Model Averaging PartiaL Effect (MAPLE) Estimation with Large Dimensional Data

Yundong Tu*
Department of Economics
University of California, Riverside

October 2011

Abstract

This paper studies the estimation of the marginal effect of one economic variable on another in the presence of large amount of other economic variables—a problem frequently faced by applied researchers. The estimation is motivated via model uncertainty so that random components should be included to describe the economy according to the state of the world. A condition named "Conditional Mean Independence" is shown to be sufficient to identify the partial effect parameter of interest. In the case that the parameter of interest can be identified in more than one approximating model, we propose two estimators for such a parameter: generalized-method-of-moment-based model averaging partial effect (gMAPLE) estimator and entropy-based model averaging partial effect (eMAPLE) estimator. Consistency and asymptotic normality of the MAPLE estimators are established under a suitable set of conditions. Thorough simulation studies reveal that MAPLE estimators outperform factor based, variable selection based and other model averaging estimators available in the literature. An economic example is taken to illustrate the use of MAPLE estimator to evaluate the effect of inherited control on firms' performance.

Key Words: Partial Effect; Treatment Effect; Model Averaging; Bayesian Model Averaging; Jackknife Model Averaging; FOGLeSs; Variable Selection; Factor Models; Inherited Control.

JEL Classification: C13; C51; C63; L25; M13.

*Department of Economics, University of California, Riverside, CA 92521, E-mail: yundong.tu@email.ucr.edu . The author would like to thank Tae-Hwy Lee, Aman Ullah, Jerry Hausman, Victoria Zinde-Walsh, Amos Golan, Halbert White, Raffaella Giacomini, Andres Santos, Karim Chalak, Bruce Hansen, Jim Powell, Hong Han, Shu Shen, Li Gan for helpful discussions, participants of Midwest Econometrics Conference 2011, California Econometrics Conference and seminar participants at UC, Riverside for valuable comments.

1 Introduction

We live in a world full of valuable information recorded by thousands of economic and financial variables. Economic researchers, policy makers and financial analysts are faced with these overwhelming economic signals. In theoretical macroeconomics, agents are forced to process all available quantities when they form expectations for future. In program evaluation, experts incorporate individual features such as gender, education, marriage status, family size, health status, etc. to analyze the treatment effect. In labor economics, newly available sources of data are called forth to advance theory and inform policy. In finance, equity premium is studied with thousands of financial variables, indices and macro policy variables.

This paper, to our knowledge, serves as the first work to study the marginal effect of one variable on another in the large dimensional data setting, with the use of model averaging. This problem is typical in any field of economics, since artistic economic theory would suggest plenty of variables that would be potentially related to the variable of interest (Sala-i-Martin et al 2006). When it comes to estimation of such partial effect, the omission of other variables from the model would lead to biased estimate, fallible inference and result in misleading policy recommendation. In the following subsections, we first make clear the problem of estimation in the presence of large dimensional data, then review the related literature and finally spell out the contributions of the paper.

1.1 Large Dimensional Data v.s. Small Models

With the advancement of computer technology, economic and financial data are more easily collected, shared and utilized in studies. Resources for Economists on the Internet[1] provides a wide range of economic topics with links to many different data sources. National Bureau of Economic Research[2] provides links to various data sources including macro data, industrial data, hospital data, demographic and vital statistics, patent and scientific papers data, and so forth. Penn World Table[3] provides purchasing power parity and national income accounts converted to international prices for 188 countries for years 1950-2004. In finance and business, Datastream by Thompson Financial[4] and Wharton research Data Services[5] provide researchers worldwide with instant access to financial and marketing series. Yahoo[6] and the Federal Reserve Bank of St Louis maintain free data access to a wide variety of financial time series.

Economic models are introduced and estimated to analyze the linkage among economic variables and characterize the relationship of interest. With the principle of parsimony, researchers usually start with small models that focus on salient features of economic phenomena. For

[1] http://rfe.org/

[2] http://www.nber.org/data/

[3] http://pwt.econ.upenn.edu/

[4] http://www.thomsonone.com/

[5] http://wrds-web.wharton.upenn.edu/wrds/

[6] http://finance.yahoo.com/

example, Keynes (1936) hypothesized that the major influence on individual consumption is personal income; Phillips (1958) and numerous work afterwards described an inverse relationship between money wage changes and unemployment in British economy; Mincer (1976) studied the direction of labor mobility resulting from minimum-wage imposition; Ashenfelter (1978) attribute current earning to past earnings and job training.

While these models are argued to explain economic phenomena, economists usually implicitly or explicitly require the environment under investigation hold *ceteris paribus*. This superiority is appreciated together with Occam's Razor in all scientific exploration. However, such a parsimony principle is better interpreted as a heuristic rather than an irrefutable principle of logic (Gernert, 2007). It has been maintained in economic modeling for mainly two reasons: First, analyzing the full model with all available economic variables would result in difficulties in parameter identification, estimation and model evaluation, driving us astray from the economic analysis originally designated.[7] The second reason that leads to simple models is that economics is more complex than it appears. Modeling methods available in mainstream science aim to separate important linkage from abounded noisy signals. This intrinsic feature limits inference in the presence of large data sets.

1.2 Related Literature

Dimensionality reduction techniques have been proposed and frequently used in forecasting literature when large dimensional data are present.The first line of research assumes that the data is generated by some underlying factors of smaller dimension and approaches the estimation of the common factors in a way fitting the problem at hand.[8] For recent work in this direction, see Bai and Ng (2010) and references therein. Another direction to achieve dimensionality reduction is variable selection. Selection is conducted by minimizing some objective loss functions, such as Akaike information criterion (AIC) or Bayesian information criterion (BIC). Early examples are forward variable selection, backward selection and stepwise selection etc. (Miller 2002). More recently the literature is overwhelmed by more sophisticated methods.[9] See Fan and Lv (2010) for a review.

Though popular in forecasting literature, dimensionality reduction methods have their own limitations when applying to partial effect estimation. In factor analysis, partial effect param-

[7] With advancement in fuzzy analysis, set identification and inference has achieved significant progress. In economic applications, see Manski (1995, 2003, 2007), Imbens and Manski (2004), Santos (2011), Romano and Shaikh (2008, 2010), to name a few. Inference with large dimensional data is still left open.

[8] Popular examples are Principal Component Analysis (PCA) invented by Pearson (1901), factor analysis pioneered by Spearman (1904), Partial Least Square (PLS) developed by Wold (1966), Principal Covariate Regression (PCovR) proposed by De Jong and Kiers (1992), Supervised Factor Model (SFM) introduced by Tu and Lee (2011), and so forth.

[9] Examples include LASSO (Tibishirani 1996), SCAD (Fan and Li 2001), Elastic Net (Zou and Hastie 2005), group LASSO (Zou 2006), bridge estimator (Huang, Horowitz and Ma 2008) and so on.

eters are not estimated and factor loadings are hard to interpret. On the other side, variable selection is mostly concerned only with the explanation of the dependent variable by choosing a subset of regressors, but not with the estimation of the partial effect. The key variable whose effect is of interest may be excluded from a variable selection procedure. Even though oracle properties of variable selection procedures (e.g. Huang, Horowitz and Ma 2008) have been established, these oracle properties do not provide a satisfactory answer in finite sample. First, when the variable of interest is not selected, oracle selection procedure such as bridge estimator would estimate the partial effect as zero. In this case, there is no way to do further inference such as constructing confidence interval or testing for the partial effect. That is, variable selection procedures would be over confident that the partial effect is zero when it is actually not. Second, even when the variable of interest is kept after the model selection procedure, the asymptotic distribution of the partial effect estimator depends on the true value of the partial effect and thus hard to provide valid inference in finite sample. See Leeb and Pötscher (2005, 2006, 2008abc, 2009), Pötscher (2009) and Pötscher and Schneider (2009, 2010) for problems that involve inferences with model selection procedures. Theoretical investigation of partial effect estimation with dimensionality reduction methods demands more effort before they become the working force.

Statisticians have long noticed that "all models are wrong but some are useful" (Box, 1979). This famous quote vividly describes a dilemma with which theoretical researchers are forced to face: models are misspecified. Taken as granted, we're in a position to estimate parameters of interest in misspecified models. For example, program evaluation researchers are evaluating the effects of the treatment with their misspecified model. The partial effect thus computed potentially suffers from model misspecification bias. Macro policy makers are predicting the effects of a counterfactual policy on the performance of economy, using a misspecified model. The prediction is as accurate as the model itself. Luckily for researchers that are concerned with partial effect parameters, potentially of low dimension, they are free of this misspecification problem, as to be pointed out by this paper. We specify a condition under which researchers who are interested in learning the partial effect parameters can well proceed with a misspecified model. Nevertheless, the parameter of interest should be correctly identified within the model. This is a big step, following White and Lu (2010), towards the estimation of economic sensible parameters rather than some statistical projection coefficients. It is important to point out that the identified partial effect parameters have the causal effect interpretation but the regression coefficients do not (e.g., White and Chalak 2006, White and Lu 2010). In a word, classical modeling and estimation approaches are contaminated with bias and new estimation techniques are called upon to derive more efficient estimators. This paper suggests the use of model averaging to achieve this aim.

Model averaging, advocated by Bates and Granger (1969), works as an alternative to the factor approach or variable selection in the forecasting paradigm. Simple model averaging gains a lot of popularity in financial market forecasts, for example, Rapach et al (2010). Recently, Hansen (2007, 2008, 2009, 2010) proposed model averaging with Mallow's criterion to select the combining weights, while Hansen and Racine (2011) proposed Jackknife model averaging. Model averaging is shown to be promising in forecasting exercises due to at least three facts: First, averaging reduces variances while incurring small bias. Whenever the bias is relatively small compared to the variance reduction, model averaging performs better than individual models in Mean Squared Error (MSE) sense. Secondly, individual models are likely to be misspecified and exclude information that is incorporated in averaging models. This loss of information potentially degrades the power of a single model. Thirdly, model uncertainty is somehow reduced in averaging model attributing to the observation that it incorporates individual models as special cases by properly assigning the weights, spanning a larger model space and reducing the chance of misspecification.

However, the power of model averaging for parameter estimation has not been fully explored. Hansen (2009) applied model averaging for parameter estimation in a structural break setting. The idea of averaging estimator dates back to Breiman (1996), where a **b**ootstrap method is implemented together with model **agg**regat**ing** (bagging, hereafter).[10] There is a large literature on Bayesian Model Averaging (BMA).[11] BMA takes a different perspective that the parameter of interest is random rather than have a true value. A prior on the parameter is required and a computing algorithm (e.g. MCMC) is needed to derive the BMA estimator. The dependence of the results on the prior and the algorithm adopted usually weaken the conclusions therefore arrived. More than often, convergence of the computing algorithms available (e.g. Metropolis-Hastings, Gibbs sampler, or MC^3, etc.) is hard to check in practice. See Hoeting et al (1999) for more details about these challenges faced by Bayesian researchers.

1.3 Contributions

This paper contributes to the literature in the following regards: First, we lay out the conditions that help to identify the partial effect parameter of interest in a large dimensional model. We show that Conditional Mean Independence (CMI) is sufficient for this purpose. This is a weaker condition than conditional independence used in White and Lu (2010). When CMI does not hold, we state a weaker condition, Weak Conditional Mean Independence, that identifies the partial effect parameter when the number of observation is large. CMI conditions can be either implied by conditional independence (White and Chalak 2010, Su and White 2011) or easily checked using the nonparametric tests proposed by Li and Wang (1998) or Hsiao, Li and Racine (2007). An information-based approach that is easy to implement is also suitable to test CMI.[12] We emphasize that such estimated coefficients would have economic

[10] Breiman (1996) shows that bagging estimator has a smaller MSE in the i.i.d. case for the the purpose of prediction. Bulman and Yu (2002) establish the theoretical properties of bagging estimators, followed by Lee, Tu and Ullah (2011ab) and Tu (2011a) that adopt bagging for constrained parameter estimation in nonparametric setting.

[11] In economics, recent work on BMA includes Sala-i-martin et al (2004), Eicher et al (2009) and so on.

[12] See Tu (2011b) for more details.

interpretation like causal effects only under identification. However, this identification issue is often ignored by empirical researchers, especially those who experiment with including and excluding explanatory variables till they get coefficient estimates agree with initial intuition.

Second, we consider the situation in which the parameter of interest can be identified in more than one model. This is often the case when we have large dimensional data. We propose two **m**odel **a**veraging **p**artia**l** **e**ffect (MAPLE) estimators in this setting. One of the estimators is generalized-method-of-moment based MAPLE (gMAPLE) and the other is entropy-based MAPLE (eMAPLE). The estimators are constructed from model averaging point of view, utilizing more than one model (potentially misspecified) to quantify such partial effect. They utilize more information than partial effect estimator derived from each individual model. Averaging in this way helps to wipe out the large bias lying in individual estimator and reduces variances, especially in small sample.

The gMAPLE estimator is constructed through combining all the moment conditions specified by individual models. A GMM-like objective function is used to derive the gMAPLE estimator. This estimator is different from the classical GMM estimator proposed by Hansen (1982) in the sense that each model has its own unique parameters other than the common partial effect parameter. This estimator looks similar to but differs from the GMM estimator of Seemingly Unrelated Regression models because of the common partial effect parameter in each model. gMAPLE estimator is the first attempt, as far as we know in the literature, to use moment conditions of more than one model to conduction inference on parameters of interest, while treating other parameters as pseudo ones.

The eMAPLE estimator is motivated from the Maximum Entropy point of view, i.e., to maximize the uncertainty of the model and data that is consistent with the moment conditions that identify the partial effect parameter. The main intuition is that the same set of data would occur with different probability if they are generated from different models. We introduce the concept of entropy of a model class in line with the classical notion of entropy of a random variable. We similarly define the conditional and joint entropy between a model class and random variables generated from that model class. Our eMAPLE estimator is constructed such that the conditional entropy of the model class given the observations is maximized. That is, the uncertainty of the model class is maximized given that the data is observed. Model averaging with entropy-based weights opens a new area of Maximum Entropic Econometrics (MEE). Other than estimating the probability of each observation in classical MEE, model averaging raises the question of probability of each individual model, instead of assigning equal probabilities. eMAPLE estimation is a novel statistical inference approach in that it introduces model uncertainty and model averaging into the entropy paradigm for parameter estimation. The inference based on the objective function (joint entropy) to construct confidence intervals or testing restrictions for the parameters of interest is easy to carry out and often better resembles the asymptotic results in finite sample than competing methods.

The third contribution of the paper is the theoretical study of the two MAPLE estimators. We set up conditions under which our MAPLE estimators are consistent and asymptotically normal. The conditions for gMAPLE estimator are similar to those in the GMM literature. The conditions for eMAPLE estimator resemble those used in the Generalized Empirical Likelihood (GEL) literature.[13] Testing of nonlinear restrictions on the parameters is also considered. We show that the Wald, Rao's Score and Likelihood-ratio type tests based on our MAPLE estimators are asymptotically chi-squared distributed.

The fourth contribution is the thorough simulation study conducted to compare various partial effect estimators, including MMA, JMA, FOGLeSs etc.. Our gMAPLE and eMAPLE estimator are shown to have appealing finite sample properties in various Data Generating Processes, including factor model, large dimensional models, models with large number of irrelevant regressors and models with heterogeneous errors etc.. Evaluation measures including Mean Squared Errors, Mean Absolute Errors, Bias, Variance, Inter Quantile Range are used to compare the competing estimators. Our MAPLE estimators clearly stand out, especially in small samples, and even achieve the oracle efficiency lower bound in MSE in some designs (true design is small dimensional without heterogeneity, but with a large dimensional covariates). We also conduct simulations to examine the performance of the MAPLE based test statistics. Generally, these tests enjoy sizes closer to theoretical ones than other testing procedures including, e.g., FOGLeSs based tests.

Finally, we illustrate the use of MAPLE estimator in an economic application to evaluate the effect of inherited control on firm performance. We find that our MAPLE estimates confirm earlier findings by Pérez-González (2006) and White and Lu (2010) that there is a negative effect, i.e., firms with family related CEOs tend to underperform those with family unrelated CEOs. However, confidence intervals constructed based on MAPLE estimators are much narrower than those based on FOGLeSs estimator, which indicates the superority the proposed approach.

Structure of the rest of this paper is planned as follows: Section 2 presents the model and discusses the identification issues. Section 3 proposes the gMAPLE estimator, introduces the concept of entropy of models in the presence of model uncertainty and proposes the eMAPLE estimator. Section 4 presents the theoretical properties of the proposed MAPLE estimators. Section 5 studies the finite sample properties, via simulation experiments, of our estimator together with other competitors. Section 6 provides an illustration of our estimation approach with the dataset of Pérez-González(2006) in the study of the impact of inherited control on firm performance. Section 7 concludes and comments on future studies. All the technical proofs are collected in the Appendix.

2 The Model and Identification Condition

In this section, we introduce the model with large dimensional data and illustrate with six economic examples. We discuss the identification of the partial effect parameter of interest and

[13] See Kitamura (2006) for a review.

present the key condition, Conditional Mean Independence (CMI), that serves for identification purpose. Other approaches for identification are also discussed in a concise way. In the end, We point out other related issues.

2.1 The Model

We present the model after introducing notations. Let $\mathbf{y}$ denote the $n \times 1$ dependent variable, $\mathbf{x}$ the $n \times 1$ exogenous independent variable whose partial effect on $\mathbf{y}$ is of major interest, and $\mathbf{z}$ the large dimensional independent variables.

Assumption A.1 (linearity):

$$y_i = \mathbf{x}_i^{\tau} \beta + \mathbf{z}_i^{\tau} \gamma + \varepsilon_i \ , (i = 1, 2, \cdots, n) \tag{2.1}$$

where β is the partial effect vector of interest, γ is a large-dimensional coefficient vector and ε_i is the disturbance term.
Assumption A.2 (α-mixing stationarity): The large dimensional vector stochastic process $\{d_i\}_{i=1}^n \equiv \{y_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i\}_{i=1}^n$ is a stationary α-mixing process with mixing coefficients $\alpha\,(j)$ satisfying $\sum_{j=1}^{\infty} j^2 \alpha^{\epsilon/(\epsilon+2)}\,(j) < \infty$ for $0 < \epsilon \leq 1$, where $\mathbf{w_i}$ is some instrumental vector.
Assumption A.3 (moment restriction): All the instruments are orthogonal to the contemporaneous error term: $E\,(\mathbf{w}_{ik} \cdot \varepsilon_i) = \mathbf{0}$, for all i and k $(= 1, 2, \cdots, \dim(\mathbf{w}_i))$.
Assumption RC (rank condition): $E[\mathbf{w}_i\,(\ \mathbf{x}_i^{\tau},\ \mathbf{z}_i^{\tau})]$ is finite with full column rank.

We comment on the strength of Assumption A and RC before presenting some economic examples.

Remark A: Assumption A.1 assumes that the relationship between y and the covariates is linear. As we see later on, we require that y be linear in unknown parameters. This assumption is not restrictive and can be extended in various ways. However, we will maintain this assumption only to clarify the presentation of our identification and estimation approach. In addition, the model as specified does not contain an intercept term. This is not restrictive either since a demean of the data would remove the intercept. We emphasize that the model is structural in the sense that the parameters, e.g., β carry causal effect interpretation. More than often, a low dimensional parameter such as β has economic policy implication but not others contained in γ. Our inference is mainly concerned with β. Assumption A.2 is classical since that certain type of nonstationary process can be made stationary via transformations such as differencing or detrending. Dependence across observations is allowed by the α-mixing condition. Assumption A.3 would meet since we include all possible explanatory variables in the regression. Anything else that is not explained in the dependent variable should be due to the pure random error term.

Remark RC: The rank condition is needed to identify all the unknown parameters, but often fails when z_i is of large dimension. This is especially the case for economic models, contrasted to statistical models, since all economic variables are closely intertwined. In the case that a few economic variables are linearly dependent RC fails to hold. However, as argued earlier, economists more than often are concerned with only the partial effect parameter, β, but not the other coefficient vector γ. This observation is momentum since its implication is that we only need focus on identification and inference on β. This alleviates the need for Assumption RC and allows us to proceed with weaker condition such as Conditional Mean Independence. We will introduce CMI for identification after presenting some economic examples that highlight the importance of partial effect estimation in large dimensional data.

2.2 Examples

We briefly discuss some examples from macroeconomics, program evaluation and labor economics.

Example 2.1 *(**Phillips Curve**) The famous historical inverse relationship between the rate of unemployment and the rate of inflation in the economy, usually termed as Phillips Curve (Fisher 1926; Phillips 1958), has been the focus of macro economy since its birth. Yet this is short run phenomena. A cursory analysis of U.S. inflation and unemployment data 1953-92 reveals that there is no single curve that fits the data. However, this argument ignore the fact that the macro economy has been evolving over time and factors such as technological developments, institutional factors including macro policy are also affecting the curve. These factors might prove to be important, but do not change the relationship between unemployment rate and inflation rate. Therefore, the estimation of the Phillips Curve should incorporate other macro variables.*

Example 2.2 *(**Consumption Hypothesis**) Keynes (1936) developed his theory of consumption and detailed the relationship between consumption and income in his famous book* “The General Theory of Employment, Interest and Money” *(Keynes, 1936). A function that relates consumption and income is usually estimated and Keynes' consumption theory is tested. The marginal propensity to consume (MPC), i.e., the rate at which consumption changes as income is changing, is the slope of the consumption function. According to Keynes, MPC should be in between 0 and 1. However, a consumption function that only has income as an explanatory cause suffers from potential misspecification bias. It bases consumption only on current income, but neglects other factors that also have important effects. One such factor is future income, which leads to Friedman's (1957) Permanent Income Hypothesis.*

Example 2.3 *(**Treatment Effect**) Ashenfelter (1978) studied effect of training programs on earnings where individual characteristics such as gender, race, past earnings together with training variable.*

Example 2.4 *(**Wage Equation**) Kruger (1993) examined the role of computers on the wage structure. A long list of variables such as gender, education, race, age, occupation, union status, hours, marriage status, experience and region are considered as important factors when studying the effect of computers on wage.*

Example 2.5 *(**Inherited Control**) Pérez-González (2006) used a large data set from 355 management transitions of publicly traded U.S. corporations to examine whether firms with family related incoming chief executive officers (CEOs) underperform in terms of operating profitability relatives to firms with unrelated incoming CEOs. 34 covariates are used including firm size, firm's past performance, board's R&D expenditure, departing CEO's separation conditions and incoming CEO's ownership, incoming CEO's characteristics, together with 17 year dummies. We will provide more analysis with this example in the empirical exercise in Section 6.*

Example 2.6 *(**Economic Growth**) Sala-i-Martin et al (2004) studied the determinants of economic growth with 67 variables that correlate with economic growth with only 80 observations. This job would be in vain since we have a large number of unknowns compared to the number of observations. However, growth economists are interested to know whether a particular variable, e.g., human capital, is a determinant of economic growth, in the presence of large number of other covariates.*

2.3 Identification and Conditional Mean Independence

In this subsection, we look into the identification issue of the partial effect parameter β. We distinguish the identification problem for two cases: (i) when Assumption RC holds; and (ii) when Assumption RC fails. It is to be shown that Assumption RC, together with Assumption A.1, A.2 and A.3, are sufficient for β to be identified. When Assumption RC fails, a further condition called conditional mean independence (CMI) is introduced to identify β. Tests to verify CMI and lower level conditions that imply CMI are reviewed.

Note that under Assumption RC, $E\left[\mathbf{w}_i\left(\ \mathbf{x}_i^{\tau},\ \mathbf{z}_i^{\tau}\right)\right]$ is of full column rank. The moment restriction Assumption A.3 implies that,

$$E\left[\mathbf{w}_i\left(y_i - \mathbf{x}_i^{\tau}\beta - \mathbf{z}_i^{\tau}\gamma\right)\right] = \mathbf{0},$$

which is equivalent to

$$E\left[\mathbf{w}_i\left(\ \mathbf{x}_i^{\tau},\ \mathbf{z}_i^{\tau}\right)\right]\left[\begin{array}{c}\beta\\ \gamma\end{array}\right] = E\left(\mathbf{w}_i y_i\right). \tag{2.2}$$

There is a unique solution to the above equation. This completes the identification of β.

2.3.1 Conditional Mean Independence

If, on the other hand, Assumption RC fails, then $E\left[\mathbf{w}_i\left(\ \mathbf{x}_i^{\tau},\ \mathbf{z}_i^{\tau}\right)\right]$ is singular. This leads to multiple solutions of β in equation (2.2). Consequently, β is underidentified. We find the following condition is needed for β to be identified.

Assumption CMI (conditional mean independence):

$$E\left(\mathbf{z}_i^2|\mathbf{x}_i,\mathbf{z}_i^1\right)=E\left(\mathbf{z}_i^2|\mathbf{z}_i^1\right) \tag{2.3}$$

where $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$ forms a partition of $\mathbf{z}_i$, i.e., $\mathbf{z}_i^{\tau}=\left[\mathbf{z}_i^{1\tau},\mathbf{z}_i^{2\tau}\right]$,for $i=1,2,\cdots,n$.

CMI condition is quite commonly adopted in the literature of parameter identification. A similar form of CMI is used in Stock and Watson (2010, pp.232) to distinguish the role of variables of interest and control variables. Under CMI, the coefficient of the variable of interest is argued to have an interpretation of causal effect. In the case that $\mathbf{z}_i^2$ is univariate, tests of Li and Wang (1998) and Hsiao, Li and Racine (2007) can be easily adjusted to verify CMI condition. When $\mathbf{z}_i^2$ is multivariate, element-wise tests would apply.

Conditions stronger than CMI are, for example, conditional exogeneity and conditional independence. They have been imposed by Hahn (1998, 2004), White and Lu (2010) and White, Chalak and Lu (2010), to name a few, as major tools to study identification, treatment effect and Granger-Causality. Su and White (2007ab) suggest tests of conditional independence that are based on Hellinger metrics and empirical likelihood. White and Chalak (2010) provided tests for conditional exogeneity. See White and Chalak (2010), Su and White (2008) and references therein for details.

Lemma 2.7 *β is identified under Assumption 1, 2, 3, and CMI.*

The proof of Lemma 1 is given in Appendix A. More than often in economic modeling, we will assume the existence of a partition of $\mathbf{z}$ such that Assumption CMI is satisfied. As a result of Lemma 1, the partial effect parameter β is identified.

When the set of $\mathbf{z}$ contains a large dimensional data, it is possible that more than one decomposition can be found such that (2.3) is satisfied. This is the case if $\mathbf{z}$ are linearly dependent. In this circumstance, we have competing models that all can identify β according to Lemma 1. However, each model will produce a different estimate of β, for a given sample of observations. In practice, it is hard to tell which estimate is closer to the true value. An average estimate that aggregate these estimated values can be constructed, with weights inversely proportional to each individual variance. See White and Lu (2010) for example. However, the construction of this estimate requires the knowledge of variance of individual estimators. Estimates of the variance can be used in practice. Nevertheless, this estimation procedure is deemed to be inefficient since the estimation of β takes into account different model specifications one-at-a-time and that the variance estimates are usually not accurate in finite sample. In the next section, we propose a model averaging estimator that could potentially circumvent such difficulties and result in a more efficient estimator.

Next we investigate the more interesting situation when there is no such partition of $\mathbf{z}$ such that CMI holds. The direct consequence is that β is not identifiable. We consider two cases (i) weak identification and (ii) no identification.

2.3.2 Weak Conditional Mean Independence

Assumption WCMI (weak conditional mean independence):

$$E\left(\mathbf{z}_i^2|\mathbf{x}_i,\mathbf{z}_i^1\right)=E\left(\mathbf{z}_i^2|\mathbf{z}_i^1\right)+\eta\mathbf{x}_i \tag{2.4}$$

where $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$ forms a partition of $\mathbf{z}_i$, i.e., $\mathbf{z}_i^\tau=\left[\mathbf{z}_i^{1\tau},\mathbf{z}_i^{2\tau}\right]$,for $i=1,2,\cdots,n$, and η is a matrix of the same dimension as $\mathbf{z}_i^2$, with Euclidean norm

$$||\eta||=o\left(n^{-1/2}\right)$$

Under Assumption WCMI, β is weakly identified. As sample size increase, the dependence of $\mathbf{z}_i^2$ on $\mathbf{x}_i$ becomes weaker and weaker. In the limit, condition WCMI becomes condition CMI. Therefore, β is identified in the limit. This type of condition has been used in Belloni, et al (2011) to approximate the factor estimation.

Lemma 2.8 *β is weakly identified (identified in the limit) under Assumption 1, 2, 3, and WCMI.*

When WCMI condition fails, β cannot be identified in any approximating models. This is a more interesting case, since the true model cannot be approximated arbitrarily well as we intend to. Our proposed estimator based on model averaging, tends to perform well for this difficult case, as shown in our simulation results in Section 5.

Before we proceed, a few things should be noted in sequence. First, when β is not identified, estimators for β using methods such as OLS are not targeting the correct partial effect. Inevitably, estimators are biased and their properties are hard to evaluate. In this circumstance, hardly any effort can be made towards the estimation of partial effect. Second, partial identification approaches advocated by Manski (2003) could be employed when WCMI fails, which is beyond the focus of this paper. Third, the estimation of β can be put into a general framework in which conditional moment restrictions summarize the model information. These restrictions take the form

$$E\left[g\left(y_i,\mathbf{x}_i,\mathbf{z}_i;\beta,\gamma\right)\right]=0, \tag{2.5}$$

where $g\left(y_i,\mathbf{x}_i,\mathbf{z}_i;\beta,\gamma\right)$ has a known functional form, β is the partial effect parameter of interest and γ is a vector of pseudo parameters. Note that first, $g(\cdot)$ may be derived from a nonlinear model, thus it is not restricted to the model specified in (2.1). γ can also be an infinite dimensional parameter such as a nonparametric function. Identifications of this type have been studied by Chen, et al (2011). A separate paper is written to study estimation in this semiparametric framework and leaves us to focus on the case when γ is the coefficient of $\mathbf{Z}$.

3 Model Averaging PartiaL Effect Estimation

3.1 Model Uncertainty and Moment Uncertainty

We motivate the estimation of partial effect from the point of view of model uncertainty. Model (2.1) can be viewed as aggregated models from $\mathcal{M}$ with certain probabilities. For example, in state s, the dependent variable is generated through the following equation,

$$y_i = \mathbf{x}_i^{\tau} \beta + \mathbf{z}_{i,s}^{\tau} \gamma_s + \varepsilon_{i,s} \ (i = 1, 2, \cdots, n) \tag{3.1}$$

where $\mathbf{z}_{i,s}$ is a subset of $\mathbf{z}_i$ and γ_s denotes corresponding coefficient vector. Denote the above model as M_s and denote a collection of such models as $\mathcal{M}$. We emphasize that in (3.1), β is identified via the CMI condition.

Ideally, if the observed data can be classified according to the state from which they are generated, we can estimate the coefficients β and γ within each state via LS whenever it applies. A second averaging procedure may be implemented after $\hat{\beta}_s$ is computed in state s $(= 1, 2, \cdots, S)$ to derive a more efficient estimator $\hat{\beta}$ using an auxiliary regression. See White and Lu (2010) for such a construction via a pseudo regression of $\hat{\beta}_s$ on β. Nevertheless, classification of data into states is neither practical nor necessary. First, classification requires further information and renders the estimation even more complex. Inference after data classification or model selection raises challenging issues such as those in data snooping (White, 2000). See Berk et al (2009) and Berk et al (2011) for recent studies on this issue. Second, entropy-based inference is already suitable for this type of so-called ill-posed "inverse" problems. Partial effect estimation of β amounts to estimating the model probability distribution $\mathbf{p}$ and model coefficients β and γ. We present procedures that circumvent the classification difficulty as notified and achieves the estimation objective.

To put the analysis in a general framework, we present the estimation of β from the model information characterized by moment constraints in the form of (2.5). To facilitate the presentation, we simplify our notations. Note that first, model (3.1) in state s can be summarized by corresponding moment condition

$$E\left[g_s\left(d; \theta_0\right)\right] = 0, \tag{3.2}$$

where expectation is taken over random vector $d = (y, \mathbf{x}, \mathbf{z})$, with $g_s(\cdot; \cdot)$ denoting the moment restriction in M_s, and $\theta_0 = (\beta, \gamma_1, \ldots \gamma_S)$ collecting all the unknown parameters in S models. We emphasize that β is the partial effect parameter of interest that is identified in each model, but not the projection coefficient vectors $\gamma_s, s = 1, \ldots S$.

3.2 gMAPLE

Facing parameter estimation problems identified by moment conditions via (3.2), it is natural to adopt the Generalized Method of Moment (GMM) approach proposed by Hansen (1982). We present the GMM estimator in the current setting. Denote $\bar{g}_s(d, \theta) = \frac{1}{n} \sum_{i=1}^{n} g_s(d_i, \theta)$ and

$\bar{g}(d,\theta) = [\bar{g}_1^\tau(d,\theta), \dots, \bar{g}_S^\tau(d,\theta)]^\tau$. The one-step GMM estimator with a weighting matrix W is defined as

$$\hat{\theta}_{gMAPLE} = \arg\min_{\theta} \bar{g}^\tau(d,\theta) W \bar{g}(d,\theta). \tag{3.3}$$

The solution to this convex minimization problem can be easily found through numerical methods.

We need some notation to proceed. Define $\nabla_\theta g_s(d,\theta) = \partial g_s^\tau(d,\theta)/\partial\theta$, where $\partial g_s^\tau(d,\theta)/\partial\theta$ is the transpose of $\partial g_s(d,\theta)/\partial\theta$. Denote $G(s,\theta) = E[\nabla_\theta g_s(d,\theta)]$ and $V(s,\theta) = E[g_s(d,\theta) g_s^\tau(d,\theta)]$. Define $G(\theta) = (G^\tau(1,\theta), \dots, G^\tau(S,\theta),)^\tau$, $V(\theta)$ =diag$(V(1,\theta), \dots, V(S,\theta))$ and use short notation $G = G(\theta_0)$, $V = V(\theta_0)$, $\Omega = E[g(d,\theta_0) g^\tau(d,\theta_0)]$. Following the GMM literature (see, e.g., Newey and McFadden, 1994), it is easy to establish the following theorem, under suitable set of additional assumptions on the moment conditions (3.2).

Theorem 3.1 *The GMM estimator defined in (3.3) has the following properties:*

(a) $\hat{\theta}_{gMAPLE} \xrightarrow{p} \theta_0$.

(b) $\sqrt{n}\left(\hat{\theta}_{gMAPLE} - \theta_0\right) \xrightarrow{d} N\left(0, (G^\tau W G)^{-1} G^\tau W \Omega W G (G^\tau W G)^{-1}\right)$

An efficient two-step GMM estimator can be derived based on a first step estimator $\hat{\theta}_{gMAPLE1}$ that solves (3.3) by setting $W = I$, the identity matrix. The optimal weight matrix can be shown to be $W_{opt} = \Omega^{-1}$ that can be consistently estimated by

$$\hat{W}_{opt} = \left[\frac{1}{n}\sum_{i=1}^{n} g\left(d_i, \hat{\theta}_{gMAPLE1}\right) g^\tau\left(d_i, \hat{\theta}_{gMAPLE1}\right)\right]^{-1}. \tag{3.4}$$

Theorem 3.2 *The GMM estimator defined in (3.3) with $W = \hat{W}_{opt}$ have the following properties:*

(a) $\hat{\theta}_{gMAPLE} \xrightarrow{p} \theta_0$.

(b) $\sqrt{n}\left(\hat{\theta}_{gMAPLE} - \theta_0\right) \xrightarrow{d} N\left(0, I^{-1}(\theta_0)\right)$, with $I(\theta_0) = G^\tau \Omega^{-1} G$.

Note that (3.4) is a very large dimensional matrix in the current setting. Earlier results (e.g., Altonji and Segal 1994) show that GMM estimator with estimated optimal weighting matrix does not perform well in finite sample. The two-step optimal GMM can be even beaten by the one-step GMM that uses the naive identity weighting matrix. In practice, iterative GMM estimator and continuously updating GMM estimators can be used, see Hansen et al (1996).

3.3 eMAPLE

This section introduce entropy-based model averaging. We start by defining the entropy for model uncertainty of a given class of models. We then extend this concept and account for model uncertainty in the presence of random variables that are generated from the models. Similar concepts, such as entropy, joint entropy and conditional entropy, exist in the entropy literature, for example, as in Cover and Thomas (2006) or Golan et al (1996). However, to our knowledge, it is the first time to define these concepts for random models.

3.3.1 Entropy

Imagine a world that is comprised of a finite number of states $s = 1, 2, \cdots, S$. In each state, the data generating process is described by a mechanism, called *model.* We denote $\mathcal{M}$ as a collection of such models, i.e., $\mathcal{M} = \{M_s : s = 1, 2, \cdots, S\}$, where M_s describes the world in state s. Each state of the world, s, is associated with a probability q_s. We denote the probability space by the simplex $\triangle^S = \left\{ \mathbf{q} \in \mathbf{R}^S : q_s \geq 0, \sum\limits_{s=1}^{S} q_s = 1 \right\}$.

Definition 3.3 *Consider a class of models $\mathcal{M} = \{M_s : s = 1, 2, \cdots, S\}$, from which data are generated with probability distribution $\mathbf{q}(\mathcal{M}) = (q_1, q_2, \cdots, q_S)$. The* ***entropy*** *that characterizes the information uncertainty associated with $\mathcal{M}$ is defined as*

$$H(\mathbf{q}) = -\sum_{s=1}^{S} q_s \log q_s,$$

where the convention $0 \cdot \log 0 = 0$ is taken.

Here $\mathbf{q}(\mathcal{M}) = (q_1, q_2, \cdots, q_S)$ is the the probability mass function of models M_1, M_2,$\cdots$, M_S that are contained in $\mathcal{M}$. It is abbreviated as $\mathbf{q}$ whenever no confusion occurs. As defined, $H(\mathbf{q})$ is a measure of the amount of uncertainty in the probability mass $\mathbf{q}(\mathcal{M})$ that describes the states of the world. It reaches a maximum when $q_s = 1/S$, for all $s = 1, 2, \cdots, S$, i.e., when the probability is uniform. This definition is consistent with entropy of a discrete random variable. See, for example, Cover and Thomas (2006) or Golan et al (1996) for more details.

Next, we extend the measure of uncertainty when there is an additional set of observations from the potential class of models $\mathcal{M}$. The following definition parallels that of joint entropy of two random variables. Let a random vector D be defined on $\mathcal{D}$.

Definition 3.4 *The* ***joint entropy*** *$H(\mathcal{M}, D)$ of the model class $\mathcal{M}$ and the random vector D with a joint distribution $p(M, D)$ is defined as*

$$\begin{aligned} H(\mathcal{M}, D) &= -\sum_{M \in \mathcal{M}} \sum_{d \in \mathcal{D}} p(M, d) \log p(M, d) \qquad (3.5) \\ &= -E \log p(\mathcal{M}, D). \end{aligned}$$

We further define the conditional entropy of a model class given a random vector as the expected value of the entropies of the conditional distributions averaged over the conditioning random vector.

Definition 3.5 *The* ***conditional entropy*** $H\left(\mathcal{M}|D\right)$ *of the model class* $\mathcal{M}$ *given the random vector* D *with a joint distribution* $p\left(M,d\right)$ *is defined as*

$$\begin{aligned} H\left(\mathcal{M}|D\right) &= \sum_{d\in\mathcal{D}} p\left(d\right) H\left(\mathcal{M}|D=d\right) \\ &= -\sum_{d\in\mathcal{D}} p\left(d\right) \sum_{M\in\mathcal{M}} p\left(M|d\right) \log p\left(M|D=d\right) \\ &= -\sum_{M\in\mathcal{M}} \sum_{d\in\mathcal{D}} p\left(M,d\right) \log p\left(M|d\right) \\ &= -E \log p\left(\mathcal{M}|D\right). \end{aligned} \tag{3.6}$$

Similarly, we can define the conditional entropy $H\left(D|\mathcal{M}\right)$ of the random vector D given the model class $\mathcal{M}$.

Definition 3.6 *The* ***conditional entropy*** $H\left(D|\mathcal{M}\right)$ *of the random vector* D *given the model class* $\mathcal{M}$ *with a joint distribution* $p\left(M,d\right)$ *is defined as*

$$\begin{aligned} H\left(D|\mathcal{M}\right) &= \sum_{M\in\mathcal{M}} p\left(M\right) H\left(D|\mathcal{M}=M\right) \\ &= -\sum_{M\in\mathcal{M}} p\left(M\right) \sum_{M\in\mathcal{M}} p\left(d|M\right) \log p\left(d|\mathcal{M}=M\right) \\ &= -\sum_{d\in D} \sum_{M\in\mathcal{M}} p\left(M,d\right) \log p\left(d|M\right) \\ &= -E \log p\left(D|\mathcal{M}\right). \end{aligned} \tag{3.7}$$

The following theorem shows that the difference between the joint entropy defined in (3.5) and conditional entropy defined in (3.6) is the entropy of the conditioning random vector D. Similar result holds if we switch the model class and the random vector.

Theorem 3.7 *(Chain rule)*

$$\begin{aligned} H\left(\mathcal{M},D\right) &= H\left(\mathcal{M}|D\right) + H\left(D\right) \\ &= H\left(D|\mathcal{M}\right) + H\left(\mathcal{M}\right). \end{aligned}$$

Proof: The proof follows closely from that of Theorem 2.2.1 in Cover and Thomas (2006, p.17).

3.3.2 eMAPLE estimator

Instead of approximating the expectation in (3.2) with a simple sample average as in GMM, we adopt

$$\sum_{i=1}^{n} p_{is} g_s\left(d_i, \theta_0\right) = 0$$

where p_{is} is defined as probability of observing d_i given that the model is M_s. That is, $p_{is} = p\left(d_i | M_s\right)$, $s = 1, \ldots, S$. Requirement of probability states that

$$p_{is} \geq 0 \text{ and } \sum_{i=1}^{n} p_{is} = 1, i = 1, \ldots, n, s = 1, \ldots, S.$$

For each parameter vector $\theta \in \mathbf{\Theta}$, define the set of probability measures:

$$\mathcal{P}\left(\theta\right) \equiv \left\{ \mathbf{p} = \left(p_1^\tau, \ldots, p_S^\tau\right)^\tau : \sum_{i=1}^{n} p_{is} g_s\left(d_i; \theta\right) = 0, \sum_{i=1}^{n} p_{is} = 1, p_s^\tau = \left(p_{is}\right) \geq 0, s = 1, \ldots, S. \right\}. \tag{3.8}$$

and

$$\mathcal{Q}\left(\theta\right) \equiv \left\{ \mathbf{q} = \left(q_1, \ldots, q_S\right)^\tau : \sum_{s=1}^{S} q_s = 1 \right\},$$

where $q_s = p\left(M_s\right)$, $s = 1, \ldots S$. To estimate the probabilities, it is natural to consider the following maximization problem,

$$\max_{\left[p^\tau, q^\tau\right]^\tau \in \mathcal{P}(\theta) \times \mathcal{Q}(\theta)} H\left(\mathcal{M} | D\right), \tag{3.9}$$

for each $\theta \in \mathbf{\Theta}$. That is, we simultaneously select the conditional probabilities, p_{is}, the probability of observing d_i given model M_s, and the marginal probability of model M_s, q_s, to maximize the missing information between the class of model $\mathcal{M}$ and the observed data. This is the essential philosophy of maximum entropy econometrics.

To analyze directly the objective function in (3.9), ones needs to know the conditional entropy of a model class for a given data set. This is a conceptual challenge, since we need begin with the probability distribution of the model class for a given data set. This is exactly the difficulty researchers facing when model uncertainty presents, for example, in the Bayesian model averaging methods. However, we will circumvent this difficulty in entropy-based approach. We make use of the following theorem to rewrite the objective function.

Theorem 3.8 *The solution in (3.9) solves*

$$\max_{\left[p^\tau, q^\tau\right]^\tau \in \mathcal{P}(\theta) \times \mathcal{Q}(\theta)} - \sum_{s=1}^{S} \sum_{i=1}^{n} q_s p_{is} \log p_{is} - \sum_{s=1}^{S} q_s \log q_s.$$

Proof. Note first that the joint entropy

$$\begin{aligned} H(\mathcal{M}, D) &= H(D|\mathcal{M}) + H(\mathcal{M}) \\ &= -\sum_{s=1}^{S}\sum_{i=1}^{n} q_s p_{is} \log p_{is} - \sum_{s=1}^{S} q_s \log q_s \end{aligned}$$

Therefore, we have

$$\begin{aligned} [\hat{p}^{\tau}, \hat{q}^{\tau}]^{\tau} &= \arg \max_{[p^{\tau},q^{\tau}]^{\tau} \in \mathcal{P}(\theta)\times\mathcal{Q}(\theta)} H(\mathcal{M}|D) \\ &= \arg \max_{[p^{\tau},q^{\tau}]^{\tau} \in \mathcal{P}(\theta)\times\mathcal{Q}(\theta)} H(\mathcal{M}|D) + H(D) \\ &= \arg \max_{[p^{\tau},q^{\tau}]^{\tau} \in \mathcal{P}(\theta)\times\mathcal{Q}(\theta)} H(D, \mathcal{M}) \\ &= \arg \max_{[p^{\tau},q^{\tau}]^{\tau} \in \mathcal{P}(\theta)\times\mathcal{Q}(\theta)} H(D|\mathcal{M}) + H(\mathcal{M}) \\ &= \arg \max_{[p^{\tau},q^{\tau}]^{\tau} \in \mathcal{P}(\theta)\times\mathcal{Q}(\theta)} -\sum_{s=1}^{S}\sum_{i=1}^{n} q_s p_{is} \log p_{is} - \sum_{s=1}^{S} q_s \log q_s. \end{aligned}$$

This completes the proof. ■

Lagrange multipliers can be used to solve (3.9). The Lagrangian is

$$\begin{aligned} \mathcal{L} &= -\sum_{s=1}^{S}\sum_{i=1}^{n} q_s p_{is} \log p_{is} - \sum_{s=1}^{S} q_s \log q_s - \mu \left[\sum_{i=1}^{n} p_{is} - 1\right] \\ &\quad -\sum_{s=1}^{S} \eta_s^{\tau} \sum_{i=1}^{n} g_s(d_i;\theta) p_{is} - \xi \left[\sum_{s=1}^{S} q_s - 1\right], \end{aligned} \tag{3.10}$$

where μ, η_s^{τ} and ξ are Lagrange multipliers.

In the appendix, we show that the solution to (3.9) are

$$\hat{q}_s = \frac{1}{\sum\limits_{s=1}^{S} \exp\left(-\sum\limits_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is}\right)} \exp\left(-\sum_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is}\right), \tag{3.11}$$

and

$$\hat{p}_{is} = \frac{1}{\Upsilon_s(\lambda_s, \theta)} \exp\left[-\lambda_s^{\tau} g_s(d_i;\theta)\right] \tag{3.12}$$

where

$$\Upsilon_s(\lambda_s, \theta) = \sum_{i=1}^{n} \exp\left[-\lambda_s^{\tau} g_s(d_i;\theta)\right], \tag{3.13}$$

with $\lambda_s^{\tau} = \eta_s^{\tau}/\hat{q}_s, \lambda = (\lambda_1^{\tau}, \ldots, \lambda_S^{\tau})^{\tau}$. In addition, each $\theta \in \mathbf{\Theta}$, $\hat{\lambda}_s^{\tau}$ solves

$$\sum_{i=1}^{n} g_s(d_i;\theta) \exp\left[-\hat{\lambda}_s^{\tau} g_s(d_i;\theta)\right] = 0, \tag{3.14}$$

for all $s = 1, \ldots, S$.

We define the profile joint entropy (JE) at θ as

$$\begin{aligned} \mathrm{JE}(\theta) &= -\sum_{s=1}^{S}\sum_{i=1}^{n}\hat{q}_s\hat{p}_{is}\log\hat{p}_{is} - \sum_{s=1}^{S}\hat{q}_s\log\hat{q}_s \\ &= \log\Upsilon(\lambda,\theta), \end{aligned} \tag{3.15}$$

where the last equality is show in the Appendix, with

$$\Upsilon(\lambda,\theta) = \sum_{s=1}^{S}\Upsilon_s(\lambda_s,\theta) = \sum_{s=1}^{S}\sum_{i=1}^{n}\exp\left[-\lambda_s^{\tau}g_s(d_i;\theta)\right],$$

and $\lambda = (\lambda_1^{\tau}, \ldots, \lambda_S^{\tau})^{\tau}$.

Our **e**ntropy-based **m**odel **a**veraging **p**artia**l** **e**ffect (eMAPLE) estimator of θ is thus defined as

$$\begin{aligned} \hat{\theta}_{eMAPLE} &= \arg\max_{\theta\in\boldsymbol{\Theta}}\mathrm{JE}(\theta) \qquad (3.16)\\ &= \arg\max_{\theta\in\boldsymbol{\Theta}}\frac{1}{nS}\exp\{JE(\theta)\} \\ &= \arg\max_{\theta\in\boldsymbol{\Theta}}\frac{1}{nS}\Upsilon(\lambda,\theta) \\ &= \arg\max_{\theta\in\boldsymbol{\Theta}}\mathrm{JE}_n(\theta), \end{aligned}$$

where

$$\mathrm{JE}_n(\theta) = \frac{1}{nS}\Upsilon(\lambda,\theta) = \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n}\exp\left[-\lambda_s^{\tau}g_s(d_i;\theta)\right]$$

To implement our estimator, it's easily seen that the λ_s^{τ} solving (3.14) can be alternatively found as

$$\lambda_s^{\tau} = \arg\max_{\varsigma\in R^{\dim(g_s(x,\theta))}}\Upsilon_s(\varsigma,\theta),$$

Note that this is a well-defined finite dimensional unconstrained convex maximization problem that has a unique solution. Algorithms such as Newton-Raphson method can be easily applied. Once λ_s^{τ} is solved as a function of θ, it can be substituted to (3.12) and (3.11), and consequently, (3.16) can be solved easily through numerical methods.

3.4 Alternative methods

An alternative is to optimally combine the estimators of parameters common to all models via an artificial regression (White and Lu, 2010). Denote the GLS estimator of β_0 (parameter of interest) from model s as $\tilde{\beta}_s$ (we suppress the dependence on sample size n) and that of θ_0 as $\tilde{\theta}_s$.

Denote $\tilde{\beta}_n = \left[\tilde{\beta}_1^\tau, \ldots, \tilde{\beta}_S^\tau\right]^\tau$, $\tilde{\theta}_n = \left[\tilde{\theta}_1^\tau, \ldots, \tilde{\theta}_S^\tau\right]^\tau$ and Λ a selection matrix such that $\tilde{\beta}_n = \Lambda\tilde{\theta}_n$. A combined estimator of β_0 can be formulated through the following regression,

$$\sqrt{n}\tilde{\beta}_n = \sqrt{n}\mathcal{I}\beta_0 + e, \tag{3.17}$$

where the $S \dim(\beta_0) \times S \dim(\beta_0)$ matrix of artificial regressor $\mathcal{I} \equiv \iota \otimes I_{\dim(\beta_0)}$, with ι being the $S \times 1$ vector of ones and $I_{\dim(\beta_0)}$ being the identity matrix of the same dimension as β_0, $e \sim N(0, \Sigma^*)$ is the artificial regression error with $\Sigma^* = \left((\Lambda G)^\tau \Omega^{-1} \Lambda G\right)^{-1}$. The Feasible Optimally combined GLS (FOGLeSs) estimator of White and Lu (2010) is defined as the FGLS estimator of (3.17):

$$\tilde{\beta}_n^* = \left(\mathcal{I}^\tau \hat{\Sigma}^{*-1} \mathcal{I}\right)^{-1} \mathcal{I}^\tau \hat{\Sigma}^{*-1} \tilde{\beta}_n, \tag{3.18}$$

where $\hat{\Sigma}^*$ is a consistent estimator of Σ^* and satisfies

$$\sqrt{n}\left(\tilde{\beta}_n^* - \beta_0\right) \xrightarrow{d} N\left(0, \left(\mathcal{I}^\tau \Sigma^{*-1} \mathcal{I}\right)^{-1}\right).$$

Note that $\left(\mathcal{I}^\tau \Sigma^{*-1} \mathcal{I}\right)^{-1} = \left(\mathcal{I}^\tau \left((\Lambda G)^\tau \Omega^{-1} \Lambda G\right) \mathcal{I}\right)^{-1} \geq \Lambda \left(G^\tau \Omega^{-1} G\right)^{-1} \Lambda^\tau$. FOGLeSs estimator is not as efficient as optimal GMM estimator. We emphasize that, to implement FOGLeSs, $\hat{\Sigma}^*$ is needed to compute $\tilde{\beta}_n^*$ in (3.18).

However, it is important to explore our proposed entropy-based estimation approach for the following reasons. First, the above two-step estimation procedures require the first step consistent estimator of θ_0, which will be used for the estimation of the optimal weighting matrix. Inevitably, this would introduce finite sample bias for the second stage estimation. Second, the weighing matrix (either is of large dimension whose accuracy is more than often a concern when available data sample size is small.

4 Theoretical Properties

In this section, we present the theoretical properties of the eMAPLE estimator.

4.1 Consistency

We introduce some notations before stating the needed assumptions. Define $\nabla_\theta g_s(d, \theta) = \partial g_s^\tau(d, \theta) / \partial\theta$, where $\partial g_s^\tau(d, \theta) / \partial\theta$ is the transpose of $\partial g_s(d, \theta) / \partial\theta$. Denote $G(s, \theta) = E\left[\nabla_\theta g_s(d, \theta)\right]$ and $V(s, \theta) = E\left[g_s(d, \theta) g_s^\tau(d, \theta)\right]$.

Assumption B.1 For each $\theta \neq \theta_0$ there exists a sub class $\mathcal{M}_\theta \subseteq \mathcal{M}$ such that $\Pr(M_s \in \mathcal{M}_\theta) > 0$, and $E\{g_s(d, \theta)\} \neq 0$ for each $M_s \in \mathcal{M}_\theta$.
Assumption B.2 $E\{\sup_{\theta \in \boldsymbol{\Theta}} ||g_s(d, \theta)||^m\} < \infty$ for some $m \geq 8$, for all $s = 1, \ldots, S$.
Assumption B.3 For all $s = 1, \ldots, S$,

(i) $\theta_m \to \theta \in \boldsymbol{\Theta} \Longrightarrow g_s(d, \theta_m) \to g_s(d, \theta)$ for almost every d;

(ii) $E\left[\sup_{\theta\in\boldsymbol{\Theta}} \parallel \partial g_s(d,\theta)/\partial\theta' \parallel\right] < \infty$, for all $s = 1, \ldots, S$.

(iii) $\sup_{\theta\in\mathcal{B}} \left|\partial g_s^{(i)}(d,\theta)/\partial\theta^{(j)}\right| \leq r(d), \sup_{\theta\in\mathcal{B}} \left|\partial^2 g_s^{(i)}(d,\theta)/\partial\theta^{(j)}\partial\theta^{(k)}\right| \leq t(d)$, $w.p.1$ for some real valued functions $r(d)$ and $t(d)$ such that $Ed^{\upsilon} < \infty$ for some $\upsilon \geq 4$ and $Et(d) < \infty$.

Assumption B.4 There is a closed ball around θ_0, $\mathcal{B}$, such that for all $s = 1, \ldots, S$

(i) $G(s, \cdot)$ and $V(s, \cdot)$ are continuous w.p.1. on $\mathcal{B}$.

(ii) $\inf_{(\varsigma,s,\theta)} \varsigma' V(s,\theta)\varsigma > 0$ and $\sup_{(\varsigma,d,\theta)} \varsigma' V(s,\theta)\varsigma < \infty$ with $\theta \in \mathcal{B}$.

Assumption B.5 $\lambda_s \in \left\{\gamma : \|\gamma\| \leq an^{-1/m}\right\}$ for some $a > 0$ and m as in Assumption B.2.

Remark: Similar assumptions to the above ones are adopted in the EL literature (e.g., Kitamura, Tripathi and Ahn (2004)). Assumptions B.1 states that θ_0 is identified jointly in all S models. Assumption B.2 is needed to prove a Lemma C.1 in line with Lemma 3 of Owen (1990) or Lemma D.2 of Kitamura, Tripathi and Ahn (2004). Assumption B.3 impose regularity conditions on the moment function. Assumption B.4 imposes conditions on the first derivative of the moment condition and the variance-covariance matrix. Assumption B.5 is a technical assumption that leads to the asymptotic normality of e-MAPLE estimator.

Theorem 4.1 *(consistency) Under Assumption A.1-3, B.1-4, e-MAPLE estimator is consistent, i.e.,* $\hat{\theta}_{eMAPLE} \to^p \theta_0$.

Proof: See the Appendix.

Remark: Theorem 4.1 shows that e-MAPLE estimator is consistent. The consistency comes as a result of the moment conditions that identify θ_0 (that includes β). We emphasize that γ_s in θ_0 is pseudo projection coefficient vector in model s and do not carry any economic interpretation.

4.2 Asymptotic Normality

Theorem 4.2 *(asymptotic normality) Under additional Assumption B.5,*

$$\sqrt{n}\left[\hat{\theta}_{eMAPLE} - \theta_0\right] \xrightarrow{d} N\left(0, J^{-1}(\theta_0)\, I(\theta_0)\, J^{-1}(\theta_0)\right),$$

where $J(\theta_0) = G^{\tau}V^{-1}G$, $I(\theta_0) = G^{\tau}V^{-1}\Omega V^{-1}G$.

Proof: See the Appendix.

Remark: The theorem shows that, when there is no correlation among models in different states, i.e., $\Omega = V$, our e-MAPLE estimator $\hat{\theta}_{eMAPLE}$ is asymptotically efficient and it achieves the asymptotic variance lower bound, $G^{\tau}\Omega^{-1}G$, which is the expectation of the inverse of Fisher information matrix averaged across states of the world. Note that this lower bound agrees the variance of optimally weighted GMM estimator, where the optimal weight is used for each individual model. When $\Omega \neq V$, i.e., there is correlation among models in different states, the e-MAPLE estimator agrees with the GMM estimator that adopts weighting matrix $W = V$. This is suboptimal in the sense that it efficiently uses information of in each model only. As pointed out earlier, e-MAPLE estimator avoids the estimation of the large dimensional variance-covariance matrix, which makes it appealing in finite sample.

4.3 Hypothesis Testing

To construct tests of the possible nonlinear restrictions as follows:

$$H_0 : R\left(\theta_0\right) = r, \tag{4.1}$$

where r is a $k \leq m$ dimensional vector of constants and $R\left(\cdot\right)$ is a known parametric function. Impose this restriction in the optimization of (3.16). Denote the constrained solution by $\hat{\theta}^c$ and the Jacobian matrix of R evaluated at θ_0 as A, which is assumed to be of full row rank. We have the following theorem for the Wald, Rao's Score, and Likelihood Ratio-like test statistics.

Theorem 4.3 *Test statistics of the restrictions (4.1),*

$$\begin{aligned} Wald_n &= n\left[R\left(\hat{\theta}\right) - r\right]' \left[A\hat{I}^{-1}\left(\theta_0\right)A\right]^{-1} \left[R\left(\hat{\theta}\right) - r\right], \\ LM_n &= ng\left(d, \hat{\theta}^c\right)\hat{V}^{-1}\hat{G}\hat{I}^{-1}\left(\theta_0\right)\hat{G}'\hat{V}^{-1}g\left(d, \hat{\theta}^c\right), \\ LR_n &= 2\left[JE_n\left(\hat{\theta}\right) - JE_n\left(\hat{\theta}^c\right)\right] \end{aligned}$$

are asymptotically χ^2_k*, where* $\hat{V}, \hat{G}, \hat{I}^{-1}\left(\theta_0\right)$, $\hat{A}$*, are consistent estimates of* $V, G, I\left(\theta_0\right)$ *and* A*.*

Proof: The results follows from Theorem 3 and Amemiya (1985).

Remark: Tests based on g-MAPLE estimators can be similarly constructed, without any difficulty. However, in practice, we recommend the LR_n test based on e-MAPLE estimator due to its easy implementation and nice finite sample properties as to be shown in the next section.

5 Finite Sample Investigations

In this section, we conduct simulation studies to examine the finite sample properties of e-MAPLE estimator, with a comparison to other estimators available in the literature. We include the ordinary least square estimator, Generalized Least Square (GLS) estimator with perfect knowledge of heterogeneity function, Feasible GLS with knowledge of heterogeneity functional form,1-step GMM (GMM1) estimator, 2-step optimal GMM (GMM2) estimator, the FOGLeSs estimator of White and Lu (2010), LASSO estimator of Tibshirani (1996), the factor based estimator of Galbraith etc (2010) and the Mallows model averaging (MMA) estimator of Hansen (2007).

We perform sequentially five experiments for the investigation. We briefly describe our experiments before presenting the details. The first experiment is to study the performance in a factor model setting. The second experiment is look into the classical regression model with a large number of covariates. The third experiment is to amplify the role of efficient estimation in the presence of heterogeneous errors. The next experiment is to investigate the effects of the irrelevant covariates with homogeneous disturbance, which is replaced by heterogeneity in the last experiment. For all experiments considered, we consider sample size $n = 50, 100, 150, 200, 250$ and replicate the process 1,000 times. The covariates are kept fixed for each replication. The estimator of the parameter of interest, β, is the partial effect of x on y. We report different criteria to evaluate estimators under investigation, including the Mean Squared Error (MSE), the Mean Absolute Error (MAE), Squared Bias (Bias^2), Variance (Var) and Inter Quantile Range (IQR) over 1,000 replications.

5.1 Experiment 1: factor model

We first consider a factor model, in which all the observed covariates are generated from some underlying factors f_i, according to the following DGP.

$$\text{DGP1}: \begin{cases} y_i = x_i^\tau \beta + z_i' \gamma + e_{1i}, \\ Z_i = f_i^\tau \xi_z + e_{2i}, \\ x_i = f_i^\tau \xi_x + e_{3i}, \end{cases}$$

$i = 1, ..., n$. We consider $n_f = \dim(f_i) = 3$ and generate f_i from a multivariate normal distribution with random mean vector and random covariance matrix. $p = \dim(z_i) = 0.8n - 2$, ξ_z is generated in a similar way as f_i but is normalized to a unit vector. ξ_x is generated from uniform $U[0, 3]$. We set $\beta = [2, 3]'$, and γ is generated from $U[0, 0.3]$. $e_{ji}, j = 1, 2, 3$, is independent standard normal error.

5.2 Experiment 2: regression model, case 1

We next consider the classical regression model that has a large dimensional observed covariates.

$$\text{DGP2}: \ \ y_i = x_i^{\tau}\beta + z_i'\gamma + e_i,$$

where x_i and z_i are generated from independent standard normal distribution. x_i and z_i of the same dimension as those in DGP 1, and so are the values of β and γ. e_i is the independent standard normal error.

5.3 Experiment 3: regression model, case 2

While heterogeneity is more often the case than exception in economic data, we incorporate such a feature into DGP3.

$$\text{DGP3}: \ \ y_i = x_i^{\tau}\beta + z_i'\gamma + v_i^m \cdot e_i,$$

We generate x_i, z_i, β, γ and e_i in the same way as in DGP2. We consider heterogeneity function v_i^m for three different forms

$$\begin{aligned} v_i^1 &= \sqrt{x_{1i}^2 + x_{1i}^2} \\ v_i^2 &= \sqrt{x_{1i}^2 + 2x_{1i}^2} \\ v_i^3 &= \exp\left(-x_{1i}^2\right) \end{aligned}$$

5.4 Experiment 4: regression model, case 3

Note that in earlier experiments, the large dimensional covariate vector (x_i, z_i) are causal in generating the dependent variable y. We next consider the case in which a large dimensional irrelevant covariates are available.

$$\text{DGP4}: \ \ y_i = x_i^{\tau}\beta + z_{1i}'\gamma + e_i,$$

where Z_{1i} is a subset of Z_i. x_i, Z_i, β, γ and e_i are generated in the same way as in DGP2.

5.5 Experiment 5: regression model, case 4

We further consider effects of heterogeneity in error term.

$$\text{DGP5}: \ \ y_i = x_i^{\tau}\beta + z_{1i}'\gamma + u_i^m \cdot e_i,$$

where we consider three different form of heterogeneity,

$$\begin{aligned} u_i^1 &= \sqrt{x_{1i}^2 + 2x_{1i}^2}, \\ u_i^2 &= \sqrt{x_{1i}^2 + 5x_{1i}^2}, \\ u_i^3 &= \sqrt{x_{1i}^2 + 5x_{1i}^2 + 2\cos\left(x_{1i}x_{2i}\right)}. \end{aligned}$$

Table 4-21 present the simulation results for experiment 1-5. To save space, we only report the squared bias and MSE for the estimators considered, with sample size 50 and 200. Other simulation results resemble and are available from the author upon requests. The findings are summarized as follows.

1. In experiment 1, factor estimator and MMA estimator suffer from huge bias. This leads to its bad performance in MSE. FOGLeSs, GMM and eMAPLE estimator incur a small bias but enjoy a big reduction in variance, as shown as their MSE are much smaller than GLS estimator. Our proposed estimators are generally better than FOGLeSs estimator. Lasso is very attractive in small sample, due to the correlation in the factors and the regressors. However, it is much worse than MAPLE estimators when $n = 200$.

2. In experiment 2, MMA, JMA and LASSO estimators perform pretty well and even beat the oracle GLS estimator. The advantage of MAPLE estimator becomes clear when sample size is $n = 200$. MAPLE outperforms FOGLeSs estimator in all cases.

3. In experiment 3, when heterogeneity presents, the performance of the estimators considered is quite mixed. A smaller MSE of one parameter estimator is usually glued with a larger MSE of the other parameter estimator. However, in the third heterogeneity case, MAPLE estimators outperform others in small sample.

4. In experiment 4, we see the clear dominance of MAPLE estimators over other competing ones. Especially, eMAPLE estimator is performing as if it is the oracle GLS estimator in terms of MSE. All competitors perform quite close to GLS. LASSO becomes the worst among all methods.

5. In experiment 5, the dominance of MAPLE estimator remains when heterogeneity exists. Although they are not as good as the oracle GLS, but they are quite competing with the FGLS. LASSO remains the worst among all competitor, but perform slightly better than the naive OLS estimator.

5.6 Experiment 6: rejection probability

We consider to evaluate the size of the tests based on different estimators. We include GLS, FOGLeSs, GMM and our eMAPLE estimators. For tests based on eMAPLE estimator, we only include the LR type test that is appealing in its computation. Other tests are based on the usual t-test statistic. We consider the following data generating process.

$$\text{DGP6-1} : y_i = 2x_i + z'_{1i}\gamma + e_i,$$

where z_{1i} is generated in the same way as that in experiment 4. We report the results based on 1000 replications for sample size n=50 and 200.

Table 1: Rejection Probability: Homogeneous Case

	$n = 50$			$n = 200$		
α	0.01	0.05	0.10	0.01	0.05	0.10
OLS	0.259	0.389	0.468	0.246	0.374	0.456
GLS	0.023	0.076	0.133	0.012	0.061	0.115
FOGLeSs	0.081	0.179	0.252	0.015	0.077	0.148
gMAPLE1	0.034	0.079	0.152	0.013	0.055	0.116
gMAPLE2	0.049	0.100	0.176	0.014	0.061	0.119
eMAPLE	0.027	0.075	0.132	0.013	0.053	0.115

We consider to evaluate the size of the tests. Table 1 presents the rejection probability when there is no heterogeneity. Test based on eMAPLE estimator tends to outperform all other competitors, including that based on oracle GLS estimator. Its rejection probabilities are very close to their nominal levels for both sample sizes. FOGLeSs estimator suffers from big size distortion.

To incorporate heterogeneity, we consider the following design,

$$\text{DGP6-2} : \quad y_i = x_i^{\tau}\beta + z'_{1i}\gamma + u_i \cdot e_i,$$

with

$$u_i = \log\left(3x_i^2\right).$$

The performance of eMAPLE estimator remains satisfactory in the presence of heterogeneity. Although there is a distortion when sample size is 50, it beats GLS estimator when sample size becomes 200. FOGLeSs perform slightly better with heterogeneity, but still have serious size distortion.

6 Empirical Illustration

We illustrate the use of MAPLE estimator in the study of the impact of inherited control on firm performance. We adopt the data set that is originally analyzed by Pérez-González(2006)

Table 2: Rejection Probability: Heterogeneous Case

	$n = 50$			$n = 200$		
α	0.01	0.05	0.10	0.01	0.05	0.10
OLS	0.240	0.366	0.455	0.249	0.386	0.463
GLS	0.017	0.068	0.126	0.013	0.040	0.086
FOGLeSs	0.124	0.235	0.306	0.030	0.090	0.153
gMAPLE1	0.034	0.103	0.166	0.013	0.057	0.111
gMAPLE2	0.044	0.125	0.194	0.015	0.052	0.114
eMAPLE	0.021	0.094	0.159	0.012	0.053	0.114

and subsequently examined by White and Lu (2010). Pérez-Gonzálezuses data from 335 management transitions of publicly traded U.S. corporations to examine whether firms with family related incoming chief executive officers (CEOs) underperform in terms of operating profitability relative to firms with unrelated incoming CEOs. In this application, x equals to 1 if the incoming CEO is related to the departing CEO, to the founder, or to a large shareholder by blood or marriage and otherwise it equals to 0. Operating return on assets (OROA) is used as a measure of firm performance. y is the difference in OROA calculated as the three-year average after succession minus the three-year average before succession. We direct detailed data description to White and Lu (2010).

Following White and Lu (2010), we classify the covariates into firm size, firm's past performance, board characteristics, firm's R&D expenditure, departing CEO's separation conditions and incoming CEO's ownership, and incoming CEO's characteristics. We follow White and Lu (2010) to consider 5 model specifications that correspond to 5 states of the world. We report the estimated weights and e-MAPLE estimate in TABLE 7, together with associated the t-statistic. In TABLE 3, we include the estimator of White and Lu (2010) for comparison.

Table 3: Empirical results: Inherited control

	FOGLeSs	gMPL 1	gMPL2	eMPL
Estimate	-0.0246	-0.0283	-0.0283	-0.0221
95% C.I.	(-0.04409,-0.00510)	(-0.04805,-0.00862)	(-0.04606,-0.01057)	(-0.03300,-0.01200)
95% C.I. length	0.03899	0.03943	0.03550	0.02100
		eMAPLE model probability		
0.1996	0.2001	0.2001	0.2003	0.1999

We find that all the estimates are negative and all 95% confidence intervals are to the left of zero. The implication is that the effect of inherited control on firm performance is significant. This agree with the findings of White and Lu (2010) and Pérez-González(2006). A second finding from Table 3 is that confidence interval based on our eMAPLE estimator is much narrower than those based on FOGLeSs and gMAPLE estimators. Combined with findings in our simulation results, the eMAPLE estimator provides more accurate inference analysis.

7 Conclusion and Future Work

This paper studies the estimation of marginal effect of one economic variable on another, in the presence of large amount of other economic variables. The paper first points out that only small dimensional partial effect parameters have economic policy implication and therefore are economically sensible. Then we set up conditions to identify partial effect parameter of interest in high dimensional structural model. Based on identification of the parameter of interest, we consider the case that the partial effect parameter may be identified in more than one model. I propose two new model averaging estimator to estimate the partial effect estimator based on a GMM-like objective function and an entropy objective function. The two estimators are termed as gMAPLE and eMAPLE estimators. Asymptotic properties of MAPLE estimators are established under a suitable set of conditions. Simulation results show that the MAPLE estimator outperform other competitors in finite sample. An application of the MAPLE estimator to study the effect of inherited control on firm's performance is carried out to illustrate its use. We found that a negative effect does exist which is consistent with earlier findings in the literature. The gain in using MAPLE estimator compared to FOGLeSs is revealed through the shorter confidence interval length.

This paper opens directions for future studies in model averaging in numerous ways. It emphasizes the estimation of parameter of interest in large dimensional model via identification conditions and model averaging techniques. An information based test of the key identification condition, conditional mean independence, is under investigation by the author. A second direction is to apply MAPLE to study the determinants of economic growth following the work of Sala-i-Martin et al (2004). MAPLE can also be extended to the nonparametric and semiparametric models. the only challenge is the identification condition. As an alternative to entropy based approach, empirical likelihood (Owen 1988, 1990, 1991) based approach can be used for MAPLE as well. Moreover, information based variable selection and estimation is another direction to extend the current paper.

Appendix

A Proof of Lemmas

Proof of Lemma 2.7. Partition the coefficient of $\mathbf{z}_i$ as γ_i^τ=$[\gamma_{1i}^\tau, \gamma_{2i}^\tau]$ corresponding to the partition of $\mathbf{z}_i^\tau = \left[\mathbf{z}_i^{1\tau}, \mathbf{z}_i^{2\tau}\right]$ .Under Assumption CMI,

$$\begin{aligned} E\left[y_i|\mathbf{x}_i, \mathbf{z}_{1i}\right] &= \alpha + \mathbf{x}_i\beta + \mathbf{z}_{1i}^\tau\gamma_{1i} + E\left[\mathbf{z}_{2i}^\tau\gamma_{2i}|\mathbf{x}_i, \mathbf{z}_{1i}\right] + E\left[\varepsilon_i|\mathbf{x}_i, \mathbf{z}_{1i}\right] \\ &= \alpha + \mathbf{x}_i\beta + \mathbf{z}_{1i}^\tau\gamma_{1i} + E\left[\mathbf{z}_{2i}^\tau|\mathbf{z}_{1i}\right]\gamma_{2i} \qquad \text{(A.1)} \end{aligned}$$

where $E\left[\varepsilon_i|\mathbf{x}_i, \mathbf{z}_{1i}\right] = 0$ due to strict exogeneity of the regressors. Note that (A.1) implies that β is identified in the regression of y_i on $\mathbf{x}_i$ and $\mathbf{z}_{1i}$, with conditions as specified in Robinson (1988).

Proof of Lemma 2.8. Similar to the proof of Lemma 1, under Assumption WCMI, we can derive that

$$E\left[y_i|\mathbf{x}_i, \mathbf{z}_{1i}\right] = \alpha + \mathbf{x}_i\left(\beta + \eta\right) + \mathbf{z}_{1i}^\tau\gamma_{1i} + E\left[\mathbf{z}_{2i}^\tau|\mathbf{z}_{1i}\right]\gamma_{2i}.$$

Thus $(\beta + \eta)$ would be identified in the regression of y_i on $\mathbf{x}_i$ and $\mathbf{z}_{1i}$, with conditions as specified in Robinson (1988). Since $||\eta|| = o\left(n^{-1/2}\right)$, with sample size gets large, Robinson's (1988) estimator of $(\beta + \eta)$ will converge to β.

B Derivation of Some Equations

This Appendix provides derivation of equation (3.11), (3.12), (3.14), (3.15).

The FOCs of the Lagrangian in (3.10) are:

$$\begin{aligned} \frac{\partial L}{\partial p_{is}} &= -\hat{q}_s \log \hat{p}_{is} - \hat{q}_s - \hat{\mu}_s - \hat{\eta}_s^\tau g_s\left(d_i, \theta\right) = 0, && \text{(B.1)} \\ \frac{\partial L}{\partial q_s} &= -\sum_{i=1}^n \hat{p}_{is} \log \hat{p}_{is} - 1 - \log \hat{q}_s - \hat{\xi} = 0, && \text{(B.2)} \\ \frac{\partial L}{\partial \mu_s} &= \sum_{i=1}^n \hat{p}_{is} - 1 = 0, && \text{(B.3)} \\ \frac{\partial L}{\partial \eta_s} &= \sum_{i=1}^n \hat{p}_{is} g_s\left(d_i; \hat{\theta}\right) = 0, && \text{(B.4)} \\ \frac{\partial L}{\partial \xi} &= \sum_{s=1}^S \hat{q}_s - 1 = 0, && \text{(B.5)} \\ \frac{\partial L}{\partial \theta} &= \sum_{s=1}^S \hat{\eta}_s^\tau \sum_{i=1}^n \hat{p}_{is} \nabla_\theta g_s\left(d_i, \hat{\theta}\right) = 0. && \text{(B.6)} \end{aligned}$$

(i) Derivation of (3.11). From (B.2), we get

$$\hat{q}_s = \exp\left(-\sum_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is} - 1 - \hat{\xi}\right).$$

Combined this equation with (B.5), it gives (3.11).

(ii) Derivation of (3.12). Using (B.1), It's straightforward to show that

$$\hat{p}_{is} = \exp\left(-\frac{\hat{q}_s - \hat{\mu}_s - \hat{\eta}_s^{\tau} g_s\left(d_i, \theta\right)}{\hat{q}_s}\right),$$

With normalization in (B.3), we have

$$\begin{aligned}
\hat{p}_{is} &= \frac{\exp\left(\frac{-\hat{q}_s - \hat{\mu}_s - \hat{\eta}_s^{\tau} g_s(d_i,\theta)}{\hat{q}_s}\right)}{\sum_{i=1}^{n} \exp\left(\frac{-\hat{q}_s - \hat{\mu}_s - \hat{\eta}_s^{\tau} g_s(d_i,\theta)}{\hat{q}_s}\right)} \\
&= \frac{\exp\left(\frac{-\hat{\eta}_s^{\tau} g_s(d_i,\theta)}{\hat{q}_s}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{-\hat{\eta}_s^{\tau} g_s(d_i,\theta)}{\hat{q}_s}\right)} \\
&= \frac{1}{\Upsilon_s\left(\lambda_s, \theta\right)} \exp\left[-\lambda_s^{\tau} g_s\left(d_i; \theta\right)\right],
\end{aligned}$$

with $\lambda_s^{\tau} = \hat{\eta}_s^{\tau} / \hat{q}_s$, and $\Upsilon_s\left(\lambda_s, \theta\right) = \sum_{i=1}^{n} \exp\left[-\lambda_s^{\tau} g_s\left(d_i; \theta\right)\right]$. This proves (3.12).

(iii) Derivation of (3.14). Plugging (3.12) into (B.4) results

$$\sum_{i=1}^{n} \frac{g_s\left(d_i; \theta\right)}{\Upsilon_s\left(\lambda_s, \theta\right)} \exp\left[-\lambda_s^{\tau} g_s\left(d_i; \theta\right)\right] = 0.$$

Since $\Upsilon_s\left(\lambda_s, \theta\right) > 0$, this leads to (3.14).

(iv) Derivation of (3.15). We show this results in two steps. (a) Note that

$$\begin{aligned}
-\sum_{s=1}^{S} \hat{q}_s \log \hat{q}_s &= -\sum_{s=1}^{S} \hat{q}_s \log \left[\frac{\exp\left(-\sum\limits_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is}\right)}{\sum\limits_{s=1}^{S} \exp\left(-\sum\limits_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is}\right)}\right] \\
&= \sum_{i=1}^{n} \hat{q}_s \sum_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is} \\
&\quad + \sum_{s=1}^{S} \hat{q}_s \log \sum_{s=1}^{S} \exp\left(-\sum_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is}\right) \\
&= \sum_{i=1}^{n} \hat{q}_s \sum_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is} + \log \sum_{s=1}^{S} \exp\left(-\sum_{i=1}^{n} \hat{p}_{is} \log \hat{p}_{is}\right).
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{JE}\left(\theta\right) &= -\sum_{s=1}^{S}\sum_{i=1}^{n}\hat{q}_s\hat{p}_{is}\log\hat{p}_{is}-\sum_{s=1}^{S}\hat{q}_s\log\hat{q}_s \\
&= -\sum_{s=1}^{S}\sum_{i=1}^{n}\hat{q}_s\hat{p}_{is}\log\hat{p}_{is}+\sum_{i=1}^{n}\hat{q}_s\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}+\log\sum_{s=1}^{S}\exp\left(-\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}\right) \\
&= \log\sum_{s=1}^{S}\exp\left(-\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is}\right).
\end{aligned}$$

(b) Next,

$$\begin{aligned}
& -\sum_{i=1}^{n}\hat{p}_{is}\log\hat{p}_{is} \\
&= -\sum_{i=1}^{n}\hat{p}_{is}\log\frac{\exp\left[-\lambda_s^{\tau}g_s\left(d_i;\theta\right)\right]}{\Upsilon_s\left(\lambda_s,\theta\right)} \\
&= \lambda_s^{\tau}\sum_{i=1}^{n}\hat{p}_{is}g_s\left(d_i;\theta\right)+\sum_{i=1}^{n}\hat{p}_{is}\log\Upsilon_s\left(\lambda_s,\theta\right) \\
&= \log\Upsilon_s\left(\lambda_s,\theta\right),
\end{aligned}$$

where we have used (B.4).

Putting (a) and (b) together leads to

$$\mathrm{JE}\left(\theta\right)=\log\sum_{s=1}^{S}\Upsilon_s\left(\lambda_s,\theta\right)=\log\Upsilon\left(\lambda,\theta\right),$$

which proves (3.15).

C Proof of Auxiliary Lemmas

Lemma C.1 Under Assumption B.1-5, $\sup_{\theta\in\boldsymbol{\Theta},s=1,\ldots,S,d_i}\left|\lambda_s^{\tau}g_s\left(d_i;\theta\right)\right|=o_p\left(1\right)$.

Proof. It follows from Lemma 3 of Owen (1990) or Lemma D.2 of Kitamura, Tripathi and Ahn (2004).

Lemma C.2 Under Assumption B.1-5, $\sup_{\theta\in\boldsymbol{\Theta}}||\lambda_s^{\tau}\left(\theta\right)-V^{-1}\left(s,\theta\right)Eg_s\left(d,\theta\right)||=o_p\left(\|\lambda_s\|\right).$

Proof. By (3.14), λ_s^{τ} solves

$$\sum_{i=1}^{n}g_s\left(d_i;\theta\right)\exp\left[-\lambda_s^{\tau}g_s\left(d_i;\theta\right)\right]=0.$$

By Taylor's Theorem, there exists $\bar{\lambda}_s$ lying between 0 and λ_s such that

$$0 = \sum_{i=1}^{n} g_s(d_i;\theta)\left\{1 - \lambda_s^\tau g_s(d_i;\theta) + \frac{\left(\bar{\lambda}_s^\tau g_s(d_i;\theta)\right)^2}{2}\right\}.$$

Rearranging terms leads to

$$\begin{aligned}\lambda_s &= \left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1}\sum_{i=1}^{n} g_s(d_i;\theta)/n + \left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]\sum_{i=1}^{n} g_s(d_i;\theta)\frac{\left(\bar{\lambda}_s^\tau g_s(d_i;\theta)\right)^2}{2n}\\ &\equiv l_1 + l_2,\end{aligned}$$

where

$$\begin{aligned} l_1 &= V^{-1}(s,\theta)\, Eg_s(d,\theta) + \left[\hat{V}^{-1}(s,\theta) - V^{-1}(s,\theta)\right] Eg_s(d,\theta) + \hat{V}^{-1}(s,\theta)\left[\sum_{i=1}^{n} g_s(d_i;\theta)/n - Eg_s(d,\theta)\right]\\ &= V^{-1}(s,\theta)\, Eg_s(d,\theta) + o_p(1),\end{aligned}$$

by Assumption B.2 and B.4, and

$$\begin{aligned}\|l_2\| &= \left\|\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1}\right\|\left\|\sum_{i=1}^{n} g_s(d_i;\theta)\frac{\left(\bar{\lambda}_s^\tau g_s(d_i;\theta)\right)^2}{2n}\right\|\\ &\le \left\|\hat{V}^{-1}(s,\theta)\right\|\left\|\sum_{i=1}^{n} g_s^2(d_i;\theta)\right\|^{1/2}\left\|\sum_{i=1}^{n}\left[\bar{\lambda}_s^\tau g_s(d_i;\theta)\right]^4\right\|^{1/2}/n\\ &\le \left\|\hat{V}^{-1}(s,\theta)\right\|\left\|\sum_{i=1}^{n} g_s^2(d_i;\theta)\right\|^{1/2}\sup\left\|\left[\bar{\lambda}_s^\tau g_s(d_i;\theta)\right]^4\right\|\\ &\le \left\|\hat{V}^{-1}(s,\theta)\right\|\left\|\sum_{i=1}^{n} g_s^2(d_i;\theta)\right\|^{1/2}\sup\left\|\bar{\lambda}_s^\tau\right\|^4\left\|g_s(d_i;\theta)\right\|^4\\ &\le \left\|\hat{V}^{-1}(s,\theta)\right\|\left\|\sum_{i=1}^{n} g_s^2(d_i;\theta)\right\|^{1/2}\left(\sup\left\|\lambda_s\right\|\left\|g_s(d_i;\theta)\right\|\right)^4\\ &= o(1),\end{aligned}$$

by Cauchy-Schwartz's inequality, Assumption B.2 and Lemma C.1.

Lemma C.3 Under Assumption B.1-5, $\sup_{\theta\in\mathbf{\Theta}}||\nabla_\theta\lambda_s^\tau(\theta) - V^{-1}(s,\theta)\, D(s,\theta)|| = o_p(1)$.

Proof. By (3.14), λ_s^τ solves

$$\sum_{i=1}^{n} g_s(d_i;\theta)\exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] = 0.$$

Differentiating both sides with respect to θ gives

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \nabla_\theta g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \\
&\quad - \sum_{i=1}^{n} g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \nabla_\theta \lambda_s^\tau(\theta) g_s(d_i;\theta) \\
&\quad - \sum_{i=1}^{n} g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \lambda_s^\tau(\theta) \nabla_\theta g_s(d_i;\theta) \\
&\equiv l_1 - l_2 \nabla_\theta \lambda_s^\tau(\theta) + l_3.
\end{aligned}$$

The proof is completed after showing that (i) $\sup_{\theta\in\mathbf{\Theta}} ||l_1/n - D(s,\theta)|| = o_p(1)$; (ii) $\sup_{\theta\in\mathbf{\Theta}} ||l_2/n - V(s,\theta)|| = o_p(1)$;(iii) $\sup_{\theta\in\mathbf{\Theta}} ||l_3/n|| = o_p(1)$ and an application of triangular inequality.

We show (i) first. Note that

$$\begin{aligned}
l_1/n &= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g_s(d_i;\theta) + o_p(1) \\
&= D(s,\theta) + o_p(1),
\end{aligned}$$

by Lemma C.1 and a Law of Large Numbers.

We then show (ii). It is easily seen that

$$\begin{aligned}
l_2/n &= \frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] g_s(d_i;\theta) \\
&= \frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta) g_s^\tau(d_i;\theta) + o_p(1) \\
&= V(s,\theta) + o_p(1),
\end{aligned}$$

by Lemma C.1 and a Law of Large Numbers.

Finally, we show (iii).

$$\begin{aligned}
\|l_3/n\| &= \left\| \frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta) \exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \lambda_s^\tau(\theta) \nabla_\theta g_s(d_i;\theta) \right\| \\
&\leq \left\| \frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta) \lambda_s^\tau(\theta) \nabla_\theta g_s(d_i;\theta) \right\| + o_p(1) \\
&\leq \left\| \frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta) \right\| \sup_{\theta\in\mathbf{\Theta}} \|\lambda_s^\tau(\theta)\| \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g_s(d_i;\theta) \right\| + o_p(1) \\
&= o_p(1)
\end{aligned}$$

by Lemma C.1, Assumption B.3, B.4 and a Law of large numbers.

D Proof of Main Theorems

Proof of Theorem 4.1.

Define $\mathrm{JE}_0\left(\theta\right) = -\frac{1}{S}\sum_{s=1}^{S} Eg_s^\tau\left(d_i;\theta\right) V\left(s,\theta\right) Eg_s\left(d_i;\theta\right) \equiv -\frac{1}{S}\sum_{s=1}^{S} h_s\left(s,\theta\right)$. By Theorem 4.1.1 of Amemiya (1985), to prove $\hat{\theta} \to_p \theta_0$, we need only show that (i) $\mathrm{JE}_0\left(\theta\right)$ is uniquely maximized at $\theta = \theta_0$ and (ii) $\sup_{\theta\in\mathbf{\Theta}}\left|\mathrm{JE}_n\left(\theta\right) - \mathrm{JE}_0\left(\theta\right)\right| \to_p 0$.

We first prove (i). By Assumption B.1 and B.4, $h_s\left(s,\theta\right) > 0$ for any $\theta \in \mathbf{\Theta}\backslash\left\{\theta_0\right\}$. However, $h_s\left(s,\theta_0\right) = Eg_s^\tau\left(d_i;\theta_0\right) V\left(s,\theta_0\right) Eg_s\left(d_i;\theta_0\right) = 0$ by (3.2). Thus $\mathrm{JE}_0\left(\theta\right) \geq 0$ with the unique minimizer $\theta = \theta_0$.

Next we show (ii). Applying Lemma C.2, write

$$\sum_{i=1}^{n} \lambda_s^\tau\left(\theta\right) g_s\left(d_i;\theta\right) = \sum_{i=1}^{n} Eg_s^\tau\left(d_i;\theta\right) V^{-1}\left(s,\theta\right) g_s\left(d_i;\theta\right) + o_p\left(1\right).$$

This leads to

$$\begin{aligned}
\left|\mathrm{JE}_n\left(\theta\right) - \mathrm{JE}_0\left(\theta\right)\right| &= \frac{1}{S}\left|\sum_{s=1}^{S} Eg_s^\tau\left(d_i;\theta\right) V\left(s,\theta\right) Eg_s\left(d_i;\theta\right) - \sum_{s=1}^{S} Eg_s^\tau\left(d_i;\theta\right) V^{-1}\left(s,\theta\right)\frac{1}{n}\sum_{i=1}^{n} g_s\left(d_i;\theta\right) + o_p\left(1\right)\right| \\
&= \frac{1}{S}\left|\sum_{s=1}^{S} Eg_s^\tau\left(d_i;\theta\right) V\left(s,\theta\right)\left[Eg_s\left(d_i;\theta\right) - \frac{1}{n}\sum_{i=1}^{n} g_s\left(d_i;\theta\right)\right] + o_p\left(1\right)\right| \\
&\leq \frac{1}{S}\sum_{s=1}^{S}\left|Eg_s^\tau\left(d_i;\theta\right) V\left(s,\theta\right) o_p\left(1\right) + o_p\left(1\right)\right| = o_p\left(1\right),
\end{aligned}$$

where we have used Assumption B.2, B.4 and a Law of Large numbers.

Proof of Theorem 4.2. Note that FOC of (3.16) is

$$\nabla_\theta \mathrm{JE}_n\left(\hat{\theta}\right) = 0.$$

By Taylor's Theorem, there exisit $\bar{\theta}$ lying between $\hat{\theta}$ and θ_0, s.t.,

$$0 = \nabla_\theta \mathrm{JE}_n\left(\hat{\theta}\right) = \nabla_\theta \mathrm{JE}_n\left(\theta_0\right) + \nabla_{\theta\theta}\mathrm{JE}_n\left(\bar{\theta}\right)\left(\hat{\theta} - \theta_0\right).$$

This leads to

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\left[\nabla_{\theta\theta}\mathrm{JE}_n\left(\bar{\theta}\right)\right]^{-1}\left[\sqrt{n}\nabla_\theta \mathrm{JE}_n\left(\theta_0\right)\right].$$

We complete the proof by showing that (i) $\sqrt{n}\nabla_\theta \mathrm{JE}_n\left(\theta_0\right) \to N\left(0, \frac{1}{S^2} I\left(\theta_0\right)\right)$ and (ii) $-\nabla_{\theta\theta}\mathrm{JE}_n\left(\bar{\theta}\right) \to_p \frac{1}{S} J\left(\theta_0\right)$ and an application of Slutsky's Theorem.

(1) We first prove $\sqrt{n}\nabla_\theta \mathrm{JE}_n\left(\theta_0\right) \to N\left(0, \frac{1}{S^2} I^{-1}\left(\theta_0\right)\right)$. Note first that by (3.14), λ_s^τ solves

$$\sum_{i=1}^{n} g_s\left(d_i;\theta\right)\exp\left[-\lambda_s^\tau g_s\left(d_i;\theta\right)\right] = 0.$$

Thus, we have

$$\begin{aligned}
\nabla_\theta \mathrm{JE}_n(\theta_0) &= \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n} \nabla_\theta \lambda_s^\tau(\theta)\, g_s(d_i;\theta)\exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \\
&\quad + \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n} \lambda_s^\tau(\theta)\,\nabla_\theta g_s(d_i;\theta)\exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \\
&= \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n} \lambda_s^\tau(\theta)\,\nabla_\theta g_s(d_i;\theta)\exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \\
&= \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n} g_s^\tau(d_i;\theta)\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1}\nabla_\theta g_s(d_i;\theta)\exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] + o_p(1) \\
&\equiv \hat{U} + o_p(1)
\end{aligned}$$

We need to show that $n^{1/2}\hat{U} \to N\left(0, \frac{1}{S^2} I(\theta_0)\right)$.

Since $\exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] = 1 - \lambda_s^\tau g_s(d_i;\theta) + o_p(1)$ by Assumption B.5. We have

$$\begin{aligned}
n^{1/2}\hat{U} &= n^{-1/2}\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n} g_s^\tau(d_i;\theta)\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1}\nabla_\theta g_s(d_i;\theta) \\
&\quad - n^{-1/2}\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n} g_s^\tau(d_i;\theta)\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1}\nabla_\theta g_s(d_i;\theta)\,\lambda_s^\tau g_s(d_i;\theta) \\
&= n^{-1/2}\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n} g_s^\tau(d_i;\theta)\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1}\nabla_\theta g_s(d_i;\theta) + o_p(1) \\
&= n^{-1/2}\hat{U}_1 + o_p(1).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
n^{-1/2}\hat{U}_1 &= \frac{1}{\sqrt{n}}\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n}\left(\sum_{i=1}^{n} g_s^\tau(d_i;\theta)\right)\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1}\nabla_\theta g_s(d_i;\theta)/n \\
&= \frac{1}{\sqrt{n}}\frac{1}{S}\sum_{i=1}^{n}\left\{\sum_{s=1}^{S} g_s^\tau(d_i;\theta)\left[\frac{1}{n}\sum_{i=1}^{n} g_s(d_i;\theta)\, g_s^\tau(d_i;\theta)\right]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta g_s(d_i;\theta)\right)\right\} \\
&= \frac{1}{\sqrt{n}}\frac{1}{S}\sum_{i=1}^{n}\left\{\sum_{s=1}^{S} g_s^\tau(d_i;\theta)\, V^{-1}(s,\theta)\, G(s,\theta)\right\} + o_p(1) \\
&\equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_i
\end{aligned}$$

where

$$\zeta_i = \frac{1}{S}\sum_{s=1}^{S} g_s^\tau(d_i;\theta)\, V^{-1}(s,\theta)\, G(s,\theta).$$

It is easily seen that ζ_i is an m.d.s. with variance

$$\begin{aligned}
E\left[\zeta_i\zeta_i^{\tau}\right] &= E\left\{\sum_{s=1}^{S} g_s^{\tau}\left(d_i;\theta\right)V^{-1}\left(s,\theta\right)G\left(s,\theta\right)\right\}^2 . \\
&= \frac{1}{S^2}\sum_{s,t=1}^{S} G^{\tau}\left(s,\theta\right)V^{-1}\left(s,\theta\right)E\left[g_s\left(d_i;\theta\right)g_t^{\tau}\left(d_i;\theta\right)\right]V^{-1}\left(t,\theta\right)G\left(t,\theta\right) \\
&= \frac{1}{S^2}G^{\tau}V^{-1}\Omega V^{-1}G \qquad \left(=\frac{1}{S^2}G^{\tau}V^{-1}G\right) \\
&\equiv \frac{1}{S^2}I\left(\theta_0\right).
\end{aligned}$$

By a CLT for vector ergodic stationary m.d.s. (see, for example, Billingsley, 1961), we have

$$n^{1/2}\hat{U}_1 \rightarrow^d N\left(0,\frac{1}{S^2}I\left(\theta_0\right)\right).$$

(2) We then show that $-\nabla_{\theta\theta}\mathrm{JE}_n\left(\bar{\theta}\right)\rightarrow_p J\left(\theta_0\right)$. First,

$$\begin{aligned}
-nS\nabla_{\theta\theta}\mathrm{JE}_n\left(\theta_0\right) &= \sum_{s=1}^{S}\sum_{i=1}^{n}\nabla_{\theta}\lambda_s^{\tau}\left(\theta\right)g_s\left(d_i;\theta\right)\exp\left[-\lambda_s^{\tau}g_s\left(d_i;\theta\right)\right]\nabla_{\theta}\left[\lambda_s^{\tau}\left(\theta\right)g_s\left(d_i;\theta\right)\right] \\
&\quad +\sum_{s=1}^{S}\sum_{i=1}^{n}\nabla_{\theta}\lambda_s^{\tau}\left(\theta\right)\nabla_{\theta}g_s\left(d_i;\theta\right)\exp\left[-\lambda_s^{\tau}g_s\left(d_i;\theta\right)\right] \\
&\quad +\sum_{s=1}^{S}\sum_{i=1}^{n}\lambda_s^{\tau}\left(\theta\right)\nabla_{\theta\theta}g_s\left(d_i;\theta\right)\exp\left[-\lambda_s^{\tau}g_s\left(d_i;\theta\right)\right] \\
&\equiv u_1+u_2+u_3.
\end{aligned}$$

We show that (i) $\left\|u_1/n\right\|=o_p\left(1\right)$, (ii) $\left\|u_2/\left(nS\right)-J\left(\theta_0\right)\right\|=o_p\left(1\right)$, and (iii) $\left\|u_3/n\right\|=o_p\left(1\right)$.

We first show (i) $\left\|u_1/n\right\|=o_p\left(1\right)$.

$$\begin{aligned}
\left\|u_1/n\right\| &= \left\|\frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n}\nabla_{\theta}\lambda_s^{\tau}\left(\theta\right)g_s\left(d_i;\theta\right)\exp\left[-\lambda_s^{\tau}g_s\left(d_i;\theta\right)\right]\nabla_{\theta}\left[\lambda_s^{\tau}\left(\theta\right)g_s\left(d_i;\theta\right)\right]\right\| \\
&\leq \frac{1}{S}\sum_{s=1}^{S}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{\theta}\lambda_s^{\tau}\left(\theta\right)g_s\left(d_i;\theta\right)\nabla_{\theta}\left[\lambda_s^{\tau}\left(\theta\right)g_s\left(d_i;\theta\right)\right]\right\|+o_p\left(1\right) \\
&\leq \frac{1}{S}\sum_{s=1}^{S}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{\theta}\lambda_s^{\tau}\left(\theta\right)\right\|\left\|\frac{1}{n}\sum_{i=1}^{n}g_s\left(d_i;\theta\right)\right\| \\
&\quad \times\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_{\theta}\lambda_s^{\tau}\left(\theta\right)\right]g_s\left(d_i;\theta\right)\right\|+\left\|\frac{1}{n}\sum_{i=1}^{n}\lambda_s^{\tau}\left(\theta\right)\nabla_{\theta}g_s\left(d_i;\theta\right)\right\|\right\}+o_p\left(1\right) \\
&\leq o_p\left(1\right),
\end{aligned}$$

by Assumption B.3, B.4 and Lemma C.1, C.2.

We next show (ii) $\|u_2/(nS) - J(\theta_0)\| = o_p(1)$. Note that by Lemma C.3, we have

$$\begin{aligned}
u_2/(nS) &= \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n}\nabla_\theta\lambda_s^\tau(\theta)\nabla_\theta g_s(d_i;\theta)\exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right] \\
&= \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n}\nabla_\theta\lambda_s^\tau(\theta)\nabla_\theta g_s(d_i;\theta) + o_p(1) \\
&= \frac{1}{nS}\sum_{s=1}^{S}\sum_{i=1}^{n}G^\tau(s,\theta)V^{-1}(s,\theta)\nabla_\theta g_s(d_i;\theta) + o_p(1) \\
&= \frac{1}{S}\sum_{s=1}^{S}G^\tau(s,\theta)V^{-1}(s,\theta)\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta g_s(d_i;\theta)\right) + o_p(1) \\
&= \frac{1}{S}\sum_{s=1}^{S}G^\tau(s,\theta)V^{-1}(s,\theta)G(s,\theta) + o_p(1), \\
&= \frac{1}{S}G^\tau V^{-1}G + o_p(1) = \frac{1}{S}J(\theta_0) + o_p(1).
\end{aligned}$$

by Assumption B.4 and a Law of Large Numbers.

Finally we show (iii) $\|u_3/n\| = o_p(1)$.

$$\begin{aligned}
\|u_3/n\| &= \left\|\frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n}\lambda_s^\tau(\theta)\nabla_{\theta\theta}g_s(d_i;\theta)\exp\left[-\lambda_s^\tau g_s(d_i;\theta)\right]\right\| \\
&\leq \sum_{s=1}^{S}\left\|\frac{1}{n}\sum_{i=1}^{n}\lambda_s^\tau(\theta)\nabla_{\theta\theta}g_s(d_i;\theta)\right\| + o_p(1) \\
&\leq \sum_{s=1}^{S}\left\|\frac{1}{n}\sum_{i=1}^{n}\lambda_s^\tau(\theta)\right\|\left\|\frac{1}{n}\nabla_{\theta\theta}g_s(d_i;\theta)\right\| + o_p(1) \\
&\leq o_p(1),
\end{aligned}$$

by Assumption B.4 and Lemma C.2.

References

Altonji, J. G., and Segal, L. M. (1996), “Small Sample Bias in GMM Estimation of Covariance Structures,” *Journal of Business and Economic Statistics*, Vol. 14, 353-366.

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

Ashenfelter, O. (1978), “Estimating the Effect of Training Programs on Earnings”, *The Review of Economics and Statistics*, Vol. 60, No. 1, 47-57

Bai, J. and S. Ng (2010), “Instrumental Variable Estimation in a Data Rich Environment,” *Econometric Theory*, 26:6, 1577-1606.

Bates, J. and C.W.J., Granger (1969), “The Combination of Forecasts,” *Operations Research Quarterly* 20 (4): 451–468.

Belloni, A., V. Chernozhukov, and C. Hansen (2011), “LASSO Methods for Gaussian Instrumental Variables Models”, Working paper, MIT, Econ. Dept.

Berk, R., L., Brown, A., Buja, K., Zhang and L. Zhao (2011). “Valid Post-Selection Inference,” Statistics Department, Wharton School, University of Pennsylvania, Discussion Paper.

Berk, R., L., Brown and L., Zhao (2009). “Statistical inference after model selection,” *Journal of Quantitative Criminology*, 26, 217-236.

Billingsley, P. (1961), “The Lindeberg-Levy Theorem for Martingales,” in *Proceedings of the American Mathematical Society*, Vol. 12, 788-792.

Box, G.E.P. (1979), “Robustness in the strategy of scientific model building,” in *Robustness in Statistics*, R.L. Launer and G.N. Wilkinson, Editors. Academic Press: New York.

Breiman, L. (1996): “Bagging Predictors,” *Machine Learning*, 36, 105–139.

Bühlmann, P., and B. Yu (2002): “Analyzing Bagging,” *The Annals of Statistics*, 30, 927–961.

Chen, X., E., Tamer, and A., Torgovitsky (2011). “Sensitivity Analysis in a Semiparametric Likelihood Model: A Partial Identification Approach,” Department of Economics, Yale University, Working paper.

Clyde, M. and E.I., George (2004), “Model Uncertainty,” Statistical Science, Vol. 19, No. 1, 81-94.

Cover, T.M. and J.A., Thomas (2006), *Elements of Information Theory*, John Wiley & Sons, Inc.

De Jong, S., and H.A.L. Kiers (1992), “Principal covariate regression,” *Chemometrics and Intelligent Laboratory Systems* 14, pp. 155-164.

Diebold, F.X., Rudebusch, G.D. and Aruoba, B. (2006), “The Macroeconomy and the Yield Curve: A Dynamic Latent Factor Approach,” *Journal of Econometrics*, 131, 309-338.

Eicher, T.S., A., Lenkoski and A.E., Raftery (2009), “Bayesian Model Averaging and Endogeneity Under Model Uncertainty: An Application to Development Determinants,” Working Paper no. 94 Center for Statistics and the Social Sciences, University of Washington.

Fan, J. and R. Li (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of American Statistical Association.* 96 1348-1360.

Fisher, I. (1973), "I Discovered the Phillips Curve: A Statistical Relation between Unemployment and Price Changes," *The Journal of Political Economy*, Vol. 81, 496-502. Reprint of 1926 article by Irving Fisher.

Friedman, M. (1957), *A Theory of the Consumption Function*, Princeton University Press.

Galbraith, J.W. and V. Zinde-Walsh (2010), "Reduced-Dimension Control Regression," Department of Economics, McGill University, Working Paper.

Gernert, D. (2007), "Ockham's Razor and Its Improper Use," *Journal of Scientific Exploration*, Vol. 21, No. 1, pp. 135-40.

Golan, A., G., Judge, and D., Miller (1996), *Maximum Entropy Econometrics: Robust Eestimation with Limited Data*, John Wiley & Sons

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, Vol. 66, 315-331.

Hahn, J. (2004), "Functional Restriction and Efficiency in Causal Inference," *The Review of Economics and Statistics*, 86, 73-76.

Hansen, B.E. (2007), "Least Squares Model Averaging," *Econometrica*, Vol. 75, 1175-1189.

Hansen, B.E. (2008), "Least-square Forecast Averaging," *Journal of Econometrics*, Vol. 146, 342-350.

Hansen, B.E. (2009), "Averaging Estimators for Regression with a Possible Structure Break," *Econometric Theory* 35, 1498-1514.

Hansen, B.E. (2010), "Averaging Estimators for Autoregressions with a Near Unit Root," *Journal of Econometrics*, Vol. 158, 142-155.

Hansen, B.E., and J. Racine (2011), "Jackknife Model Averaging," *Journal of Econometrics*, forthcoming.

Hansen, L. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

Hansen, L., J. Heaton and A. Yaron (1996), "Finite Sample Properites of Alternative GMM Estimators," *Journal of Business and Economic Statistics*, Vol. 14, 262-280.

Hoeting, J.A., D, Madigan, A.E., Raftery and C.T., Volinsky (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, Vol. 14, No. 4, 382-417.

Hsiao, C., Q., Li and J., Racine (2007), "A consistent model specification test with mixed discrete and continuous data," *Journal of Econometrics* 140, 802-826.

Huang, J., Horowitz, J. and Ma, S. (2008). "Asymptotic properties of bridge estimators in sparse high-dimensional regression models." *The Annals of Statistics* 36, 587-613.

Imbens, G. and C. Maski (2004), "Confidence Intervals for Partially Identified Parameters," *Econometrica*, Vol. 72, 1845-1857.

2004, pp. 1845-1857.

Imbens, G., R.H., Spady, and P. Johnson (1998), "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333-357.

Keynes, J. M. (1936), *The General Theory of Unemployment, Interest and Inflation*

Kitamura, Y., (2006), "Empirical Likelihood Methods in Econometrics: Theory and Practice," *Advances in Economics and Econometrics: Theory and Applications*, Nineth World Congress, Edited by R. Blundell, W., Newey, and T., Persson.

Kitamura, Y. and M., Stutzer (1997), "An Information-theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861-874.

Kitamura, Y., G., Tripathi, and H., Ahn (2004), "Empirical Likelihood-based Inference in Conditional Moment Restriction Models," *Econometrica*, Vol. 72, 1667-1714.

Kruger, A.B., (1993), "How Computers Have Changed the Wage Structure: Evidence from Microdata," 1984-1989, *The Quarterly Journal of Economics*, Vol. 108, No. 1, 33-60.

Lee, T.H., Y., Tu and A., Ullah (2011a),"Nonparametric and Semiparametric Regressions Subject to Monotonicity Constraints: Estimation and Forecasting," Working paper, UC, Riverside, Econ. Dept.

Lee, T.H., Y., Tu and A., Ullah (2011b),"Forecasting Equity Premium: Global Historical Mean Versus Local Historical Mean and Constraints," Working paper, UC, Riverside, Econ. Dept.

Leeb, H. and B.M., Pötscher (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory* 21, 29-59.

Leeb, H. and B.M., Pötscher (2006), "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators ?", *Annals of Statistics* 34, 2554-2591

Leeb, H. and B.M., Pötscher (2008a), "Recent Developments in Model Selection and Related Areas," *Econometric Theory*, 24, 319-322.

Leeb, H. and B.M., Pötscher (2008b), "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators ?" *Econometric Theory*, 24, 338-376.

Leeb, H. and B.M., Pötscher (2008c), "Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator," *Journal of Econometrics* 142, 201-211.

Leeb, H. and B.M., Pötscher (2009), "On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding," *Journal of Multivariate Analysis* 100, 2065-2082.

Li, Q., Wang, S. (1998), "A simple consistent bootstrap test for a parametric regression function," *Journal of Econometrics*, 87, 145-165.

Manski, C.F. (1995), *Identification Problems in the Social Sciences*, Harvard University Press

Manski, C.F. (2003), *Partial Identification of Probability Distributions*, Springer-Verlag

Manski, C.F. (2007), *Identification for Prediction and Decision*, Harvard University Press

Miler, Alan (2002), *Subset Selection in Regression*, Chapman & Hall/CRC.

Mincer, J. (1976). "Unemployment Effects of Minimum Wages," *Journal of Political Economy* 84, 87-104.

Newey, W. K., and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics* Vol. 4, Edited by Daniel McFadden and Robert Engle. Amsterdam, The Netherlands: Elsevier, North-Holland, 1999, chapter 36, pp. 2113-2245.

Owen, A. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika* 75, 237-49.

Owen, A. (1990), "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics* 18, 90-120.

Owen, A. (1991), "Empirical Likelihood for Linear Models," *The Annals of Statistics* 19, 1725-1747.

Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine* 2 (6): 559-572.

Pérez-González, F. (2006), "Inherited Control and Firm Perfance," American Economic Review 96, 1559-1588.

Phillips, A. W. (1958), "The Relationship between Unemployment and the Rate of Change of Money Wages in the United Kingdom 1861-1957," *Economica* 25 (100): 283-99.

Pötscher, B.M. (2009), "Confidence Sets Based on Sparse Estimators Are Necessarily Large," *Sankhya* 71-A, 1-18.

Pötscher, B.M. and U., Schneider (2009), "On the Distribution of the Adaptive LASSO Estimator," *Journal of Statistical Planning and Inference* 139, 2775-2790.

Pötscher, B.M. and U., Schneider (2010), "Confidence Sets Based on Penalized Maximum Likelihood Estimators," *Electronic Journal of Statistics* 10, 334-360.

Qin J. and J. Lawless (1994), "Empirical Likelihood and General estimating Equations," *The Annals of Statistics* 22, 300-325.

Racine, J. and B. Hansen (2011), "Jackknife Model Averaging,", *Journal of Econometrics, forthcoming*

Rapach, D., J., Strauss and G. Zhou (2010), "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy," *Review of Financial Studies* 23, 821-862.

Robinson, P.M. (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, Vol. 56, 931-954.

Romano, J.P. and A. Shaikh (2008), "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference*, Vol. 138, 2786-2807.

Romano, J.P. and A. Shaikh (2010), "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, Vol. 78, 169-211.

Sala-i-Martin, X., G., Doppelhofer and R., Miller (2004), "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *American Economic Review*, 94(4): 813–835.

Santos A. (2011), "Inference in Nonparametric Instrumental Variables with Partial Identification," *Econometrica*, forthcoming.

Spearman, C. (1904), "General intelligence, objectively determined and measured," *American Journal of Psychology*, 15, 201-293.

Stock, J.D., and M. Watson (2010), *Introduction to Econometrics*, 3ed. Addison Wesley.

Su, L. and H. White (2007a), "A Consistent Characteristic Function-Based Test for Conditional Independence," *Journal of Econometrics*, 141, 807-834.

Su, L. and H. White (2007b), "Testing Conditional Independence via Empirical Likelihood," UCSD Dept. of Economics Discussion Paper.

Su, L. and H. White (2008), "A Nonparametric Hellinger Metric Test for Conditional Independence," *Econometric Theory*, 24, 829-864.

Tibshirani, R. (1996), "Regression shrinkage and selection via the Lasso," *Journal of Royal Statistical Society* Ser. B 58 267–288.

Tu, Y. (2011a), "Nonparametric Estimation with Economic Constraints via Trimming," Working paper

Tu, Y. (2011b), "Information-based Tests of Conditional Mean Independence," Working paper

Tu, Y. and T.H. Lee (2011), "Forecasting Using Supervised Factor Model," Working paper

White H. (2000), "A Reality Check for Data Snooping," *Econometrica*, Vol. 68, 1097-1126.

White, H. and K. Chalak (2006), "A Unified Framework for Defining and Identifying Causal Effects," Working paper, UCSD, Economics Dept.

White, H., K. Chalak, and X. Lu (2011), "Linking Granger Causality and the Pearl Causal Model with Settable Systems," *Journal of Machine Learning Research Workshop and Conference Proceedings*, 12, 1-29.

White, H. and X. Lu (2010), "Robustness Checks and Robustness Tests in Applied Economics," UCSD Dept. of Economics Discussion Paper.

Wold, Herman (1966). "Estimation of principal components and related models by iterative least squares," In P.R. Krishnaiaah (Ed.). *Multivariate Analysis.* (pp. 391-420) New York: Academic Press.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101, 1418-1429.

Zou, H. and T. Hastie (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, *Ser. B.* 67 301-320.

Table 4: Squared Bias (×100): DGP 1

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	0.01	0.01	0.01	580.45	2650.07	5672.03	7.25	2.69	10.66	10.66	11.01
	θ_2	0.00	0.00	0.00	1744.79	1353.36	2344.46	0.00	0.39	3.41	3.41	2.93
n= 200	θ_1	0.00	0.00	0.00	5445.91	6859.52	7348.23	6.11	3.95	0.04	0.04	0.00
	θ_2	0.00	0.00	0.00	2910.09	7132.12	7690.01	6.04	10.69	2.22	2.22	1.69

Table 5: Mean Squared Error (×100): DGP 1

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	4.12	4.12	4.12	581.03	2653.12	5672.88	7.57	3.07	10.70	10.70	11.06
	θ_2	0.91	0.91	0.91	1744.84	1354.10	2344.68	0.09	0.51	3.42	3.42	2.94
n= 200	θ_1	0.56	0.56	0.56	5445.99	6859.71	7348.34	6.17	4.03	0.04	0.04	0.01
	θ_2	0.81	0.81	0.81	2910.28	7132.44	7690.17	6.10	10.77	2.23	2.23	1.70

Table 6: Squared Bias (×100): DGP 2

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	0.00	0.00	0.00	2.68	1.83	1.99	0.11	6.57	5.31	5.88	5.24
	θ_2	0.02	0.02	0.02	5.30	1.24	2.44	0.00	3.80	3.18	4.34	3.24
n= 200	θ_1	0.00	0.00	0.00	0.09	0.04	0.13	0.04	0.00	0.17	0.22	0.17
	θ_2	0.01	0.01	0.01	2.98	0.76	0.78	0.02	0.13	0.19	0.07	0.18

Table 7: Mean Squared Error (×100): DGP 2

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	7.64	7.64	7.64	5.11	5.83	5.25	6.75	10.01	7.75	8.57	7.68
	θ_2	7.05	7.05	7.05	7.24	4.71	4.95	4.06	6.94	5.14	6.44	5.20
n= 200	θ_1	2.84	2.84	2.84	0.56	2.05	1.40	1.61	0.73	0.63	0.69	0.63
	θ_2	3.90	3.90	3.90	3.52	3.44	2.40	1.93	0.87	0.72	0.61	0.71

Table 8: Squared Bias (×100): DGP 3-1

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	0.01	0.02	0.03	10.83	5.26	8.19	0.79	9.20	11.18	12.13	11.22
	θ_2	0.01	0.00	0.00	0.83	5.46	9.23	1.30	16.25	17.73	18.64	17.76
n= 200	θ_1	0.00	0.01	0.01	0.60	0.00	0.04	0.07	0.17	0.58	0.70	0.59
	θ_2	0.00	0.00	0.00	1.84	1.17	1.97	0.34	0.86	0.34	0.36	0.34

Table 9: Mean Squared Error (×100): DGP 3-1

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	20.01	15.26	15.49	18.33	15.17	16.66	20.09	17.81	18.38	19.36	18.43
	θ_2	24.85	22.58	23.26	9.18	16.99	18.32	18.16	25.21	24.49	25.22	24.53
n= 200	θ_1	4.43	3.79	3.80	2.45	2.76	2.20	3.10	2.22	2.43	2.59	2.43
	θ_2	5.48	4.99	5.04	3.63	4.78	4.71	4.27	3.00	2.29	2.40	2.28

Table 10: Squared Bias (×100): DGP 3-2

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	0.05	0.00	0.01	11.84	11.74	15.00	0.81	0.75	0.15	0.09	0.14
	θ_2	0.06	0.02	0.00	0.02	0.01	0.01	0.00	1.66	2.99	3.97	3.06
n= 200	θ_1	0.00	0.00	0.00	17.70	2.58	5.97	0.99	13.76	15.75	15.57	15.72
	θ_2	0.00	0.00	0.00	1.49	0.24	0.39	0.07	1.81	1.95	2.17	1.95

Table 11: Mean Squared Error (×100): DGP 3-2

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	29.75	20.48	21.57	19.99	22.91	24.33	29.58	12.74	8.88	9.12	8.86
	θ_2	19.72	17.79	18.88	13.68	13.66	13.50	18.22	17.42	15.91	16.92	15.98
n= 200	θ_1	8.22	6.85	6.91	20.19	6.60	9.03	5.94	16.60	18.22	18.06	18.20
	θ_2	7.70	6.31	6.37	4.76	5.26	4.56	5.83	5.66	5.24	5.49	5.24

Table 12: Squared Bias (×100): DGP 3-3

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	0.00	0.00	400.00	0.00	3.15	1.33	0.16	0.18	0.06	0.19	0.09
	θ_2	0.00	0.01	900.00	13.22	13.36	12.29	0.08	0.08	0.10	0.01	0.09
n= 200	θ_1	0.00	0.00	400.00	1.97	0.07	0.36	0.34	4.06	3.16	3.45	3.16
	θ_2	0.00	0.00	900.00	0.09	0.02	0.00	0.20	0.54	0.25	0.28	0.25

Table 13: Mean Squared Error (×100): DGP 3-3

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	3.32	2.47	400.00	0.10	4.62	1.84	1.79	0.56	0.17	0.35	0.20
	θ_2	1.88	1.26	900.00	13.59	13.95	12.61	1.18	0.86	0.47	0.50	0.48
n= 200	θ_1	0.38	0.26	400.00	1.99	0.40	0.61	0.57	4.19	3.18	3.47	3.18
	θ_2	0.71	0.56	900.00	0.30	0.62	0.48	0.62	0.82	0.45	0.49	0.45

Table 14: Squared Bias (×100): DGP 4

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	0.01	0.00	0.00	0.05	0.01	0.01	0.16	0.25	0.14	0.11	0.14
	θ_2	0.00	0.00	0.00	0.04	0.08	0.08	0.26	0.27	0.16	0.12	0.16
n= 200	θ_1	0.00	0.00	0.00	0.02	0.07	0.08	0.07	0.02	0.00	0.01	0.00
	θ_2	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.01	0.01	0.01

Table 15: Mean Squared Error (×100): DGP 4

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	6.79	1.88	1.88	1.82	2.27	1.97	4.88	2.43	1.89	1.94	1.88
	θ_2	8.88	2.36	2.36	2.33	3.19	2.73	6.28	3.03	2.37	2.46	2.37
n= 200	θ_1	2.30	0.53	0.53	0.56	0.62	0.61	1.15	0.60	0.53	0.54	0.53
	θ_2	2.21	0.60	0.60	0.58	0.64	0.61	1.11	0.71	0.59	0.60	0.59

Table 16: Squared Bias (×100): DGP 5-1

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	0.01	0.00	0.01	9.18	4.12	5.22	0.25	2.41	0.81	0.38	0.72
	θ_2	0.02	0.00	0.02	2.67	0.70	1.23	0.15	1.66	0.86	0.22	0.74
n= 200	θ_1	0.00	0.00	0.00	0.04	0.08	0.12	0.09	0.04	0.01	0.02	0.01
	θ_2	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.06	0.01	0.00	0.01

Table 17: Mean Squared Error (×100): DGP 5-1

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	42.30	4.91	6.55	17.77	16.16	15.94	31.91	12.26	9.72	9.24	9.57
	θ_2	23.13	6.41	8.38	16.16	13.98	14.53	21.18	15.10	13.43	12.59	13.23
n= 200	θ_1	8.22	1.24	1.28	2.53	2.66	2.67	4.48	2.73	2.48	2.49	2.48
	θ_2	7.70	1.64	1.65	3.27	3.48	3.47	5.54	3.41	3.30	3.30	3.30

Table 18: Squared Bias (×100): DGP 5-2

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	0.01	0.01	0.02	0.06	0.16	0.54	0.12	0.01	0.01	0.01	0.01
	θ_2	0.00	0.01	0.04	6.99	11.86	20.58	0.18	0.09	0.20	0.25	0.19
n= 200	θ_1	0.03	0.00	0.00	0.16	0.10	0.02	0.01	0.02	0.04	0.02	0.04
	θ_2	0.00	0.00	0.00	0.01	0.00	0.08	0.11	0.00	0.01	0.00	0.01

Table 19: Mean Squared Error (×100): DGP 5-2

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	106.66	9.44	21.20	26.45	29.91	27.52	78.98	24.91	25.06	20.46	24.38
	θ_2	140.79	14.21	21.09	67.04	68.65	88.93	143.22	40.32	50.58	43.07	50.53
n= 200	θ_1	9.97	1.54	1.80	3.85	3.88	3.71	7.40	3.88	3.76	3.59	3.75
	θ_2	16.38	3.73	4.07	7.56	7.66	7.85	11.80	8.46	7.71	7.41	7.71

Table 20: Squared Bias (×100): DGP 5-3

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	0.00	0.04	0.06	0.26	0.16	0.04	0.22	0.27	0.05	0.00	0.05
	θ_2	0.26	0.01	0.00	0.14	1.29	3.39	0.67	0.05	0.04	0.00	0.04
n= 200	θ_1	0.02	0.01	0.01	0.09	0.06	0.01	0.01	0.03	0.04	0.09	0.04
	θ_2	0.06	0.00	0.00	0.02	0.04	0.14	0.01	0.00	0.01	0.00	0.01

Table 21: Mean Squared Error (×100): DGP 5-3

	θ_0	OLS	GLS	FGLS	factor	MMA	JMA	LASSO	FOGLS	gMPL1	gMPL2	eMPL
n= 50	θ_1	71.76	19.48	25.39	22.83	27.98	25.91	66.38	29.01	24.19	25.02	24.21
	θ_2	108.90	22.82	27.67	27.68	35.08	35.65	84.66	33.65	28.85	30.36	28.88
n= 200	θ_1	17.61	3.02	3.28	4.76	5.03	4.90	10.57	5.06	4.77	4.82	4.78
	θ_2	29.56	5.57	6.01	8.49	9.05	9.12	19.83	8.53	8.41	8.43	8.41