mathcal{X}$.

Assumption 6.2. *There is a* $v_n^* = (v_\theta^*, -\Pi_n w^* \times v_\theta^*) \in \mathcal{A}_n - \alpha_0$ *such that* $\|v_n^* - v^*\|_F = O(n^{-1/4})$.

Following AC, let $\mathcal{N}_{0n} \equiv \{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| = o(1),\ \|\alpha_n - \alpha_0\|_F = o(n^{-1/4})\}$ and define $\mathcal{N}_0$ the same way with $\mathcal{A}_n$ replaced by $\mathcal{A}$. Also, for any $v \in \overline{\mathbf{V}}$, denote

$$\frac{dg(Z,\alpha)}{d\alpha}[v] \equiv \left.\frac{dg(Z,\alpha+tv)}{dt}\right|_{t=0} \quad \text{a.s. } Z$$

and

$$\frac{dm(Z,\alpha)}{d\alpha}[v] \equiv E\left\{\frac{dg(Z,\alpha)}{d\alpha}[v] \;\middle|\; X\right\} \quad \text{a.s. } Z$$

Assumption 6.3. *For all* $\alpha \in \mathcal{N}_0$, *the pathwise first derivative* $(dg(Z,\alpha(t))/d\alpha)[v]$ *exists a.s.* $Z \in \mathcal{Z}$. *Moreover, (i) each element of* $(dg(Z,\alpha(t))/d\alpha)[v_n^*]$ *satisfies the envelope condition and is Hölder continuous in* $\alpha \in \mathcal{N}_{0n}$; *(ii) each element of* $(dm(Z,\alpha(t))/d\alpha)[v_n^*]$ *is in* $\Lambda_c^\gamma(\mathcal{X})$, $\gamma > d_x/2$ *for all* $\alpha \in \mathcal{N}_0$.

The following result is proved in the Appendix.

Theorem 6.1. *Under Assumptions 4.1-4.8, 5.1-5.6 and 6.1-6.3,* $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0,\Omega)$ *where*

$$\begin{aligned}\Omega &= E\left[Var\left(\frac{d\varphi(X,Z,\alpha_0)}{dg(Z,\alpha)}D_{w^*}(Z)\middle|X\right)\right] \\ &= \left[E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle|X\right]\right\}\right]^{-1}\end{aligned} \tag{4.33}$$

Note that if $D_w(Z)$ and $g(Z,\alpha_0)$ are independent conditional on X then the expression (4.33) reduces to the asymptotic variance-covariance formula (22) in AC that is shown to

be asymptotically efficient by these authors. A consistent estimator of Ω can be obtained in the following way: First estimate $W_0(x_i, z_j)^{-1}$ with

$$\begin{aligned} w_{ij} &= p^{k_n}(x_j)' \left(P'P\right)^{-1} p^{k_n}(x_i) \\ \widehat{\Sigma}(x_i, \widehat{\alpha}_n) &= \sum_{j=1}^{n} w_{ij} g(z_j, \widehat{\alpha}_n) g'(z_j, \widehat{\alpha}_n) \\ \widehat{W_0}(x_i, z_j)^{-1} &= \widehat{\Sigma}(x_i, \widehat{\alpha}_n)^{-1} g(z_j, \widehat{\alpha}_n) g'(z_j, \widehat{\alpha}_n) \widehat{\Sigma}(x_i, \widehat{\alpha}_n)^{-1} \end{aligned} \tag{4.34}$$

Then for each $s = 1, ..., d_\theta$ estimate w_s^* with $\widehat{w}_s^*$ which is a solution to the minimization problem

$$\begin{aligned} \min_{w_s \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(\frac{dg(z_j, \widehat{\alpha}_n)}{d\theta_s} - \frac{dg(z_j, \widehat{\alpha}_n)}{dh} [w_s] \right)' \widehat{W_0}(z_j, x_i)^{-1} \\ \times \left(\frac{dg(z_j, \widehat{\alpha}_n)}{d\theta_s} - \frac{dg(z_j, \widehat{\alpha}_n)}{dh} [w_s] \right) \end{aligned}$$

and let $\widehat{w}^* = (\widehat{w}_1^*, ..., \widehat{w}_{d_\theta}^*)$ implying

$$\widehat{D}_{\widehat{w}^*}(z_j) = \frac{dg(z_j, \widehat{\alpha}_n)}{d\theta_s} - \frac{dg(z_j, \widehat{\alpha}_n)}{dh} [\widehat{w}^*] \tag{4.35}$$

Finally, use (4.34) and (4.35) in a finite-sample analog of (4.33) to obtain

$$\widehat{\Omega} = \left[\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}' \widehat{D}_{\widehat{w}^*}(z_j)' \widehat{W_0}(x_i, z_j)^{-1} \widehat{D}_{\widehat{w}^*}(z_j) \right]^{-1}$$

We note that for linear sieves computing $\widehat{w}_s^*$ does not require nonlinear optimization and thus the covariance estimator is easy to compute.

CONCLUSION

In this Chapter extended the LWCEL estimator proposed in the previous Chapter to the semiparametric environment defined by models of conditional moment restrictions containing both θ and infinite dimensional unknown functions h, formally $E\left[g\left(Z,\alpha_0\right)|X\right]=0$. We established consistency of the new estimator $\widehat{\alpha}_n$, convergence rates of $\widehat{\alpha}_n$ under the Fisher norm, and asymptotic normality of the finite-dimensional component $\widehat{\theta}_n$. The new Sieve-based LWCEL estimator (SLWCEL) is a direct alternative to the Sieve Minimum Distance estimator considered by AC that is based on an optimization principle similar to the one of GMM. As shown by Newey and Smith (2004), GEL-type estimators, such as EL, outperform the GMM estimator in terms of higher-order properties in parametric models $E\left[g\left(Z,\theta_0\right)|X\right]=0$. We conjecture that a similar type of improvements is likely to occur also in the semiparametric context of $E\left[g\left(Z,\alpha_0\right)|X\right]=0$.

APPENDIX

Appendix 4.1: Proofs of Main Results

Discussion of Consistency

In outlining our consistency proof, we follow the discussion as given by KTA and extend it to our case of infinite dimensional parameter space. For a standard extremum estimation procedure (for example via maximization), consistency can be shown by considering the sample objective function and its population counterpart and arguing in the following manner. Consider an arbitrary neighborhood of the true parameter value. Check that:

(A) Outside the neighborhood, the sample objective function is bounded away from the maximum of the population objective function achieved at the true parameter value, w.p.a. 1.

(B) The maximum of the sample objective function is by definition not smaller than its value at the true parameter value. The latter converges to the population objective function evaluated at the true value, due to the LLN.

By (A) and (B) the maximum of the sample objective function is unlikely to occur outside the (arbitrarily defined) neighborhood for large samples. This shows the consistency.

While Newey and Powell (2003) were able to recast their estimator as an argmin of a quadratic form delivering (A), in Chen (2005) (Theorem 3.1) (A) is assumed. In our problem, however, such approach cannot be applied directly. Specifically, showing (A) is problematic here, since the objective function G_n defined in (4.10) contains the Lagrange multiplier $\lambda(\alpha_n)$ which is endogenously determined at each α_n. Therefore, in our proof we follow the KTA approach binding G_n with a dominating function and then check (A) for the latter by comparing the convergence rates of G_n at α_0 and outside a $\delta-$neighborhood of α_0. The convergence rate of $G_n(\alpha_0)$ is a new result which differs from the one of KTA since the definition of our G_n contains an additional term σ_i arising from the use of a different weighting scheme and due to our estimator being based on series rather than kernel weights. In our proof, a Uniform Law of Large Numbers (ULLN) for the dominating function is used only for α_n outside the $\delta-$neighborhood of α_0.

Regarding the complications incurred by considering an infinite dimensional parameter space α, we note that our consistency proof differs from the ones used in Newey and Powell (2003) (Theorem 1) and Chen (2005) (Theorem 3.1) for M-estimators with α. Using a ULLN over the sieve space, these authors show that the sample objective function G_n and its expectation are, w.p.a 1, within a $\delta-$neighborhood of each other when evaluated at a parameter $\widetilde{\alpha}_n$ in the sieve space that converges to the true parameter value α_0. Existence of such parameter $\widetilde{\alpha}_n$ is guaranteed by the definition of the sieve space. This approach, however, would necessitate evaluating the convergence rates of $G_n(\widetilde{\alpha}_n)$ to its expectation which is problematic in our saddle-point case since it is difficult to capture the behavior of the endogenous $\lambda_i(\alpha)$ away from α_0. Recall that $\widehat{\alpha}_n$ is defined as maximizing $G_n(\alpha_n)$ over the sieve space $\mathcal{A}_n$ and thus using $G_n(\alpha)$, $\alpha \in \mathcal{A}$ for estimation purposes would yield an unfeasible estimator. Nonetheless, the function $g(z_j, \alpha)$ and hence the functions $G_n(\alpha)$ and $\Sigma_n(x_i, \alpha)$ can theoretically be evaluated at a generic parameter value $\alpha \in \mathcal{A}$ not restricted to the sieve space. Hence the asymptotic rate of convergence of $G_n(\alpha_0)$ at the true parameter value can be derived to facilitate asymptotic analysis.

Further Notation

Let us introduce some additional notation. Let $\|\cdot\|_E$ denote the Euclidean norm. Define

$$\begin{aligned} a_i &\equiv \sigma_i - 1 \\ &= \sum_{j=1}^{n} w_{ij} - 1 \\ &= \mathbf{i}'P\left(P'P\right)^{-1} p^{k_n}(x_i) - 1 \end{aligned}$$

For generic n vectors z and a vector x we drop the subscript i and use

$$a_x \equiv \mathbf{i}'P\left(P'P\right)^{-1} p^{k_n}(x) - 1 \tag{4.36}$$

Further define $B(\alpha_0, \delta)$ and $B_n(\alpha_0, \delta)$ as $\delta-$neighborhoods around α_0 with $B(\alpha_0, \delta) \subset A$ and $B_n(\alpha_0, \delta) \subset A_n$, respectively. Consider the function $\psi(X, \alpha)$ as defined in (4.14). Denote

$$\begin{aligned} \psi_n(x_i, \alpha) &\equiv \sum_{j=1}^{n} w_{ij}\varphi\left(x_i, z_j, \alpha\right) \\ &= \sum_{j=1}^{n} w_{ij} \ln\left\{\sigma_i + \lambda_i' g\left(z_j, \alpha\right)\right\} \end{aligned} \tag{4.37}$$

$$\begin{aligned} G_n(\alpha_n) &\equiv -\frac{1}{n}\sum_{i=1}^{n} \psi_n(x_i, \alpha) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\varphi\left(x_i, z_j, \alpha\right) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \ln\left\{\sigma_i + \lambda_i' g\left(z_j, \alpha_n\right)\right\} \end{aligned} \tag{4.38}$$

$$\begin{aligned} \Sigma_n(x_i, \alpha) &\equiv \sum_{j=1}^{n} w_{ij} g(z_j, \alpha) g'(z_j, \alpha) \\ \Sigma(X, \alpha) &\equiv E_Z\left[\Sigma_n(X, \alpha)\right] \end{aligned} \tag{4.39}$$

and recall the definition of $\Sigma_0(X) \equiv Var[g(Z, \alpha_0)|X]$ in Assumption 6.1 (iii).

Main Proofs

Proof of Theorem 4.1. Following KTA, in the asymptotic analysis we will replace $\lambda_i(\alpha)$ by

$$u\left(x_i,\alpha\right)=\frac{E\left[g\left(z,\alpha\right)|x_i\right]}{\left(1+\left\|E\left[g\left(z,\alpha\right)|x_i\right]\right\|\right)}$$

For a constant $\widetilde{c}\in(0,1)$ define a sequence of truncation sets

$$C_n=\left\{z:\sup_{\alpha\in\mathcal{A}}\left|a_x+u'\left(x,\alpha_n\right)g\left(z,\alpha_n\right)\right|\leq\widetilde{c}n^{1/m}\right\} \tag{4.40}$$

and let

$$s_n\equiv n^{-1/m}\left[a_x+u'\left(x,\alpha_n\right)g\left(z,\alpha_n\right)\right]\mathbb{I}\left\{z\in C_n\right\} \tag{4.41}$$

Let

$$\begin{aligned}q_n\left(x,z,\alpha_n\right)&=-\log\left(1+n^{-1/m}\left[a_x+u'\left(x,\alpha_n\right)g\left(z,\alpha_n\right)\right]\mathbb{I}\left\{z\in C_n\right\}\right)\\&=-\log\left(1+s_n\right)\end{aligned}$$

The modified objective function is

$$Q_n(\alpha_n)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}q_n\left(x_i,z_j,\alpha_n\right) \tag{4.42}$$

Note that

$$G_n(\alpha_n)\leq Q_n(\alpha_n) \tag{4.43}$$

for all $\alpha_n\in\mathcal{A}_n$ by the optimality of λ_i.

Then by the Taylor series expansion for logarithms

$$\begin{aligned}q_n(x,z,\alpha_n)&=-\log\left(1+s_n\right)\\&=-s_n+\frac{\widetilde{s}_n^2}{2}\\&=-s_n+\frac{s_n^2}{2(1-ts_n)}\\&=-n^{-1/m}\left[a_x+u'\left(x,\alpha_n\right)g\left(z,\alpha_n\right)\right]\mathbb{I}\left\{z\in C_n\right\}+\frac{s_n^2}{2(1-ts_n)}\end{aligned}$$

$$\begin{aligned}&=n^{-1/m}\left[a_x+u'\left(x,\alpha_n\right)g\left(z,\alpha_n\right)\right]-n^{-1/m}\left[a_x+u'\left(x,\alpha_n\right)g\left(z,\alpha_n\right)\right]\\&\quad-n^{-1/m}\left[a_x+u'\left(x,\alpha_n\right)g\left(z,\alpha_n\right)\right]\mathbb{I}\left\{z\in C_n\right\}+\frac{s_n^2}{2(1-ts_n)}\\&=-n^{-1/m}\left[a_x+u'\left(x,\alpha_n\right)g\left(z,\alpha_n\right)\right]\\&\quad+n^{-1/m}\left[a_x+u'\left(x,\alpha_n\right)g\left(z,\alpha_n\right)\right]\left(1-\mathbb{I}\left\{z\in C_n\right\}\right)+\frac{s_n^2}{2(1-ts_n)}\\&=-n^{-1/m}\left[a_x+u'\left(x,\alpha_n\right)g\left(z,\alpha_n\right)\right]+R_n(t,a_x,\alpha_n)\end{aligned} \tag{4.44}$$

where

$$\begin{aligned} R_n(t,a_x,\alpha_n) &= n^{-1/m}\left[a_x+u'(x,\alpha_n)\,g(z,\alpha_n)\right](1-\mathbb{I}\{z\in C_n\}) \\ &\quad +\frac{n^{-2/m}\left[a_x+u'(x,\alpha_n)\,g(z,\alpha_n)\right]^2\mathbb{I}\{z\in C_n\}}{2(1-tn^{-1/m}\left[a_x+u'(x,\alpha_n)\,g(z,\alpha_n)\right]\mathbb{I}\{z\in C_n\})^2} \end{aligned}$$

Note that, by the triangle and Cauchy-Schwarz inequalities

$$\begin{aligned} |R_n(t,a_x,\alpha_n)| &\leq n^{-1/m}\left[|a_x|+\left\|u'(x,\alpha_n)\right\|\|g(z,\alpha_n)\|\right](1-\mathbb{I}\{z\in C_n\}) \\ &\quad +\frac{n^{-2/m}\left[a_x^2+2\|a_x\|\|u'(x,\alpha_n)\|\|g(z,\alpha_n)\|+\|u'(x,\alpha_n)\|^2\|g(z,\alpha_n)\|^2\right]\mathbb{I}\{z\in C_n\}}{2(1-tn^{-1/m}\left[a_x+u'(x,\alpha_n)\,g_n(z,\alpha_n)\right])^2} \end{aligned}$$

and by $\|u'(x,\alpha_n)\|<1$ we obtain

$$\begin{aligned} |R_n(t,a_x,\alpha_n)| &\leq n^{-1/m}\left[|a_x|+\|g(z,\alpha_n)\|\right](1-\mathbb{I}\{z\in C_n\}) \\ &\quad +\frac{n^{-2/m}\left[a_x^2+2a_x\|g(z,\alpha_n)\|+\|g(z,\alpha_n)\|^2\right]}{2(1-tn^{-1/m}\left[a_x+u'(x,\alpha_n)\,g_n(z,\alpha_n)\right])^2} \end{aligned}$$

From (4.40) it follows that

$$\begin{aligned} \widetilde{c} &\geq n^{-1/m}\sup_{\alpha\in\mathcal{A}}\left|a_x+u'(x,\alpha_n)\,g(z,\alpha_n)\right| \\ &\geq n^{-1/m}\left|a_x+u'(x,\alpha_n)\,g(z,\alpha_n)\right| \\ &\geq tn^{-1/m}\left|a_x+u'(x,\alpha_n)\,g_n(z,\alpha_n)\right| \end{aligned}$$

and hence

$$\begin{aligned} |R_n(t,a_x,\alpha_n)| &\leq n^{-1/m}\left[|a_x|+\|g(z,\alpha_n)\|\right](1-\mathbb{I}\{z\in C_n\}) \\ &\quad +\frac{n^{-2/m}\left[a_x^2+2a_x\|g(z,\alpha_n)\|+\|g(z,\alpha_n)\|^2\right]}{2(1-\widetilde{c})^2} \\ &= n^{-1/m}\left[|a_x|+\|g(z,\alpha_n)\|\right](1-\mathbb{I}\{z\in C_n\}) \\ &\quad +n^{-2/m}\frac{a_x^2}{2(1-\widetilde{c})^2}+\frac{n^{-2/m}\left[2a_x\|g(z,\alpha_n)\|+\|g(z,\alpha_n)\|^2\right]}{2(1-\widetilde{c})^2} \end{aligned}$$

taking sup over $\mathcal{A}$ we obtain

$$\begin{aligned} \sup_{\alpha\in\mathcal{A}}|R_n(t,a_x,\alpha_n)| &\leq n^{-1/m}\left[|a_x|+\sup_{\alpha\in\mathcal{A}}\|g(z,\alpha_n)\|\right](1-\mathbb{I}\{z\in C_n\})+n^{-2/m}\frac{a_x^2}{2(1-\widetilde{c})^2} \\ &\quad +\frac{n^{-2/m}\left[2a_x\sup_{\alpha\in\mathcal{A}}\|g(z,\alpha_n)\|+\sup_{\alpha\in\mathcal{A}}\|g(z,\alpha_n)\|^2\right]}{2(1-\widetilde{c})^2} \end{aligned} \tag{4.45}$$

In view of (4.44) and (4.45) approximate $n^{1/m}Q_n(\alpha_n)$ by $n^{1/m}\overline{Q}_n(\alpha_n)$ where

$$\overline{Q}_n(\alpha_n)=-\frac{1}{n^{1+1/m}}\sum_{i=1}^{n}u'(x_i,\alpha_n)\,E\left[g(z,\alpha_n)\,|x_i\right] \tag{4.46}$$

Lemma A.2 shows that

$$n^{1/m}Q_n(\alpha_n) = n^{1/m}\overline{Q}_n(\alpha_n) + o_p(1) \qquad \text{uniformly in } \alpha_n \in \mathcal{A}_n \tag{4.47}$$

Next, we will apply a uniform law of large numbers to $n^{1/m}\overline{Q}_n(\alpha)$ over the whole parameter space $\mathcal{A}$. Under Assumptions 4.4(i), 4.5, and 4.6 $E\left[g\left(z,\alpha\right)|x_i\right]$ is continuous in $\alpha \in A$ by Corollary 4.2 of Newey (1991), and so is

$$-u'\left(x_i,\alpha\right)E\left[g\left(z,\alpha\right)|x_i\right] = -\frac{\left\|E\left[g\left(z,\alpha\right)|x_i\right]\right\|^2}{1+\left\|E\left[g\left(z,\alpha\right)|x_i\right]\right\|}$$

Under Assumption 4.5(i) $E\left[\sup_{\alpha\in\mathcal{A}}\left|-u'\left(x_i,\alpha\right)E\left[g\left(z,\alpha\right)|x_i\right]\right|\right] < \infty$. These, together with Assumption 4.4(i) satisfy the conditions of Lemma A2 of Newey and Powell (2003) implying the following uniform law of large numbers:

$$\sup_{\alpha\in\mathcal{A}}\left|n^{1/m}\overline{Q}_n(\alpha) - E\left[-u'\left(x_i,\alpha\right)E\left[g\left(z,\alpha\right)|x_i\right]\right]\right| = o_p(1) \tag{4.48}$$

where $-E\left[-u'\left(x_i,\alpha\right)E\left[g\left(z,\alpha\right)|x_i\right]\right]$ is continuous in $\mathcal{A}$. This function is bounded above by

$$-E\left[u'\left(x_i,\alpha\right)E\left[g\left(z,\alpha\right)|x_i\right]\right] \leq -E\left[\mathbb{I}\left\{x\in\mathcal{X}_{\mathcal{A}}\right\}\left\|E\left[g\left(z,\alpha\right)|x_i\right]\right\|^2/\left(1+\left\|E\left[g\left(z,\alpha\right)|x_i\right]\right\|\right)\right] \tag{4.49}$$

By Assumption 4.1, the right-hand side of this inequality is strictly negative at each $\alpha \neq \alpha_0$. Therefore, by continuity of $-E\left[u'\left(x_i,\alpha\right)E\left[g\left(z,\alpha\right)|x_i\right]\right]$ and compactness of $\mathcal{A}$, there exists a strictly positive number $H(\delta)$ such that

$$\sup_{\alpha\in\mathcal{A}\setminus B(\alpha_0,\delta)} E\left[-u'\left(x_i,\alpha\right)E\left[g\left(z,\alpha\right)|x_i\right]\right] \leq -H(\delta) \tag{4.50}$$

By (4.43), (4.47), and Assumption 4.4(ii) we have

$$\sup_{\alpha_n\in\mathcal{A}_n} n^{1/m}G_n(\alpha_n) \leq \sup_{\alpha_n\in\mathcal{A}_n} n^{1/m}Q_n(\alpha_n) = \sup_{\alpha_n\in\mathcal{A}_n} n^{1/m}\overline{Q}_n(\alpha_n) + o_p(1) \tag{4.51}$$

Together (4.51) with (4.50) and (4.48) imply that

$$\Pr\left\{\sup_{\alpha_n\in\mathcal{A}_n\setminus B_n(\alpha_0,\delta)} G_n(\alpha_n) > -n^{-1/m}H(\delta)\right\} < \delta/2 \qquad \text{eventually.} \tag{4.52}$$

Next, we evaluate G_n at the true value α_0 and show that $G_n(\alpha_0)$ converges to its expectation faster than $G_n(\alpha_n)$ with α_n outside a $\delta-$neighborhood of α_0 whose convergence rate is given in (4.52). Having established this fact the conclusion of the proof is then straightforward. This approach was taken by KTA for the finite-dimensional parameter θ and we extend it to the infinite-dimensional parameter α. Our way of deriving the rate of convergence of $G_n(\alpha_0)$ differs from KTA, though, because we do not make use of kernel-based results. Rather, based on the series literature, we derive a new result for the rate of convergence by specializing Lemma A.1(A) of AC to our case.

Using Lemma A.4 and the fact

$$1 + a_i = \sum_{j=1}^{n} w_{ij} > 0 \quad \text{for each } i$$

we obtain

$$\begin{aligned} G_n(\alpha_0) &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log\left(1+a_i+\lambda_i'(\alpha_0)\,g(z_j,\alpha_0)\right) \\ &\geq -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(a_i+\lambda_i'(\alpha_0)\,g(z_j,\alpha_0)\right) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\lambda_i'(\alpha_0)\sum_{j=1}^{n} w_{ij}g(z_j,\alpha_0) \\ &\geq -\max_{1\leq i\leq n}\|\lambda_i(\alpha_0)\|\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g(z_j,\alpha_0)\right\| \end{aligned}$$

Then by Lemmas A.1 and A8,

$$\begin{aligned} G_n(\alpha_0) &= \left\{o_p(\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-1/m}}\right)\right\}^2 \\ &= o_p(r_n^2) \end{aligned}$$

where

$$r_n \equiv o_p(\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-1/m}}\right)$$

with $\widetilde{\delta}_{1n}$ defined in Lemma A.7 and ϱ defined in 4.7. Therefore, we have the following LLN

$$\Pr\left\{G_n(\alpha_0) < -r_n^2 H(\delta)\right\} < \delta/2 \qquad \text{eventually.} \tag{4.53}$$

Denote

$$\begin{aligned} \widehat{Q}_1(\alpha) &\equiv n^{1/m}G_n(\alpha) \\ \widehat{Q}_2(\alpha) &\equiv r_n^{-2}G_n(\alpha) \\ Q_1(\alpha) &\equiv -E\left[u'(x,\alpha)E\left[g(z,\alpha)|x\right]\right] \\ Q_2(\alpha) &\equiv E\widehat{Q}_2(\alpha) \end{aligned}$$

where the last expectation is taken with respect to the joint density of (Y,X). Under Assumptions 4.4(i), 4.5, and 4.6 $Q_2(\alpha)$ is continuous in $\alpha\in\mathcal{A}$ by Corollary 4.2 of Newey (1991). Note that since $n^{1/m}r_n^2\to 0$ and $n^{1/m}G_n(\alpha)\leq 0$, by (4.48) and (4.51), w.p.a. 1,

$$\begin{aligned} r_n^{-2} &> n^{1/m} \\ \widehat{Q}_2(\alpha) &\leq \widehat{Q}_1(\alpha) \end{aligned} \tag{4.54}$$

If we retain $\lambda_i(\alpha)$ instead of $u(x,\alpha)$ in the definition of $Q_n(\alpha)$ in (4.42), using $\lambda_i(\alpha)=O_p(1)$ which follows from (4.11), we can derive an analog of $\overline{Q}_n(\alpha)$ in (4.46) as

$$\overline{Q}_{2n}(\alpha) = -\frac{1}{n^{1+1/m}}\sum_{i=1}^{n}\lambda_i'(\alpha)E\left[g(z,\alpha)|x_i\right]$$

By a corresponding analog of (4.47) and the moment restriction $E\left[g\left(z,\alpha_0\right)|x_i\right]=0$ it follows that $\overline{Q}_{2n}(\alpha_0)=0$ and $Q_2(\alpha_0)=0$. Also, by (4.49) $Q_1(\widehat{\alpha}_n)<0$ for each $\theta\neq\theta_0$ and thus

$$Q_1(\widehat{\alpha}_n)\leq 0 \tag{4.55}$$

Then, w.p.a. 1,

$$\begin{aligned} Q_1(\widehat{\alpha}_n) &\geq \widehat{Q}_1(\widehat{\alpha}_n)+H(\delta)/2 && (4.56)\\ &\geq \widehat{Q}_1(\alpha_0)+H(\delta)/2 && (4.57)\\ &\geq \widehat{Q}_2(\alpha_0)+H(\delta)/2 && (4.58)\\ &> Q_2(\alpha_0)+H(\delta) && (4.59)\\ &= H(\delta) && (4.60)\end{aligned}$$

where (4.56) holds by (4.48) and (4.51), (4.57) holds by the definition of $\widehat{\alpha}_n$, (4.58) by (4.54), (4.59) by LLN at α_0 (4.53), and (4.60) by $Q_2(\alpha_0)=0$. By (4.55) and δ being arbitrary, taking $H(\delta)\to 0$,

$$\widehat{Q}_1(\widehat{\alpha}_n)\xrightarrow{p}0$$

Then, using Assumption 4.4(ii), $\Pr\left(\left|\widehat{Q}_1(\widehat{\alpha})-Q_2(\alpha_0)\right|\geq H(\delta)\right)\to 0$ and by (4.52)
$\Pr\left(\widehat{\alpha}_n\in\mathcal{A}_n\backslash B_n\left(\theta_0,\delta\right)\right)\to 0$.

□

Proof of Theorem 5.1. In deriving the convergence rates under the Fisher norm $\|\cdot\|_F$ we will proceed in a way that is similar to the proof of Theorem 3.1 in AC. Specifically, we will use their Lemma A.1 and Corollary A.1 that hold for a generic function $m(X,\alpha)$ and the Euclidean metric. However, since our objective function and metric differs from the ones used by these authors, we need to derive the counterparts of their Corollaries A.2 and B.1 for our case.

Recall the definition of $G_n(\alpha_n)$ in (4.38)

$$G_n(\alpha_n)\equiv-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\ln\left\{\sigma_i+\lambda_i'g\left(z_j,\alpha_n\right)\right\}$$

and define

$$\overline{G}_n(\alpha_n)\equiv-\frac{1}{n}\sum_{i=1}^{n}E\left[\ln\left\{\sigma_i+\lambda_i'g\left(z,\alpha_n\right)\right\}|x_i\right] \tag{4.61}$$

Let $\delta_{0n}=o(n^{-1/4})$ and denote $\alpha_{n0}=\Pi\alpha_0$ (the orthogonal projection of α_0 onto the sieve space).

$$P\left(\|\widehat{\alpha}_n-\alpha_0\|_F\geq\delta_{0n}\right)=P\left(\sup_{\left\{\|\widehat{\alpha}_n-\alpha_0\|_F\geq\delta_{0n},\alpha_n\in\mathcal{A}_n\right\}}G_n(\alpha_n)\geq G_n(\alpha_{n0})\right)$$

Note that Assumptions 3.1-3.2, 3.6-3.8 and 4.1(iii) in AC are equivalent to our Assumptions 4.2, 4.3, 5.2, 4.5, 4.6, 5.3-5.5 and 5.6, respectively. Assumption 3.3 in AC is implied by our Assumption 4.1 and the condition (4.1). The analog of AC's Assumption 3.4 for our $\Sigma_n(x_i,\alpha)$ defined in (4.39) is satisfied by AC's Corollary A.1(i). Thus Assumptions of AC's Lemma A.1 and Corollary A.1 are satisfied.

Lemma B.1 states the counterparts of their AC's Corollaries A.2 and B.1 for our case. We note that condition (A) of our consistency proof was shown to hold for $G_n(\alpha_n)$ in Theorem 4.1.

Since $\widetilde{G}_n(\alpha_n) \leq G_n(\alpha_n)$, by (4.51) the condition also holds for $\widetilde{G}_n(\alpha_n)$. Thus the identification condition is satisfied. Satisfying Assumptions of Theorem 1 of Shen and Wong (1994) is also a necessary condition for AC's Theorem 3.1. Since the role of the pseudodistance in Theorem 1 of Shen and Wong (1994) is performed by our metric $\|\cdot\|_F^2$ in a way topologically equivalent to the AC's one, and the remaining AC's Assumptions hold as described above, this condition is also satisfied. Invocation of AC's Theorem 3.1, with their objective function and metric replaced with ours, completes the proof.

□

Proof of Theorem 6.1. Substituting (4.32) into (4.31) yields

$$\frac{dg(Z,\alpha_0)}{d\alpha}[v^*] = D_{w^*}(Z)\left[E\left\{E\left[D_w(Z)'W_0(Z,X)^{-1}D_w(Z)\middle| X\right]\right\}\right]^{-1}\tau \tag{4.62}$$

Note that by the chain rule

$$\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[v^*] = \frac{d\varphi(X,Z,\alpha_0)}{dg(Z,\alpha)}\frac{dg(Z,\alpha_0)}{d\alpha}[v^*] \tag{4.63}$$

Using Lemma C.1 and (4.62) in (4.63), we obtain

$$\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}[v^*] = \frac{d\varphi(X,Z,\alpha_0)}{dg(Z,\alpha)}D_{w^*}(Z)\left[E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle| X\right]\right\}\right]^{-1}\tau \tag{4.64}$$

We will now check the conditions for Theorem 7.1 in Appendix 3 that is an extension of Theorem 1 of Shen (1997) to our conditional case. Lemma C.2 shows that under our Assumptions, Conditions A is satisfied. Since $\{g(z,\alpha_n) : \alpha_n \in \mathcal{A}_n\} \subset \Lambda_c^\gamma(\mathcal{X})$, Condition B follows directly from Lemma B.1. Since $\|\widehat{\alpha}_n - \alpha_0\|_F = o_p(n^{-1/4})$, then $\delta_n = n^{-1/4}$ and hence for Condition C we require

$$\begin{aligned}\sup_{\{\alpha_n\in\mathcal{A}_n:\|\alpha_n-\alpha_0\|\leq\delta_n\}} \|\varepsilon_n u^* - \varepsilon_n u_n^*\| &= O_p(\delta_n^{-1}\varepsilon_n^2)\\ &= O_p(n^{-1/4})\end{aligned}$$

which is satisfied by Assumption 6.2. Condition D follows from the smoothness of $\frac{d\varphi(x_i,z_j,\alpha_0)}{d\alpha}[\alpha-\alpha_0]$ in $\mathcal{N}_{0n}$. Condition F is satisfied by the definition of $f(\widehat{\alpha}_n) \equiv \tau'\widehat{\theta}_n$, $\omega = 1$, and Assumption 6.2. Condition G is satisfied by Assumption 6.1.

By Theorem 7.1 in Appendix 3, for arbitrarily fixed $\tau \in \mathbb{R}^{d_\theta}$ with $|\tau| \neq 0$,

$$\sqrt{n}\tau'(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma_{v^*})$$

where

$$\begin{aligned}\Sigma_{v^*} &\equiv E\left[Var\left(\frac{d\varphi(X,Z,\alpha_0)}{d\alpha}\middle| X\right)\right]\\ &= \tau'\Omega\tau\end{aligned} \tag{4.65}$$

and hence

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Omega)$$

Using (4.64) in (4.65) we obtain

$$\begin{aligned}\Omega &= \left[E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle|X\right]\right\}\right]^{-1} \\ &\times E\left[Var\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}D_{w^*}(Z)\middle|X\right)\right] \\ &\times\left[E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle|X\right]\right\}\right]^{-1} \end{aligned} \tag{4.66}$$

Using Lemma C.1 and (4.66)

$$\Omega = \left[E\left\{E\left[D_{w^*}(Z)'W_0(Z,X)^{-1}D_{w^*}(Z)\middle|X\right]\right\}\right]^{-1}$$

□

Appendix 4.2: Auxiliary Results

Consistency

Lemma A.1 (B.3). *Let Assumptions 4.5 and 4.7 hold. Then, pointwise for a given $\alpha \in \mathcal{A}$,*

$$\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\| = o_p(\widetilde{\delta}_{1n}) + o_p\left(\frac{1}{n^{\varrho-1/m}}\right)$$

where $\widetilde{\delta}_{1n}$ is defined in Lemma A.7 and ϱ in Assumption 4.7.

Proof. Decompose

$$\begin{aligned}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\| &\leq \max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|\mathbb{I}_{i,n} \\ &+ \max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|\max_{1\leq i\leq n}\mathbb{I}^c_{i,n}\end{aligned}$$

Note that he results of Lemma D.3 and D.5 in KTA hold also for w_{ij} as defined in this paper. Therefore

$$\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|\max_{1\leq i\leq n}\mathbb{I}^c_{i,n} = o_p\left(\frac{1}{n^{\varrho-1/m}}\right)$$

Next, pick any $\epsilon>0$, $c_n\downarrow 0$, and observe that

$$\begin{aligned}&\Pr\left\{\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|\mathbb{I}_{i,n}>\epsilon c_n\right\} \\ \leq\ &\Pr\left\{\sup_{X\in\mathcal{X}}\left\|\sum_{j=1}^{n} w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|>\epsilon c_n\right\}\end{aligned}$$

Using Lemma A.7,

$$\Pr\left\{\sup_{X\in\mathcal{X}}\left\|\sum_{j=1}^{n}w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|>\epsilon c_n\right\}\leq\epsilon$$

if

$$c_n=\widetilde{\delta}_{1n}$$

where $\widetilde{\delta}_{1n}$ is defined in Lemma A.7. Hence

$$\max_{1\leq i\leq n}\left\|\sum_{j=1}^{n}w_{ij}g\left(z_j,\alpha\right)-E\left[g\left(z,\alpha\right)|x_i\right]\right\|\mathbb{I}_{i,n}=o_p\left(\widetilde{\delta}_{1n}\right)$$

and the desired result follows. □

Lemma A.2 (B.8). *Let Assumptions 4.5 and 4.7 hold. Then*

$$\sup_{\alpha_n\in\mathcal{A}_n}\left|Q_n(\alpha_n)-\overline{Q}_n(\alpha_n)\right|=o_p(n^{-1/m})$$

Proof. Substituting from (4.44) for $q_n\left(x_i,z_j,\alpha_n\right)$ we obtain

$$\begin{aligned}
&n^{1/m}\sup_{\alpha_n\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}q_n\left(x_i,z_j,\alpha_n\right)+\frac{1}{n^{1+1/m}}\sum_{i=1}^{n}u'\left(x_i,\alpha_n\right)E\left[g\left(z,\alpha_n\right)|x_i\right]\right|\\
\leq\;&n^{1/m}\sup_{\alpha_n\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left\{-n^{-1/m}\left[a_i+u'\left(x_i,\alpha_n\right)g\left(z_j,\alpha_n\right)\right]\right\}+\frac{1}{n^{1+1/m}}\sum_{i=1}^{n}u'\left(x_i,\alpha_n\right)E\left[g\left(z,\alpha_n\right)|x_i\right]\right|\\
&+n^{1/m}\sup_{\alpha_n\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}R_n(t,a_i,\alpha_n)\right|\\
=\;&\sup_{\alpha_n\in\mathcal{A}_n}\left|-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}a_i+\frac{1}{n}\sum_{i=1}^{n}u'\left(x_i,\alpha\right)E\left[g\left(z,\alpha_n\right)|x_i\right]-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}u'\left(x_i,\alpha_n\right)g\left(z_j,\alpha_n\right)\right|\\
&+n^{1/m}\sup_{\alpha_n\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}R_n(t,a_i,\alpha_n)\right|\\
\leq\;&-\sup_{\alpha_n\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}a_i\right|+\sup_{\alpha_n\in\mathcal{A}_n}\frac{1}{n}\sum_{i=1}^{n}\left\|E\left[g\left(z,\alpha_n\right)|x_i\right]-\sum_{j=1}^{n}w_{ij}g\left(z_j,\alpha_n\right)\right\|\\
&+n^{1/m}\sup_{\alpha_n\in\mathcal{A}_n}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}R_n(t,a_i,\alpha_n)\right|
\end{aligned}$$

The first term drops out by Lemma A.4, the second term is $o_p(1)$ by Corollary A.1(i) in AC, p. 1824, and the third term is $o_p(1)$ by Lemma A.3. □

Lemma A.3. *Let Assumptions 4.5 and 4.7 hold. Then*

$$n^{1/m} \sup_{\alpha_n \in \mathcal{A}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} R_n(t, a_i, \alpha_n) \right| = o_p(1)$$

Proof. Note that by (4.45)

$$\begin{aligned}
& n^{1/m} \sup_{\alpha_n \in \mathcal{A}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} R_n(t, a_i, \alpha_n) \right| \\
\leq \; & \frac{1}{n^{1-1/m}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sup_{\alpha_n \in \mathcal{A}_n} |R_n(t, a_i, \alpha_n)| \\
\leq \; & \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left[|a_i| + \sup_{\alpha_n \in \mathcal{A}_n} \|g(z_j, \alpha_n)\| \right] (1 - \mathbb{I}\{z_j \in C_n\}) \\
& + \frac{1}{n^{1+1/m}} \frac{1}{2(1-\widetilde{c})^2} \sum_{i=1}^{n} a_i^2 \sum_{j=1}^{n} w_{ij} \\
& + \frac{1}{n^{1+1/m}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \frac{\left[2a_i \sup_{\alpha_n \in \mathcal{A}_n} \|g(z_j, \alpha_n)\| + \sup_{\alpha \in \mathcal{A}} \|g(z_j, \alpha_n)\|^2 \right]}{2(1-\widetilde{c})^2} \\
= \; & D_1 + D_2 + D_3
\end{aligned}$$

By Assumption 4.5(i) and 4.4(ii), $\sup_{\alpha_n \in \mathcal{A}_n} \|g(z, \alpha_n)\| < \infty$. By Lemma A.5 $|a_i| < \infty$ and hence by Lemma A.6

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left[|a_i| + \sup_{\alpha_n \in \mathcal{A}_n} \|g(z_j, \alpha_n)\| \right] = O_p(1).$$

Since $\max_{1 \leq j \leq n} \mathbb{I}\{z_j \notin C_n\} = o_p(1)$, $D_1 = o_p(1)$. By Lemma A.6 $D_2 = o_p(1)$.

$$\begin{aligned}
D_3 &= \frac{1}{n^{1+1/m}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \frac{\left[2a_i \sup_{\alpha_n \in \mathcal{A}_n} \|g(z_j, \alpha_n)\| + \sup_{\alpha \in \mathcal{A}} \|g(z_j, \alpha_n)\|^2 \right]}{2(1-\widetilde{c})^2} \\
&= \frac{1}{n^{1+1/m}(1-\widetilde{c})^2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} a_i + \frac{1}{n^{1+1/m}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \frac{\sup_{\alpha_n \in \mathcal{A}_n} \|g(z_j, \alpha_n)\|^2}{2(1-\widetilde{c})^2}
\end{aligned}$$

where the first part drops out by Lemma A.4 and the second part is $o_p(1)$ by Assumption 4.5(i), 4.4(ii) and Lemma A.6. □

Lemma A.4. *Under Assumptions 4.3 and 4.4, for w_{ij} defined in (4.5) and a_i defined in (4.36), it holds that*

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} a_i = 0$$

Proof.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}a_i &= \frac{1}{n}\sum_{i=1}^{n} a_i \sum_{j=1}^{n} w_{ij} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} w_{ij} - 1\right]\sum_{j=1}^{n} w_{ij} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i)\mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i) - \mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i)p^{k_n}(x_i)'\left(P'P\right)^{-1}P'\mathbf{i} - \mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i)\right] \\
&= \mathbf{i}'P\left(P'P\right)^{-1}\left(P'P\right)\left(P'P\right)^{-1}P'\mathbf{i} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{i}'P\left(P'P\right)^{-1}p^{k_n}(x_i) \\
&= \frac{1}{n}\mathbf{i}'P\left(P'P\right)^{-1}P'\mathbf{i} - \frac{1}{n}\mathbf{i}'P\left(P'P\right)^{-1}P'\mathbf{i} \\
&= 0
\end{aligned}$$

□

Lemma A.5. *Under Assumptions 4.3 and 4.4, for w_{ij} defined in (4.5),*

$$\sum_{j=1}^{n} w_{ij} = O(1)$$

for each $X \in \mathcal{X}$.

Proof. By Assumption 4.3, for any $E\left[\rho_l\left(Z,\alpha\right)|x_i\right]$ there exists $p^{k_n}(x_i)'\pi_l = \sum_{j=1}^{n} w_{ij}g_l\left(z_j,\alpha\right)$ such that

$$E\left[E\left[g_l\left(Z,\alpha\right)|x_i\right] - \sum_{j=1}^{n} w_{ij}g_l\left(z_j,\alpha\right)\right] = O(1)$$

The result follows by boundedness of $g_l\left(z_j,\alpha\right)$. □

Lemma A.6. *Under Assumptions 4.3 and 4.4, for w_{ij} defined in (4.5),*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = O_p(1)$$

Proof. Follows directly from Lemma A.5. □

Lemma A.7. *Let*

$$\begin{aligned}
\xi_{0n} &\equiv \sup_{X\in\mathcal{X}}\left\|p^{k_n}(X)\right\|_E \\
\xi_{1n} &\equiv \sup_{X\in\mathcal{X}}\left\|\frac{\partial p^{k_n}(X)}{\partial x'}\right\|_E
\end{aligned}$$

Let $\widetilde{g}: \mathcal{Z} \to \mathbb{R}$ denote a generic measurable function of the data $Z \in \mathcal{Z}$, evaluated at a given fixed parameter α. Define $\varepsilon(Z,\alpha) = \widetilde{g}(Z,\alpha) - E[\widetilde{g}(Z,\alpha)|X]$ and $\varepsilon(\alpha) = (\varepsilon(Z_1,\alpha),...,\varepsilon(Z_n,\alpha))'$.

Suppose that Assumptions 4.2 and 4.3(i) and the following are satisfied:

(i) There exists a constant c_{1n} and a measurable function $c_1(Z): \mathcal{Z} \to [0,\infty)$ with $E[c_1(Z)^p] < \infty$ for some $p \geq 4$ such that $|\widetilde{g}(Z,\alpha)| \leq c_{1n} c_1(Z)$ for all $Z \in \mathcal{Z}$;

(ii) There exists a positive value $\widetilde{\delta}_{1n} = o_p(1)$ such that

$$\frac{n\widetilde{\delta}_{1n}^2}{\ln\left[(\frac{\xi_{1n}c_{1n}}{\delta_{1n}})^{d_x}\right]\max\left\{\xi_{0n}^2 c_{1n}^2, \xi_{0n}^{2+2/p}\delta_{1n}^{1-2/p}c_{1n}^{1+2/p}\right\}} \to \infty$$

Then

$$p^{k_n}(X)'(P'P)^{-1}P'\varepsilon(\alpha) = o_p(\delta_{1n})$$

uniformly over $X \in \mathcal{X}$.

Proof. This result specializes Lemma A.1(A) in AC, derived for the combined space $\mathcal{X} \times \mathcal{A}$ to the space $\mathcal{X}$ only, with $g(z_j,\alpha)$ evaluated at a given fixed α. Since we do not have to account for growth restrictions on the parameter space, we are able to obtain faster convergence rate δ_{1n} than AC.

Let c denote a generic constant that may have different values in different expressions. For any pair $X_1 \in \mathcal{X}$ and $X_2 \in \mathcal{X}$

$$\begin{aligned} & \left|p^{k_n}(X_1)'(P'P)^{-1}P'\varepsilon(\alpha) - p^{k_n}(X_2)'(P'P)^{-1}P'\varepsilon(\alpha)\right| \\ = & \left|\left[p^{k_n}(X_1) - p^{k_n}(X_2)\right]'(P'P)^{-1}P'\varepsilon(\alpha)\right| \end{aligned}$$

Note that

$$\left\|p^{k_n}(X_1)' - p^{k_n}(X_2)'\right\|_E^2 \leq \xi_{1n}^2 \|X_1 - X_2\|_E^2$$

It follows that

$$\left|\left[p^{k_n}(X_1) - p^{k_n}(X_2)\right]'(P'P)^{-1}P'\varepsilon(\alpha)\right| \leq \xi_{1n}^2 \|X_1 - X_2\|_E^2 \sqrt{\frac{1}{n\lambda_n}\sum_{i=1}^n \varepsilon(Z_i,\alpha)^2}$$

where λ_n denotes the smallest eigenvalues of $P'P/n$. Condition (i) implies

$$\frac{1}{n}\sum_{i=1}^n \varepsilon(Z_i,\alpha)^2 \leq \frac{c_{1n}^2}{n}\sum_{i=1}^n (c_1(Z_i) + E[c_1(Z_i)|X_i])^2$$

Assumption 4.3(i) implies $\lambda_n = O_p(1)$. Applying the weak law of large numbers and $E\left\{(E[c_1(Z_i)|X_i])^2\right\} \leq E\left\{c_1(Z)^2\right\}$, we obtain

$$\frac{1}{n}\sum_{i=1}^n (c_1(Z_i) + E[c_1(Z_i)|X_i])^2 = O_p(1)$$

Thus there exists a constant c such that

$$\Pr\left(\sqrt{\frac{1}{n\lambda_n}\sum_{i=1}^n (c_1(Z_i) + E[c_1(Z_i)|X_i])^2} > c\right) < \eta$$

for sufficiently large n.

For any small ϵ partition $\mathcal{X}$ into b_n mutually exclusive subsets $\mathcal{X}_m$, $m = 1, ..., b_n$, where $X_1 \in \mathcal{X}_m$ and $X_2 \in \mathcal{X}_m$ imply $\|X_1 - X_2\|_E^2 \leq \epsilon\widetilde{\delta}_{1n}/(c_{1n}\xi_{1n}c)$. Then with probability approaching one we have

$$\left|p^{k_n}(X_1)'(P'P)^{-1}P'\varepsilon(\alpha) - p^{k_n}(X_2)'(P'P)^{-1}P'\varepsilon(\alpha)\right| \leq \epsilon\widetilde{\delta}_{1n}$$

Let X_m denote a fixed point in $\mathcal{X}_m$. For any X there exists an m such that $\|X_1 - X_2\|_E^2 \leq \epsilon\widetilde{\delta}_{1n}/(c_{1n}\xi_{1n}c)$. Then with probability approaching one

$$\sup_{X\in\mathcal{X}}\left|p^{k_n}(X)'(P'P)^{-1}P'\varepsilon(\alpha)\right| \leq \epsilon\widetilde{\delta}_{1n} + \max_m\left|p^{k_n}(X_m)'(P'P)^{-1}P'\varepsilon(\alpha)\right|$$

Hence

$$\begin{aligned}
&\Pr\left(\sup_{X\in\mathcal{X}}\left|p^{k_n}(X)'(P'P)^{-1}P'\varepsilon(\alpha)\right| > 2\epsilon\widetilde{\delta}_{1n}\right) \\
< \;& 2\eta + \Pr\left(\max_m\left|p^{k_n}(X_m)'(P'P)^{-1}P'\varepsilon(\alpha)\right| > 2\epsilon\widetilde{\delta}_{1n}\right)
\end{aligned}$$

For some constant c, let

$$M_n = \left(\frac{c\xi_{0n}c_{1n}}{\delta_{1n}\epsilon\eta}\right)^{2/p}$$

Define $d_{in} = \mathbb{I}\{c_1(Z) \leq M_n\}$. Define $g_1(Z_i,\alpha) = d_{in}g_1(Z_i,\alpha)$ and $g_2(Z_i,\alpha) = (1-d_{in})\,g_1(Z_i,\alpha)$. Define $\varepsilon_1(Z_i,\alpha)$ and $\varepsilon_2(Z_i,\alpha)$ accordingly. It follows that

$$\begin{aligned}
&\Pr\left(\max_m\left|p^{k_n}(X_m)'(P'P)^{-1}P'\varepsilon(\alpha)\right| > 2\epsilon\widetilde{\delta}_{1n}\right) \\
\leq \;& \Pr\left(\max_m\left|p^{k_n}(X_m)'(P'P)^{-1}\sum_{i=1}^n\varepsilon_1(Z_i,\alpha)\right| > \epsilon\widetilde{\delta}_{1n}\right) \\
&+ \Pr\left(\max_m\left|p^{k_n}(X_m)'(P'P)^{-1}\sum_{i=1}^n\varepsilon_2(Z_i,\alpha)\right| > \epsilon\widetilde{\delta}_{1n}\right) \\
\equiv \;& P_1 + P_2
\end{aligned}$$

AC show that $P_2 \leq \eta$, along with

$$\sigma_m^2 \equiv nE\left\{\left[p^{k_n}(X_m)'(P'P)^{-1}\sum_{i=1}^n p^{k_n}(X_i)\varepsilon_1(Z_i,\alpha)\right]^2\right\} = O(c_{1n}^2\xi_{0n}^2)$$

and

$$\left|p^{k_n}(X_m)'(P'P/n)^{-1}p^{k_n}(X_i)\varepsilon_1(Z_i,\alpha)\right| \leq \frac{M_n\xi_{0n}^2c_{1n}}{\lambda_n}$$

Noting that

$$\begin{aligned}
&\Pr\left(\left|p^{k_n}(X_m)'(P'P)^{-1}\sum_{i=1}^n\varepsilon_1(Z_i,\alpha)\right| > \epsilon\widetilde{\delta}_{1n}\right) \\
= \;& E\left[\Pr\left(\left|p^{k_n}(X_m)'(P'P)^{-1}\sum_{i=1}^n\varepsilon_1(Z_i,\alpha)\right| > \epsilon\widetilde{\delta}_{1n} \mid X_1, ..., X_n\right)\right]
\end{aligned}$$

AC apply the Bernstein inequality for independent processes to obtain

$$\begin{aligned}&\Pr\left(\left|p^{k_n}(X_m)'(P'P)^{-1}\sum_{i=1}^{n}\varepsilon_1(Z_i,\alpha)\right|>\epsilon\delta_{1n}\right)\\ \leq\ & 2E\left[\exp\left(-n\varepsilon^2\widetilde{\delta}_{1n}^2/\left(c\sigma_m^2+M_n\xi_{0n}^2c_{1n}^2\lambda_n^{-1}\varepsilon\widetilde{\delta}_{1n}\right)\right)\right]\end{aligned}$$

where $E[\cdot]$ is taken with respect to the joint distribution of $(X_1,...,X_n)$. Hence

$$P_1<2b_nE\left[\exp\left(-n\varepsilon^2\widetilde{\delta}_{1n}^2/\left(c\sigma_m^2+M_n\xi_{0n}^2c_{1n}^2\lambda_n^{-1}\varepsilon\widetilde{\delta}_{1n}\right)\right)\right]$$

which is arbitrarily small if

$$\frac{n\widetilde{\delta}_{1n}^2}{\max\left\{\xi_{0n}^2c_{1n}^2,M_n\xi_{0n}^2c_{1n}\widetilde{\delta}_{1n}\right\}}-\ln(b_n)\to\infty$$

Since $\mathcal{X}$ is a compact subset in $\mathbb{R}^d$, we have

$$b_n=O\left(\left(\frac{\widetilde{\delta}_{1n}}{c_{1n}\xi_{1n}}\right)^{-d_x}\right)$$

Substituting for M_n and b_n we obtain

$$\begin{aligned}&\frac{n\widetilde{\delta}_{1n}^2}{\ln(b_n)\max\left\{\xi_{0n}^2c_{1n}^2,M_n\xi_{0n}^2c_{1n}\widetilde{\delta}_{1n}\right\}}\\ =\ & O\left(\frac{n\widetilde{\delta}_{1n}^2}{\ln\left[\left(\frac{\widetilde{\delta}_{1n}}{c_{1n}\xi_{1n}}\right)^{-d_x}\right]\max\left\{\xi_{0n}^2c_{1n}^2,\xi_{0n}^{2+2/p}\widetilde{\delta}_{1n}^{1-2/p}c_{1n}^{1+2/p}\right\}}\right)\end{aligned}$$

Thus, for $P_1<\eta$ for sufficiently large n by condition (ii). □

Lemma A.8 (part of B.1). *Let Assumptions 4.2-4.6 and 4.8 hold. Let also $n^{1/m}\widetilde{\delta}_{1n}\downarrow 0$ and $\rho>2/m$ where $\widetilde{\delta}_{1n}$ is defined in Lemma A.7 and ϱ in Assumption 4.7. Then*

$$\max_{1\leq i\leq n}\|\lambda_i(\alpha_0)\|=o_p(\widetilde{\delta}_{1n})+o_p\left(\frac{1}{n^{\varrho-1/m}}\right)\tag{4.67}$$

This Lemma is analogous to Lemma B.1 of KTA. However, the analysis is somewhat complicated due to the extra term σ_i. Moreover, here we do not make use of results related to kernel estimation. Thus, for example, consistency of the variance-covariance matrix $\Sigma_n(x_i,\alpha_0)$ follows from series results of AC.

Proof. In this Lemma, we will use the F.O.C.s (3.17) and (3.19) that combine to

$$\begin{aligned}\sum_{j=1}^{n}\frac{w_{ij}}{1+a_i+\lambda_i' g\left(x_j,\alpha\right)} &= \sum_{j=1}^{n}\frac{w_{ij}}{\lambda_i' g\left(x_j,\alpha\right)+\sigma_i}\\ &= \sum_{j=1}^{n}\widehat{\pi}_{ij}\\ &= 1\end{aligned} \tag{4.68}$$

Let

$$\lambda_i\left(\alpha_0\right)=\rho_i\xi_i \tag{4.69}$$

where $\rho_i\geq 0$ and $\xi_i\in\mathbb{R}^{d_g}$. It holds that

$$\begin{aligned}\sum_{j=1}^{n}w_{ij}\frac{\left[a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)\right]^2}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)} &= a_i^2\sum_{j=1}^{n}\frac{w_{ij}}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)}+\frac{2a_i\rho_i\sum_{j=1}^{n}w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)}\\ &\quad+\frac{\rho_i^2\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)}\end{aligned} \tag{4.70}$$

For the first term of the RHS sum of (4.70), using (4.68), it holds that

$$\begin{aligned}a_i^2\sum_{j=1}^{n}\frac{w_{ij}}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)} &= a_i^2\\ &= \left(\sigma_i-1\right)^2\\ &= \sigma_i^2-2\sigma_i+1\end{aligned} \tag{4.71}$$

Substituting (4.71) into (4.70) yields

$$\begin{aligned}\sum_{j=1}^{n}w_{ij}\frac{\left[a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)\right]^2}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)} &= \sigma_i^2-2\sigma_i+1+\frac{2a_i\rho_i\sum_{j=1}^{n}w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)}\\ &\quad+\frac{\rho_i^2\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}{1+a_i+\lambda_i'\left(\alpha_0\right)g\left(z_j,\alpha_0\right)}\end{aligned} \tag{4.72}$$

Note that for a generic constant c

$$\begin{aligned}\frac{c^2}{1+c} &= \frac{c^2}{1+c}+(1-c)-(1-c)\\ &= \frac{c^2}{1+c}+\frac{(1-c)(1+c)}{1+c}-(1-c)\\ &= \frac{c^2}{1+c}+\frac{1-c^2}{1+c}-(1-c)\\ &= \frac{1}{1+c}-1+c\end{aligned}$$

Using this fact, letting $c = a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0)$, we have

$$\begin{aligned}
\sum_{j=1}^{n} w_{ij} \frac{\left[a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0)\right]^2}{1 + a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0)} &= \sum_{j=1}^{n} w_{ij} \left\{ \frac{1}{1 + a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0)} - 1 + a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0) \right\} \\
&= \sum_{j=1}^{n} \frac{w_{ij}}{1 + a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0)} - \sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ij} a_i \\
&\quad + \sum_{j=1}^{n} w_{ij} \lambda_i'(\alpha_0)\, g(z_j, \alpha_0) \\
&= 1 - \sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ij} a_i + \sum_{j=1}^{n} w_{ij} \lambda_i'(\alpha_0)\, g(z_j, \alpha_0) \qquad (4.73)
\end{aligned}$$

By the definition of σ_i,

$$\begin{aligned}
1 - \sum_{j=1}^{n} w_{ij} + a_i \sum_{j=1}^{n} w_{ij} &= 1 - \sigma_i + (\sigma_i - 1)\sigma_i \\
&= \sigma_i^2 - 2\sigma_i + 1 \qquad (4.74)
\end{aligned}$$

Substituting (4.74) into (4.73) gives us

$$\sum_{j=1}^{n} w_{ij} \frac{\left[a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0)\right]^2}{1 + a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0)} = \sigma_i^2 - 2\sigma_i + 1 + \rho_i \sum_{j=1}^{n} w_{ij} \xi_i' g(z_j, \alpha_0) \qquad (4.75)$$

Combining (4.72) and (4.75) yields, after canceling $\sigma_i^2 - 2\sigma_i + 1$ from both sides,

$$\frac{2a_i \rho_i \sum_{j=1}^{n} w_{ij} \xi_i' g(z_j, \alpha_0)}{1 + a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0)} + \frac{\rho_i^2 \xi_i' \Sigma_n(x_i, \alpha_0) \xi_i}{1 + a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0)} = \rho_i \sum_{j=1}^{n} w_{ij} \xi_i' g(z_j, \alpha_0) \qquad (4.76)$$

Using Assumption 4.8, by Lemma D.2 in KTA,

$$\max_{1 \le j \le n} \|g(z_j, \alpha_0)\| = o_p(n^{1/m}) \qquad (4.77)$$

and this $o_p(n^{1/m})$ term does not depend on i, j, or $\alpha_n \in \mathcal{A}_n$. By (4.77) it holds that

$$0 \le 1 + a_i + \lambda_i'(\alpha_0)\, g(z_j, \alpha_0) \le 1 + a_i + \rho_i \|g(z_j, \alpha_0)\| = 1 + a_i + \rho_i o_p(n^{1/m}) \qquad (4.78)$$

Using (4.78) in (4.76) and canceling ρ_i yields

$$\frac{2a_i \sum_{j=1}^{n} w_{ij} \xi_i' g(z_j, \alpha_0)}{1 + a_i + \rho_i o_p(n^{1/m})} + \frac{\rho_i \xi_i' \Sigma_n(x_i, \alpha_0) \xi_i}{1 + a_i + \rho_i o_p(n^{1/m})} \le \sum_{j=1}^{n} w_{ij} \xi_i' g(z_j, \alpha_0) \qquad (4.79)$$

By Corollary D.1 of AC, $\Sigma_n(x_i,\alpha_0) = \Sigma(x_i,\alpha_0) + o_p(1)$ uniformly over $X \in \mathcal{X}$. Using the fact that $\xi_i'\Sigma(x_i,\alpha_0)\xi_i$ is bounded away from zero on $(x_i,\xi_i) \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_g}$, we can divide (4.79) by $\frac{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}{1+a_i+\rho_i o_p(n^{1/m})}$ and rearrange terms to obtain

$$\begin{aligned}\rho_i &\leq \left[1 + a_i + \rho_i o_p(n^{1/m})\right] \frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} - 2a_i \frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} \\ &= (1-a_i)\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} + \rho_i o_p(n^{1/m}) \frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}\end{aligned}$$

and hence

$$\begin{aligned}\rho_i\left(1 - o_p(n^{1/m})\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}\right) &\leq (1-a_i)\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} \\ \rho_i &\leq (1-a_i)\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} \\ &\quad \times \left(1 - o_p(n^{1/m})\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i}\right)^{-1} \qquad (4.80)\end{aligned}$$

For the last term of the RHS of (4.80), using Lemma A.1 and $\left\|\xi_i'\right\| < \infty$ for all i, it holds that

$$\begin{aligned}o_p(n^{1/m})\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} &= o_p(n^{1/m})\left\|\xi_i'\right\| \max_{1\leq i\leq n}\left\|\sum_{j=1}^n w_{ij} g\left(z_j,\alpha_0\right)\right\| \\ &= o_p(n^{1/m})O(1)\left[o_p(\widetilde{\delta}_{1n}) + o_p\left(\frac{1}{n^{\varrho-1/m}}\right)\right] \\ &= o_p(n^{1/m}\widetilde{\delta}_{1n}) + o_p\left(\frac{1}{n^{\varrho-2/m}}\right) \qquad (4.81)\end{aligned}$$

while for the first term of the RHS of (4.80), using also Lemma A.5,

$$\begin{aligned}(1-a_i)\frac{\sum_{j=1}^n w_{ij}\xi_i' g\left(z_j,\alpha_0\right)}{\xi_i'\Sigma_n(x_i,\alpha_0)\xi_i} &= O(1)\left\|\xi_i'\right\| \max_{1\leq i\leq n}\left\|\sum_{j=1}^n w_{ij} g\left(z_j,\alpha_0\right)\right\| \\ &= O(1)O(1)\left[o_p(\widetilde{\delta}_{1n}) + o_p\left(\frac{1}{n^{\varrho-1/m}}\right)\right] \\ &= o_p(\widetilde{\delta}_{1n}) + o_p\left(\frac{1}{n^{\varrho-1/m}}\right) \qquad (4.82)\end{aligned}$$

Under our assumptions, $n^{1/m}\widetilde{\delta}_{1n} \downarrow 0$ and $n^{-\varrho+2/m} \downarrow 0$ in (4.81). This used in (4.80) along with (4.82) and consistency of $\Sigma_n(x_i,\alpha_0)$, implies that

$$\max_{1\leq i\leq n}\|\rho_i\| = o_p(\widetilde{\delta}_{1n}) + o_p\left(\frac{1}{n^{\varrho-1/m}}\right)$$

which yields the desired result by the definition of ρ_i in (4.69). □

Convergence Rates

Lemma B.1. *Consider the functions $G_n(\alpha_n)$ and $\overline{G}_n(\alpha_n)$ defined in (4.38) and (4.61), respectively. Assumptions 4.1-4.3, 4.5, 4.6, 5.1-5.6 imply: (i) $G_n(\alpha_n) - \overline{G}_n(\alpha_n) = o_p(n^{-1/4})$ uniformly over $\alpha_n \in \mathcal{A}_n$; and (ii) $G_n(\alpha_n) - G_n(\alpha_0) - \{\overline{G}_n(\alpha_n) - \overline{G}_n(\alpha_0)\} = o_p(\eta_n n^{-1/4})$ uniformly over $\alpha_n \in \mathcal{A}_n$ with $\|\alpha_n - \alpha_0\|_F \leq o(\eta_n)$, where $\eta_n = n^{-\tau}$ with $\tau \leq 1/4$.*

Proof. This Lemma shows the counterpart of AC's Corollary B.1 for our case. Since $\lambda_i(\alpha_n)$ solves

$$\sum_{j=1}^{n} \frac{w_{ij} g(z_j, \alpha_n)}{\sigma_i + \lambda_i' g(z_j, \alpha_n)} = 0 \tag{4.83}$$

denote by $\lambda_{i0}(\alpha_n)$ the solution to

$$E\left[\frac{g(z_j, \alpha_n)}{\sigma_i + \lambda_i' g(z_j, \alpha_n)} \middle| x_i\right] = 0$$

Lemma A.5 and Assumption 4.5(i) suffice to satisfy the pointwise convergence condition of Lemma 3.3.5 (p. 311) in Van der Vaart and Wellner (1996) (henceforth) VW for the objective function (4.83). Note that $\{g(z, \alpha_n) : \alpha_n \in \mathcal{A}_n\} \subset \Lambda_c^{\overline{\gamma}}(\mathcal{X})$ and $\Lambda_c^{\overline{\gamma}}(\mathcal{X})$ is a Donsker class by Theorem 2.5.6 in VW. Since $\lambda_i(\alpha_n) \in R^{d_g}$, $\{\lambda_i(\alpha_n) : \alpha_n \in \mathcal{A}_n\}$ belongs to the Donsker class. By Example 2.10.8 (p. 192) in VW $\{\lambda_i' g(z, \alpha_n) : \alpha_n \in \mathcal{A}_n\}$ is Donsker. Since $0 < \sigma_i < \infty$ is a data-determined scalar by Lemma A.5, by Example 2.10.9 (p. 192) in VW (4.83) is Donsker in $\alpha_n \in \mathcal{A}_n$. Hence the Assumptions of Lemma 3.3.5 (p. 311) in VW are satisfied and we can invoke Theorem 3.3.1 (p. 310) in VW to conclude that $\|\lambda_i(\alpha_n) - \lambda_{i0}(\alpha_n)\|_E = O_p(n^{-1/2})$, uniformly over $\alpha_n \in \mathcal{A}_n$, for each i. Lemma A.1(A) of AC (defining δ_{1n}) states that $\sum_{j=1}^{n} w_{ij} g(z_j, \alpha_n) - m(x_i, \alpha_n) = o_p(\delta_{1n})$ uniformly over $\mathcal{X} \times \mathcal{A}_n$. These two rate results for $\lambda_i(\alpha_n)$ and $g(z_j, \alpha_n)$, simple law of large numbers for σ_i and continuity of the log function satisfy the satisfy the pointwise convergence condition of Lemma 3.3.5 (p. 311) in VW for the objective function $G_n(\alpha_n)$. By Theorem 2.10.6 (p. 192) in VW $\{\ln[\sigma_i + \lambda_i' g(z_j, \alpha_n)] : \alpha_n \in \mathcal{A}_n\}$ is Donsker. By Lemma A.5, $0 < \sigma_i < \infty$ for each i and thus we can renormalize σ_i by dividing by $\sup_{1 \leq i \leq n} \sigma_i$ that guarantees $\sum_{i=1}^{n} \sigma_i < 1$. By Theorem 2.10.3 (p. 190) in VW

$$\begin{aligned} \left|G_n(\alpha_n) - \overline{G}_n(\alpha_n)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \ln\left\{\sigma_i + \lambda_i' g(z_j, \alpha_n)\right\} - \frac{1}{n}\sum_{i=1}^{n} E\left[\ln\left\{\sigma_i + \lambda_{i0}' g(z, \alpha_n)\right\} | x_i\right]\right| \\ &= O_p(n^{-1/2}) \end{aligned}$$

uniformly over $\alpha_n \in \mathcal{A}_n$, which shows the result (i) in this Lemma.

In order to show part (ii) of the proof, we first derive the counterpart of AC's Corollary A.2 that is a building block for their Corollary B.1 (ii). Note that since $m(X, \alpha_0) = 0$, $\|\alpha_n - \alpha_0\|_F = o_p(1)$ and AC's result $(i.1)$ of the proof of their Corollary A.2 holds also for our $\|m(X, \alpha)\|_E^2$, we only need to show the counterpart of their part $(i.2)$. We replace Assumption 3.9 of AC by our Assumption 5.1 which applies to our metric $\|\cdot\|_F$. This Assumption together with Lemma C.1 imply that

$E\{\|m(X,\alpha)\|_E^2\}$ and $\|\alpha-\alpha_0\|_F^2$ are (topologically) equivalent. Then by Assumptions 4.1, 5.1, and 5.3(i), we have

$$E\left\{\left[\|m(X,\alpha)\|_E^2\right]^2\right\} \leq E\left\{\|m(X,\alpha)\|_E^2\right\} \times \left[\sup_{X,\alpha}\{\|m(X,\alpha)\|_E\}\right]^2 \leq const. \times \|\alpha_n-\alpha_0\|_F^2$$

satisfying part $(i.2)$. Part (ii) of AC's Corollary A.2 holds for our metric $\|\cdot\|_F$ by replacing their Assumption 3.9 with our Assumption 5.1. This, along with AC's Corollary A.1 shows (ii). □

Asymptotic Normality

Lemma C.1. *Under Assumptions 4.1-5.6,*

$$\begin{aligned} & E\left[Var\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}D_{w^*}(Z)\middle| X\right)\right] \\ = \; & E\left\{E\left[D_w(Z)'W_0(Z,X)^{-1}D_w(Z)\middle| X\right]\right\} \\ = \; & E\left\{E\left[D_w(Z)'\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg(Z,\alpha)}\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg(Z,\alpha)}\right)'D_w(Z)\middle| X\right]\right\} \end{aligned}$$

Proof. Using (4.27) and (4.25)

$$\begin{aligned} E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}D_{w^*}(Z)\middle| X\right] &= E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}\frac{dg(Z,\alpha_0)}{d\alpha}\left[v^*\right]\middle| X\right] \\ &= E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\alpha}\left[v^*\right]\middle| X\right] \\ &= E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\theta'}\left(u_\theta^*-\theta_0\right)+\frac{d\varphi\left(X,Z,\alpha_0\right)}{dh}\left[u_h^*-h_0\right]\middle| X\right] \\ &= E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{d\theta'}\middle| X\right]\left(u_\theta^*-\theta_0\right)+E\left[\frac{d\varphi\left(X,Z,\alpha_0\right)}{dh}\left[u_h^*-h_0\right]\middle| X\right] \\ &= 0 \end{aligned}$$

by the definition of α_0. Hence

$$Var\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg\left(Z,\alpha\right)}D_{w^*}(Z)\middle| X\right) = E\left[D_{w^*}(Z)'\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg(Z,\alpha)}\left(\frac{d\varphi\left(X,Z,\alpha_0\right)}{dg(Z,\alpha)}\right)'D_{w^*}(Z)\middle| X\right]$$

Taking expectation over X yields the required result. □

Lemma C.2. *Consider the notation for $v_n(\cdot)$ and $\widetilde{r}[\cdot]$ defined in Appendix 3. Then, under Assumptions 4.1-5.6,*

$$n^{-1/2}v_n\left(\widetilde{r}[\alpha_n-\alpha_0,X,Y]-\widetilde{r}[P_n\alpha^*(a_n,\varepsilon_n)-\alpha_0,X,Y]\right)=o_p(n^{-1/4})$$

Proof. This Lemma performs a similar function as Lemmas C.1 - C.3 in AC. By the definition of $v_n(\cdot)$ and $\widetilde{r}[\cdot]$,

$$\begin{aligned} & n^{-1/2}v_n\left(\widetilde{r}[\alpha_n-\alpha_0,X,Y]-\widetilde{r}[P_n\alpha^*(a_n,\varepsilon_n)-\alpha_0,X,Y]\right) \\ = \; & n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\begin{array}{c} w_{ij}\left\{\widetilde{r}[\alpha_n-\alpha_0,x_i,y_j]-\widetilde{r}[P_n\alpha^*(a_n,\varepsilon_n)-\alpha_0,x_i,y_j]\right\} \\ -E\left\{\widetilde{r}[\alpha_n-\alpha_0,X,Y]-\widetilde{r}[P_n\alpha^*(a_n,\varepsilon_n)-\alpha_0,X,Y]\right\}\end{array}\right) \\ = \; & A_1-A_2 \end{aligned}$$

$$\begin{aligned} A_1 &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\widetilde{r}[\alpha_n-\alpha_0,x_i,y_j]-E\widetilde{r}[\alpha_n-\alpha_0,X,Y] \\ A_2 &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\widetilde{r}[\alpha_n+\varepsilon_nu_n^*-\alpha_0,x_i,y_j]-E\widetilde{r}[\alpha_n+\varepsilon_nu_n^*-\alpha_0,X,Y] \end{aligned}$$

$$\begin{aligned} A_1 &= A_{11}-A_{12} \\ A_{11} &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\varphi\left(x_i,z_j,\alpha\right)-E\varphi\left(z,x,\alpha\right) \\ A_{12} &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\frac{d\varphi\left(x_i,z_j,\alpha_0\right)}{d\alpha}[\alpha-\alpha_0]-E\left\{\frac{d\varphi\left(x,z,\alpha_0\right)}{d\alpha}[\alpha-\alpha_0]\right\} \end{aligned}$$

$$\begin{aligned} A_2 &= A_{21}-A_{22} \\ A_{21} &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\varphi\left(x,z,\alpha_n+\varepsilon_nu_n^*\right)-E\varphi\left(x,z,\alpha_n+\varepsilon_nu_n^*\right) \\ A_{22} &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\frac{d\varphi\left(x_i,z_j,\alpha_0\right)}{d\alpha}[\alpha_n+\varepsilon_nu_n^*-\alpha_0]-E\left\{\frac{d\varphi\left(x,z,\alpha_0\right)}{d\alpha}[\alpha_n+\varepsilon_nu_n^*-\alpha_0]\right\} \end{aligned}$$

The goal is to show $A_{11}-A_{12}-A_{21}+A_{22}=O_p(\varepsilon_n^2)=o_p(n^{-1/4})$. Note that $A_{11}=o_p(n^{-1/4})$ and $A_{21}=o_p(n^{-1/4})$ follows from parts A and B of AC's Lemma A.1. $A_{12}=o_p(n^{-1/4})$ and $A_{22}=o_p(n^{-1/4})$ follows from the rate results for A_{11} and A_{21}, respectively, and the continuous mapping theorem. □

Appendix 4.3: Sieve Conditional Variance Proof

In this Appendix we extend Theorem 1 of Shen (1997) to our conditional case.[6] Consider the setup as in Shen (1997), with the following modifications. Suppose that the observations $\{(X_i, Y_j) : i, j = 1, ..., n\}$ are drawn independently distributed according to density $p(\alpha_0, X_i, Y_j)$.

Define

$$K(\alpha_0, \alpha) = E_0 l(\alpha_0, X_i, Y_j) - E_0 l(\alpha, X_i, Y_j)$$

Let the empirical criterion be

$$L_n(\alpha) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} l(\alpha, X_i, Y_j)$$

where $l(\alpha, Y_j, X_i)$ is the criterion based on a single observation. Consider $l(\alpha, x, y)$ for which (*analog of Shen's (4.1)*)

$$\widetilde{r}[\alpha - \alpha_0, x, y] = l(\alpha, x, y) - l(\alpha_0, x, y) - l'_{\alpha_0}[\alpha - \alpha_0, x, y] \tag{S 4.1}$$

where $l'_{\alpha_0}[\alpha - \alpha_0, x, y]$ is defined as $\lim_{t \to 0}[l(a + t[\alpha - \alpha_0], x, y) - l(\alpha_0, x, y)]/t$. Denote $\widehat{\alpha}_n$ the maximizer of $L_n(\alpha_n)$ over $\alpha_n \in \mathcal{A}_n$. We estimate a real functional of $\widehat{\alpha}_n$ denoted as $f(\alpha)$. With $\widehat{\alpha}_n$ as defined, $f(\alpha)$ is estimated by a substitution estimate $f(\widehat{\alpha}_n)$. By the definition of $\widehat{\alpha}_n$, we have (*analog of Shen's (2.1)*)

$$L_n(\widehat{\alpha}_n) \geq \sup_{\alpha \in \mathcal{A}_n} L_n(\alpha_n) - O(\varepsilon_n^2) \tag{S 2.1}$$

where $\varepsilon_n^2 \to 0$ as $n \to \infty$. For any generic function $g(X, Y)$ let

$$\nu_n(g) = n^{-1} \sum_{i=1}^{n} n^{1/2} \left\{ \sum_{j=1}^{n} w_{ij} g(X_i, Y_j) - E\left[g(X, Y) \middle| X = x_i \right] \right\}$$

be the empirical process induced by g. Let the convergence rate of the sieve estimate under $\|\cdot\|$ be $o_p(\delta_n)$ and let $\varepsilon_n^2 = o_p(n^{-1/2})$.

The following conditions are modified versions of Shen (1997)'s (p. 2568) conditions:

Condition A (Stochastic Equicontinuity). *For $\widetilde{r}[\alpha - \alpha_0, x, y]$ defined in (S 4.1),*

$$\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| \leq \delta_n\}} n^{-1/2} \nu_n \left(\widetilde{r}[\alpha_n - \alpha_0, X, Y] - \widetilde{r}[\alpha_n + \varepsilon_n u_n^* - \alpha_0, X, Y] \right) = O_p(\varepsilon_n^2)$$

Condition B (Expectation of Criterion Difference).

$$\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| \leq \delta_n\}} \left[K\left(\alpha_0, \alpha_n + \varepsilon_n u_n^*\right) - K\left(\alpha_0, \alpha_n\right) \right] - \frac{1}{2} \left[\|\alpha_n + \varepsilon_n u^* - \alpha_0\|^2 - \|\alpha_n - \alpha_0\|^2 \right] = O_p(\varepsilon_n^2)$$

[6] Measurability with respect to the underlying probability space is assumed throughout the paper and hence we do not distiguish outer expectation from the usual one.

Condition C (Approximation Error).

$$\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| \leq \delta_n\}} \|\varepsilon_n u^* - \varepsilon_n u_n^*\| = O_p(\delta_n^{-1}\varepsilon_n^2)$$

In addition,

$$\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| \leq \delta_n\}} n^{-1/2}\nu_n \left(l'_{\alpha_0}[\varepsilon_n u^* - \varepsilon_n u_n^*, X, Y]\right) = O_p(\varepsilon_n^2)$$

Condition D (Gradient).

$$\sup_{\{\alpha_n \in \mathcal{A}_n : \|\alpha_n - \alpha_0\| \leq \delta_n\}} n^{-1/2}\nu_n \left(l'_{\alpha_0}[\alpha_n - \alpha_0, X, Y]\right) = O_p(\varepsilon_n)$$

Condition E (Smoothness).

Suppose the functional f has the following smoothness property: for any $\alpha_n \in \mathcal{A}_n$

$$\left|f_{\alpha_n} - f_{\alpha_0} - f'_{\alpha_0}[\alpha_n - \alpha_0]\right| \leq u_n \|\alpha_n - \alpha_0\|_F^{\omega} \tag{S 4.2}$$

as $\|\alpha_n - \alpha_0\|_F \to 0$ where ω is the degree of smoothness of $f'_{\alpha_0}[\alpha_n - \alpha_0]$ at α_0.

Condition F (Convergence Rates and Smoothness). $u_n \delta_n^{\omega} = O_p(n^{-1/2})$.

Condition G (Variance). *$Var\left(l'_{\alpha_0}[v^*, X, Y]\right) < \infty$ is positive definite for all $X \in \mathcal{X}$, $y \in \mathcal{Y}$.*

Theorem 7.1. *Let the Conditions A-G hold. Then for the approximate substitution sieve estimate defined in (S 2.1),*

$$n^{-1/2}(f(\widehat{\alpha}_n) - f(\alpha_0)) \xrightarrow{d} N(0, E\left[Var\left(l'_{\alpha_0}[v^*, Y]\right)|X\right])$$

Proof of Theorem 7.1. Rearrange (S 4.1) as

$$l(\alpha, x, y) = \widetilde{r}[\alpha - \alpha_0, x, y] + l(\alpha_0, x, y) + l'_{\alpha_0}[\alpha - \alpha_0, x, y]$$

Subtract from (S 4.1) its expectation (under $P(\theta_0, X_i, Y_j)$ denoted by E_0), for a given (X_i, Y_j) to obtain

$$\begin{aligned} l(\alpha, x_i, y_j) - E_0 l(\alpha, x_i, y_j) \quad = \quad & l(\alpha, x_i, y_j) - E_0 l(\alpha, x_i, y_j) \\ & + l'_{\alpha_0}[\alpha - \alpha_0, x_i, y_j] - E_0 l'_{\alpha_0}[\alpha - \alpha_0, x_i, y_j] \\ & + \widetilde{r}[\alpha - \alpha_0, x_i, y_j] - E_0 \widetilde{r}[\alpha - \alpha_0, x_i, y_j] \end{aligned}$$

rearrange

$$\begin{aligned} l(\alpha, x_i, y_j) \quad = \quad & l(\alpha, x_i, y_j) - [E_0 l(\alpha, x_i, y_j) - E_0 l(\alpha, x_i, y_j)] \\ & + l'_{\alpha_0}[\alpha - \alpha_0, x_i, y_j] - E_0 l'_{\alpha_0}[\alpha - \alpha_0, x_i, y_j] \\ & + \widetilde{r}[\alpha - \alpha_0, x_i, y_j] - E_0 \widetilde{r}[\alpha - \alpha_0, x_i, y_j] \end{aligned}$$

take a weighted average over i, j with weights w_{ij}

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} l(\alpha, x_i, y_j) &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} l(\alpha_0, x_i, y_j) \\ &\quad -n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left[E_0 l(\alpha_0, x_i, y_j) - E_0 l(\alpha, x_i, y_j)\right] \\ &\quad +n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(l'_{\alpha_0}[\alpha-\alpha_0, x_i, y_j] - E_0 l'_{\alpha_0}[\alpha-\alpha_0, x_i, y_j]\right) \\ &\quad +n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(\widetilde{r}[\alpha-\alpha_0, x_i, y_j] - E_0\widetilde{r}[\alpha-\alpha_0, x_i, y_j]\right) \end{aligned}$$

and hence using the notation above, for any $P_n\alpha_n \in \{P_n\alpha_n \in \mathcal{A}_n : \|P_n\alpha_n - \alpha_0\| \le \delta_n\}$, we have

$$\begin{aligned} L_n(P_n\alpha_n) &= L_n(a_0) - K(\alpha_0, P_n\alpha_n) \\ &\quad +n^{-1/2}\nu_n(l'_{\theta_0}[P_n\alpha_n - \alpha_0, X, Y]) \\ &\quad +n^{-1/2}\nu_n(r[P_n\alpha_n - \alpha_0, X, Y]) \end{aligned} \tag{S 9.1}$$

Substituting $P_n\alpha_n$ by $\widehat{\alpha}_n$ here above, we obtain

$$\begin{aligned} L_n(\widehat{\alpha}_n) &= L_n(a_0) - K(\alpha_0, \widehat{\alpha}_n) \\ &\quad +n^{-1/2}\nu_n(l'_{\theta_0}[\widehat{\alpha}_n - \alpha_0, X, Y]) \\ &\quad +n^{-1/2}\nu_n(r[\widehat{\alpha}_n - \alpha_0, X, Y]) \end{aligned} \tag{S 9.2}$$

Subtracting (S 9.2) from (S 9.1) and substituting α_n by $\alpha^*(\widehat{\alpha}_n, \varepsilon_n)$ in (S 9.1), we have

$$\begin{aligned} & L_n(P_n\alpha^*(\widehat{\alpha}_n, \varepsilon_n)) - L_n(\widehat{\alpha}_n) \\ = \; & L_n(\alpha_0) - L_n(\alpha_0) \\ & -K(\theta_0, P_n\alpha^*(\widehat{\alpha}_n, \varepsilon_n) + K(\alpha_0, \widehat{\alpha}_n) \\ & +n^{-1/2}\nu_n(l'_{\alpha_0}[P_n\alpha^*(\widehat{\alpha}_n, \varepsilon_n) - \alpha_0, X, Y]) - n^{-1/2}\nu_n(l'_{\alpha_0}[\widehat{\alpha}_n - \alpha_0, X, Y]) \\ & +n^{-1/2}\nu_n(r[P_n\alpha^*(\widehat{\alpha}_n, \varepsilon_n) - \alpha_0, X, Y]) - n^{-1/2}\nu_n(r[\widehat{\alpha}_n - \alpha_0, X, Y]) \end{aligned}$$

which yields

$$\begin{aligned} L_n(\widehat{\alpha}_n) &= L_n(P_n\alpha^*(\widehat{\alpha}_n, \varepsilon_n)) \\ &\quad -[K(\alpha_0, \widehat{\alpha}_n) - K(\theta_0, P_n\alpha^*(\widehat{\alpha}_n, \varepsilon_n)] \\ &\quad +n^{-1/2}\nu_n(l'_{\alpha_0}[\widehat{\alpha}_n - P_n\alpha^*(\widehat{\alpha}_n, \varepsilon_n), X, Y]) \\ &\quad +n^{-1/2}\nu_n(r[\widehat{\alpha}_n - P_n\alpha^*(\widehat{\alpha}_n, \varepsilon_n), X, Y]) \end{aligned}$$

By Condition A (second line of the following)

$$\begin{aligned} & n^{-1/2}\nu_n(r[P_n\alpha^*(\widehat{\alpha}_n, \varepsilon_n) - \alpha_0, X, Y]) - n^{-1/2}\nu_n(r[\widehat{\alpha}_n - \alpha_0, X, Y]) \\ = \; & n^{-1/2}\nu_n(r[\widehat{\alpha}_n - P_n\alpha^*(\widehat{\alpha}_n, \varepsilon_n), X, Y]) \\ = \; & O_p(\varepsilon_n^2) \end{aligned}$$

Using Condition B on the difference in Ks, we obtain

$$\begin{aligned} L_n(\widehat{\alpha}_n) &= L_n(P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n)) - \frac{1}{2}\left[\|\widehat{\alpha}_n-\alpha_0\|^2 - \|P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n)-\alpha_0\|^2\right] \\ &\quad +n^{-1/2}\nu_n(l'_{\alpha_0}[\widehat{\alpha}_n - P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n),X,Y]) \\ &\quad +O_p(\varepsilon_n^2) \end{aligned}$$

By Condition C (applicable to the second line)

$$\|P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n)-\alpha^*(\widehat{\alpha}_n,\varepsilon_n)\| = O(\delta_n^{-1}\varepsilon_n^2)$$

Hence, using (S 2.1) we have

$$\begin{aligned} -O(\varepsilon_n^2) &\le -\frac{1}{2}\left[\|\widehat{\alpha}_n-\alpha_0\|^2 - \|P_n\alpha^*(\widehat{\alpha}_n,\varepsilon_n)-\alpha_0\|^2\right] \\ &\quad +n^{-1/2}\nu_n(l'_{\alpha_0}[\widehat{\alpha}_n - \alpha^*(\widehat{\alpha}_n,\varepsilon_n),X,Y]) \\ &\quad +O_p(\varepsilon_n^2) \end{aligned} \tag{S 9.3}$$

We will use the relation

$$\begin{aligned} \widehat{\alpha}_n - \alpha^*(\widehat{\alpha}_n,\varepsilon_n) &= \widehat{\alpha}_n - \widehat{\alpha}_n + \varepsilon_n\widehat{\alpha}_n - \varepsilon_n u^* - \varepsilon_n\alpha_0 \\ &= -\varepsilon_n\left(u^* - (\widehat{\alpha}_n-\alpha_0)\right) \end{aligned}$$

in $\nu_n(l'_{\alpha_0}[\widehat{\alpha}_n - \alpha^*(\widehat{\alpha}_n,\varepsilon_n),X,Y])$ to get $-\nu_n(l'_{\alpha_0}[\varepsilon_n\left(u^* - (\widehat{\alpha}_n-\alpha_0)\right),X,Y])$.

In (S 9.3) we have

$$\begin{aligned} \|P_n a^*(\widehat{a}_n,\varepsilon_n)-a_0\|^2 &= \|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)+\alpha^*(\widehat{a}_n,\varepsilon_n)-\theta_0\|^2 \\ &= \|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)+(1-\varepsilon_n)(\widehat{\alpha}_n-\alpha_0)+\varepsilon_n u^*\|^2 \\ &\le \|(1-\varepsilon_n)(\widehat{\alpha}_n-\alpha_0)\|\,\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)+\varepsilon_n u^*\| \\ &\le \|(1-\varepsilon_n)(\widehat{\alpha}_n-\alpha_0)\|\,\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\| \\ &\quad +\|(1-\varepsilon_n)(\widehat{\alpha}_n-\alpha_0)\|\,\|\varepsilon_n u^*\| \\ &= (1-\varepsilon_n)\,\|(\widehat{\alpha}_n-\alpha_0)\|\,\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\| \\ &\quad +(1-\varepsilon_n)\,\langle\widehat{\alpha}_n-\alpha_0,\varepsilon_n u^*\rangle \end{aligned}$$

We multiply $\|\widehat{a}_n-\alpha_0\|$ by the factor

$$\begin{aligned} 1-(1-\varepsilon_n)^2 &= 1-(1-2\varepsilon_n+\varepsilon_n^2) \\ &= 2\varepsilon_n-\varepsilon_n^2 \end{aligned}$$

which is a positive fraction that preserves the inequality. We also multiply $\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\theta_0\|^2$ by 2 which also preserves the inequality. Hence we obtain

$$\begin{aligned}-O(\varepsilon_n^2) \quad \leq \quad & -\frac{1}{2}\left[1-(1-\varepsilon_n)^2\right]\|\widehat{\alpha}_n-\alpha_0\|^2 \\ & +(1-\varepsilon_n)\,\|(\widehat{\alpha}_n-\alpha_0)\|\,\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\| \\ & +(1-\varepsilon_n)\,\langle\widehat{\alpha}_n-\alpha_0,\varepsilon_n u^*\rangle \\ & -n^{-1/2}\nu_n(l'_{\alpha_0}[\varepsilon_n\,(u^*-(\widehat{\alpha}_n-\alpha_0))\,,X,Y]) \\ & +O_p(\varepsilon_n^2)\end{aligned}$$

Adding $\varepsilon_n\,\|(\widehat{\alpha}_n-\alpha_0)\|\,\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\|$ still preserves the inequality. For the first line, $\varepsilon_n^2\,\|\widehat{\alpha}_n-\alpha_0\|^2=O_p(\varepsilon_n^2)$. Hence

$$\begin{aligned}-O(\varepsilon_n^2) \quad \leq \quad & -\varepsilon_n\,\|\widehat{\alpha}_n-\alpha_0\|^2+\|(\widehat{\alpha}_n-\alpha_0)\|\,\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\| \\ & +(1-\varepsilon_n)\,\langle\widehat{\alpha}_n-\alpha_0,\varepsilon_n u^*\rangle-n^{-1/2}\nu_n(l'_{\alpha_0}[\varepsilon_n\,(u^*-(\widehat{\alpha}_n-\alpha_0))\,,X,Y])+O_p(\varepsilon_n^2)\end{aligned}$$

Note that

$$\begin{aligned}-\varepsilon_n\,\|\widehat{\alpha}_n-\alpha_0\|^2 \quad &= \quad O_p(\varepsilon_n)o_p(\delta^2) \\ &= \quad o_p(\delta^2)\end{aligned}$$

By Condition C

$$\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\|=O_p(\delta^{-1}\varepsilon_n^2)$$

since

$$\|\widehat{\alpha}_n-\alpha_0\|=o_p(\delta)$$

then

$$\begin{aligned}\|\widehat{\alpha}_n-\alpha_0\|\,\|P_n\alpha^*(\widehat{a}_n,\varepsilon_n)-\alpha^*(\widehat{a}_n,\varepsilon_n)\| \quad &= \quad o_p(\delta)O_p(\delta^{-1}\varepsilon_n^2) \\ &= \quad o_p(\varepsilon_n^2)\end{aligned}$$

and using Conditions C and D

$$n^{-1/2}\nu_n(l'_{\alpha_0}[\varepsilon_n\,(u^*-(\widehat{\alpha}_n-\alpha_0))\,,X,Y])=n^{-1/2}\nu_n(l'_{\alpha_0}[u^*,X,Y])+O_p(\varepsilon_n^2)+O_p(\varepsilon_n^2)$$

Hence

$$-(1-\varepsilon_n)\,\langle\widehat{\alpha}_n-\alpha_0,u^*\rangle+n^{-1/2}\nu_n(l'_{\alpha_0}[u^*,X,Y])=o_p(n^{-1/2}) \tag{S 9.4}$$

This gives, together with the inequality in (S 9.4) with u^* replaced by $-u^*$,

$$\left|\langle\widehat{\alpha}_n-\alpha_0,u^*\rangle-n^{-1/2}\nu_n(l'_{\alpha_0}[u^*,X,Y])\right|=o_p(n^{-1/2})$$

so

$$\langle\widehat{\alpha}_n-\alpha_0,v^*\rangle=n^{-1/2}\nu_n(l'_{\alpha_0}[v^*,X,Y])+o_p(n^{-1/2})$$

Hence, by (S 4.2)

$$\begin{aligned} f_{\alpha_n} - f_{\alpha_0} &= f'_{\alpha_0}[\alpha_n - \alpha_0] + o_p(u_n \|\alpha_n - \alpha_0\|_F^{\omega}) \\ &= \langle \widehat{\alpha}_n - \alpha_0, v^* \rangle + o_p(n^{-1/2}) \\ &= n^{-1/2} \nu_n(l'_{\alpha_0}[u^*, X, Y]) + o_p(n^{-1/2}) \\ &= n^{-1} \sum_{i=1}^{n} n^{1/2} \left\{ \sum_{j=1}^{n} w_{ij} l'_{\alpha_0}[u^*, X_i, Y_j] - E\left[l'_{\alpha_0}[u^*, X, Y] \middle| X = x_i \right] \right\} \end{aligned}$$

The result then follows from the Central Limit Theorem (CLT) for triangular arrays (Proposition) in Andrews (1994, p. 2251). Note that the conditions of the Proposition are satisfied under our assumptions. In particular, $\Theta \subseteq \mathbb{R}^{d_\theta}$ is compact, finite-dimensional convergence of $n^{1/2} \sum_{j=1}^{n} w_{ij} l'_{\alpha_0}[u^*, X_i, Y_j] - E\left[l'_{\alpha_0}[u^*, X, Y] \middle| X = x_i \right]$ holds for each x_i due to the classical Lindeberg-Levy CLT, and Condition A satisfies the stochastic equicontinuity requirement of the Proposition.

□

5.0 BIBLIOGRAPHY

Ai, C. (1997). A semiparametric maximum likelihood estimator. *Econometrica 65*, 933–963.

Ai, C., A. Chatrath, and F. Song (2006). On the comovement of commodity prices. *American Journal of Agricultural Economics 88*(3), 574–588.

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*(422), 669–679.

Altonji, J. G. and L. M. Segal (1996, July). Small-sample bias in gmm estimation of covariance structures. *Journal of Business & Economic Statistics 14*(3), 353–66.

Andrews, D. W. K. (1994, May). Empirical process methods in econometrics. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4 of *Handbook of Econometrics*, Chapter 37, pp. 2248–2294. Elsevier.

Antoine, B., H. Bonnal, and E. Renault (2006). On the efficient use of the informational content of estimating equations: Implied probabilities and maximum euclidean likelihood. forthcoming in the *Journal of Econometrics*.

Bauwens, L. and N. Hautsch (20003). Dynamic latent factor models for intensity processes. Manuscript.

Bauwens, L., M. Lubrano, and J.-F. Richard (1999). *Bayesian Inference in Dynamic Econometric Models*. Oxford: Oxford University Press.

Bertschek, I. (1995). Product and process innovation as a response to increasing import and foreign direct investment. *Journal of Industrial Economics 43*(4), 341–57.

Bertschek, I. and M. Lechner (1998). Convenient estimators for the panel probit model. *Journal of Econometrics 87*(2), 329–371.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer-Verlag.

Blundell, R., X. Chen, and D. Kristensen (2006). Semi-nonparametric iv estimation of shape invariant engel curves. working paper, New York University.

Börsch-Supan, A. and V. A. Hajivassiliou (1993). Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics 58*(3), 347–368.

Börsch-Supan, A., V. A. Hajivassiliou, L. J. Kotlikoff, and J. N. Morris (1992). Health, children, and elderly living arrangements: a multiperiod-multinomial probit model with unobserved heterogeneity and autocorrelated errors. In W. D. A. (Ed.), *Topics in the Economics of Aging*. The University of Chicago Press.

Borwein, J. M. and A. S. Lewis (2006). *Convex Analysis and Nonlinear Optimization: Theory and Examples* (Second ed.). New York, NY: Springer.

Butler, J. S. and R. Moffitt (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica 50*(3), 761–764.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*, 305–334.

Chen, X. (2005). Large sample sieve estimation of semi-nonparametric models. Technical report, Department of Economics, New York University.

Chen, X. and S. Ludvigson (2006). Land of addicts? an empirical investigation of habit-based asset pricing models. Technical report, Department of Economics, New York University.

Chen, X. and X. Shen (1998, March). Sieve extremum estimates for weakly dependent data. *Econometrica 66*(2), 289–314.

Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika 85*(4), 967–972.

Csiszar, I. (1967). On topological properties of f-divergences. *Studia Scientriarum Mathematicarum Hungaria 2*, 329–339.

Danielsson, J. and J.-F. Richard (1993). Accelerated gaussian importance sampler with application to dynamic latent variable models. *Journal of Applied Econometrics 8*, 153–73.

Domínguez, M. A. and I. N. Lobato (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica 72*(5), 1601–1615.

Duncan, G. M. (1986, June). A semi-parametric censored regression estimator. *Journal of Econometrics 32*(1), 5–34.

Falcetti, E. and M. Tudela (2006). modelling currency crises in emerging markets: a dynamic probit model with unobserved heterogeneity and autocorrelated errors. *Oxford Bulleting of Economics and Statistics 68*(4), 445–471.

Gallant, A. R. and D. W. Nychka (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica 55*(2), 363–390.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, pp. 571–78. ASA.

Geweke, J. and M. Keane (2001, May). Computationally intensive methods for integration in econometrics. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5 of *Handbook of Econometrics*, Chapter 56, pp. 3463–3568. Elsevier.

Gourieroux, C. and A. Monfort (1996). *Simulation-based econometric methods*. Oxford University Press.

Greene, W. (2004). Convenient estimators for the panel probit model: Further results. *Empirical Economics 29*(1), 21–47.

Grenander, U. (1981). *Abstract Inference*. New York: Wiley.

Hajivassiliou, V. A. and D. L. McFadden (1998, July). The method of simulated scores for the estimation of ldv models. *Econometrica 66*(4), 863–896.

Hall, P. and J. L. Horowitz (1996, July). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica 64*(4), 891–916.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–54.

Hansen, L. P., J. Heaton, and A. Yaron (1996, July). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics 14*(3), 262–80.

Heckman, J. J. (1981). Statistical models for discrete panel data. In M. C. F. and M. D. (Eds.), *Structural Analysis of Discrete Data with Econometrics Applications*. The MIT Press, Cambridge, Masschusetts; London, England.

Hyslop, D. R. (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica 67*(6), 1255–1294.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Econometrica 58*, 71–120.

Imbens, G. H., R. Spady, and P. Johnson (1998). Information theoretic approaches to inference in moment condition models. *Econometrica 66*, 333–357.

Imbens, G. W. (1997, July). One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies 64*(3), 359–83.

Inkmann, J. (2000). Misspecified heteroskedasticity in the panel probit model: A small sample comparison of gmm and sml estimators. *Journal of Econometrics 97*(2), 227–259.

Jaynes, E. T. (1957). Information theory and statistical mechanics i. *Physical Review 106*(4), 620–630.

Jung, R. C. and R. Liesenfeld (2001). estimating time series models for count data using efficient importance sampling. *Allgemeines Statistisches Archiv 85*, 387–407.

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica 62*(1), 95–116.

Kitamura, Y. (2006). Empirical likelihood methods in econometrics: Theory and practice. Cowles Foundation Discussion Paper No. 1569.

Kitamura, Y. and M. Stutzer (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica 65*(4), 861–874.

Kitamura, Y., G. Tripathi, and H. Ahn (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica 72*(6), 1667–1714.

Kullback, S. (1997). *Information Theory and Statistic* (Dover ed.). Mineola, NY: Dover Publications, Inc.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*(1), 79–86.

LeBlanc, M. and J. Crowley (1995). Semiparametric regression functionals. *Journal of the American Statistical Association 90*, 95–105.

Lee, L.-F. (1997). Simulated maximum likelihood estimation of dynamic discrete choice statistical models some monte carlo results. *Journal of Econometrics 82*(1), 1–35.

Liesenfeld, R. and J.-F. Richard (2003). Univariate and multivariate stochastic volatility models: estimation and diagnostics. *Journal of Empirical Finance 10*(4), 505–531.

Liesenfeld, R. and J.-F. Richard (2006a). Classical and bayesian analysis of univariate and multivariate stochastic volatility models. *Econometric Reviews 25*(2), 1–26.

Liesenfeld, R. and J.-F. Richard (2006b). Simulation techniques for panels: Efficient importance sampling. In L. Matyas and P. Sevestre (Eds.), *Econometrics of Panel Data* (3rd ed.). Boston: Kluwer. forthcoming.

Luenberger, D. G. (1969). *Optimization by vector space methods*. New York, NY: Wiley.

McFadden, D. (1989, September). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica 57*(5), 995–1026.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica 59*(4), 1161–1167.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In G. Maddala, C. Rao, and H. Vinod (Eds.), *Handbook of Statistics*, Volume 11, pp. 2111–2245. Amsterdam: Elsevier.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–82.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Newey, W. K. and R. J. Smith (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica 72*(1), 219–255.

Nishiyama, Y., Q. Liu, and N. Sueishi (2005, December). Semiparametric estimators for conditional moment restrictions containing nonparametric functions: Comparison of gmm and empirical likelihood procedures. In A. Zerger and R. Argent (Eds.), *MODSIM 2005 In-*

ternational Congress on Modelling and Simulation, pp. 170–176. Modelling and Simulation Society of Australia and New Zealand. ISBN: 0-9758400-2-9.

Otsu, T. (2003a). Empirical likelihood for quantile regression. Manuscript.

Otsu, T. (2003b). Penalized empirical likelihood estimation of conditional moment restriction models with unknown functions. Department of Economics, University of Wisconsin-Madison.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*(2), 237–49.

Owen, A. (2001). *Empirical Likelihood*. Chapman and Hall/CRC.

Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge University Press.

Powell, J. L. (1994, May). Estimation of semiparametric models. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4 of *Handbook of Econometrics*, Chapter 41, pp. 2443–2521. Elsevier.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*(6), 1403–30.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics 22*(1), 300–325.

Ramalho, J. (2005). Small sample bias of alternative estimation methods for moment condition models: Monte carlo evidence for covariance structures. *Studies in Nonlinear Dynamics & Econometrics 9*(1).

Ramsey, J. B. (1999). The contribution of wavelets to the analysis of economic and financial data. *Phil. Trans. R. Soc. Lond. A 357*, 2593–2606.

Richard, J.-F. and W. Zhang (2000). Accelerated monte carlo integration: An application to dynamic latent variable models. In R. Mariano, M. Weeks, and T. Schuermann (Eds.),

Simulation-Based Inference in Econometrics: Methods and Applications, pp. 47–70. Cambridge: Cambridge University Press.

Richard, J.-F. and W. Zhang (2006). Efficient high-dimensional importance sampling. Manuscript.

Robinson, P. M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica 55*, 875–891.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–54.

Royden, H. L. (1987). *Real Analysis* (third ed.). Englewood Cliffs, NJ: Prentice Hall.

Schennach, S. M. (2006). Point estimation with exponentially tilted empirical likelihood. Technical report, University of Chicago. forthcoming in the Annals of Statistics.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*. Reprinted in the ACM SIGMOBILE Mobile Computing and Communications Review, Vol.5(1) (January 2001).

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics 25*(6), 2555–2591.

Shen, X. and W. H. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics 22*(2), 580–615.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Annals of Statistics 12*(3), 898–916.

Smith, R. J. (1997, March). Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *Economic Journal 107*(441), 503–19.

Smith, R. J. (2003). Local gel estimation with conditional moment restrictions. Technical report, University of Warwick.

Smith, R. J. (2005). Local gel methods for conditional moment restrictions. Cemmap working paper cwp15/05, University of Warwick.

Smith, R. J. (2006). Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics*. forthcoming.

Tibshirani, R. and T. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association 82*(398), 559–567.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics 22*, 1701–1762.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge University Press.

Van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag New York, Inc.

Vella, F. and M. Verbeek (1998). Whose wages do unions raise? a dynamic model of unionism and wage rate determination for young men. *Journal of Applied Econometrics 13*(2), 163–183.

Wong, W. H. and T. A. Severini (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. *The Annals of Statistics 19*(2), 603–632.

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Zhang, J. and I. Gijbels (2003). Sieve empirical likelihood and extensions of the generalized least squares. *Scandinavian Journal of Statistics 30*, 1–24.

Zhang, W. and L. F. Lee (2004, 06). Simulation estimation of dynamic discrete choice panel models with accelerated importance samplers. *Econometrics Journal* 7(1), 120–142.