Alternatives to GMM:
Properties of Minimum Divergence Estimators

Giuseppe Ragusa*
University of California, San Diego
gragusa@ucsd.edu

Job Market Paper

Abstract

This paper considers estimation and inference using Minimum Divergence (MD) techniques. MD estimators are obtained by minimizing a divergence between the empirical distribution and the distribution implied by moment restrictions. Methods that have received attention as possible alternatives to GMM, such as Empirical Likelihood, Exponential Tilting and Continuous Updating, are all special cases of Minimum Divergence estimators. The paper makes the following main contributions. First, it proves that there is a relationship between the Generalized Empirical Likelihood (GEL) class of estimators and the MD class that extends beyond the known cases. Second, a Bayesian interpretation of the weighting scheme through which the MD reweighs the observations is given. Third, it is shows that all the members of the MD class that have the same asymptotic bias of Empirical Likelihood are third-order efficient. This result implies that higher order efficiency is an inadequate criterion for prescribing which specific estimator should be used in applied work. I argue that in selecting from the class of third order efficient estimators, one should consider the boundedness of the influence function. Monte Carlo simulations show that test statistics based on estimators that are third order efficient and have a bounded influence function have smaller size distortion in finite sample than tests based on third order efficient estimators with unbounded influence function.

*I thank my advisors Graham Elliott, Hal White, Eli Berman and Roger Gordon for their guidance and support. I am grateful to Francesca Mazzolari, Seth Pruitt and Barbara Rossi for comments.

1 Introduction

Generalized Method of Moments (GMM) is one of the most popular techniques for estimating and conducting inference on models in economics. Nonetheless, GMM has recently been challenged by simulation evidence showing its lackluster performance in finite sample applications. The goal of this paper is to examine and extend the literature that looks for alternatives to GMM.

Over the last two decades, econometric models characterized in terms of moment restrictions have gained a prominent role in economics. The most decisive factor in favor of such models is that they avoid making strong, and often indefensible, assumptions on the probabilistic distributions of economic quantities. Instead, moment conditions arise naturally in economic models of optimizing behavior where the Euler equation from the optimal control problem implies restrictions on the correlations between economic quantities. Importantly, GMM allows a unified framework for analyzing various specific estimation methods that rely on moment conditions, as is the case for instrumental variables, dynamic panel data, and minimum distance techniques. One indication of GMM's acceptance in the profession is the choice of Hayashi (2000) to present several estimation techniques as special cases of a general GMM framework.

It is well known that GMM estimators have nice asymptotic properties (see, Gallant and White (1988) and Newey and McFadden (1994) among others). Under regularity conditions, GMM estimators are consistent, asymptotically normal and asymptotically efficient. Starting with GMM estimators, efficient test statistics can be constructed to evaluate hypotheses about parameters of interest. An important feature of GMM is that, by exploiting the overidentification of moment conditions, it allows for testing of a theoretical model or a reduced form specification.

Despite GMM's desirable asymptotic properties, there has been increasing concern over its performance in applications. In Monte Carlo simulations of model designs and sample sizes similar to those considered in real applications, evidence shows that GMM estimators are severely biased in finite samples. Given such bias, it is natural to expect that GMM's test statistics also have unsatisfying finite sample performance. In the standard asymptotic thought experiment, asymptotically efficient estimators lead to efficient tests. If, however, the small sample approximation of this thought experiment is poor, inferences based on GMM estimators lead to tests with bad size control. This intuition is in line with the findings of Monte Carlo simulations that show that GMM based tests have bad size control.

One stream of the literature has addressed the poor performance of GMM by deriving asymptotic approximations that better cope with the finite sample behavior of GMM estimators. In the context of Two Stage Least Square (TSLS) models, the last ten years have witnessed a series of papers that extend the large n thought experiments. For instance, Bekker (1994) has analyzed the properties of the asymptotic approximations when the number of instruments is let to grow with the sample size. On a related front, Staiger and Stock (1997) propose an alternative framework for analyzing asymptotic properties of TSLS estimators when the correlation between the instruments and the endogenous regressor is very small. In this case, they consider a thought experiment where the correlation coefficient of the first-stage regression is going to zero as the sample size increases. Building on these thought experiments, testing procedures that exploit weak instruments have been proposed, such as in Kleibergen (2002) and Moreira (2003), that are robust to weak instruments and/or to the presence of many instruments. A major obstacle to this approach is that the results strongly depend on specific assumptions and are usually difficult to generalize. From the same angle, Hall and Horowitz (1996) analyze the ability of the bootstrap to provide critical values for the test of overidentifying restrictions and the Wald test based on GMM. Although they show that the bootstrapped critical values have theoretical advantages, their numerical example indicates that even though the empirical size of the bootstrapped statistics are closer to their nominal values, they are still rather poor.

The aim of this paper is to extend the literature that focuses on alternatives to the traditional GMM. Motivated by the inability of the GMM to deliver estimators whose finite sample distribution adheres to the large n approximation, a new literature has emerged looking for alternative estimation techniques that possess better finite sample properties. Over the last decade, a class of alternatives has been suggested and advocated by many researchers. This class includes Empirical Likelihood (EL) (Qin and Lawless, 1994), Exponential Tilting (ET) (Kitamura and Stutzer, 1997) and Continuous Updating (CUE) (Hansen, Heaton and Yaron, 1996). There are several Monte Carlo experiments that clearly indicate that estimators obtained by these methods may have better finite sample properties than GMM. For IV estimation of a Gaussian linear equation, Judge and Mittelhammer (2001) show that EL and ET both have a smaller bias than GMM. Imbens (1997) investigates a nonlinear covariance structure model and finds that the EL has smaller bias than GMM. Imbens (2002) studies the properties of ET when applied to dynamic panel data with fixed effects, and finds that ET is superior in terms of bias and the coverage rate of confidence interval.

All these estimators exploit the same set of moment conditions that GMM uses; the key difference is how these alternatives deal with the overidentification of the model. GMM deals with the inability of exactly solving the empirical moment conditions by minimizing a weighted quadratic distance in the moment conditions. On the other hand, EL, ET, and CUE deal with the overidentification by setting the empirical moment conditions to zero through weighting the observations. There exist many weighting schemes through which the empirical moment condition can be set to zero. The idea is to pick the scheme that is closer to the empirical distribution in some meaningful sense. This meaning comes from the choice of an appropriate metric that is referred to as divergence. Differences between EL, ET, and CUE estimators arise from the different divergences these methods employ in selecting a feasible weighting scheme.

The first aim of this paper is to consider a generalization of EL, ET, and CUE estimators in which the only condition placed on their divergence is that it be a convex function. Following Corcoran (1998), who considers a similar generalization for goodness-of-fit tests, I refer to this generalization as Minimum Divergence (MD) estimators. In the context of moment conditions models, a similar generalization has only been considered for a specific family of divergences, namely the Cressie and Read (1984) family of power divergences.

Newey and Smith (2004), NS henceforth, consider a generalization that they refer to as Generalized Empirical Likelihood (GEL) estimators. They show that when one considers divergences that belong to the Cressie-Read family, GEL and MD estimators solve the same first order conditions. An interesting finding of the present paper is that there is indeed a one-to-one relationship between the MD and GEL formulation. Every MD estimator has a GEL representation and, conversely, every GEL estimator can be represented as MD estimator. This is interesting because it allows an elegant generalization of the alternatives to GMM in terms of divergence.

Another result, interesting in its own right, is that it is possible to understand the different MD-estimator-imposed weighting schemes by a probabilistic interpretation that has a clear Bayesian flavor. By imposing a prior on the data weights themselves, it can be shown that the expectation with respect to the distribution of maximum entropy to the given prior is equivalent to the MD weights when the prior is assumed to belong to a class of specific families of distributions. Although this approach can be used to include additional prior information, its core appeal is that it leads to such a theoretically established interpretation of the weighting scheme.

This interpretation may shed light on features of the MD estimators and their weighting schemes, yet the question remains as to which member of the MD class one should use in application. Comparisons between GMM and MD are particularly difficult because they all share the same asymptotic distribution with the same first order efficient variance matrix. NS compare higher order asymptotic properties of GMM and GEL. While they find that all members of GEL have lower bias than GMM, they also show that EL has the lowest bias in the GEL class. Significantly, NS show that EL is third order efficient in the sense that, after it is bias corrected, it is efficient of a higher order relative to other bias corrected estimators. A very interesting and important finding of this paper is that all MD estimators sharing the asymptotic bias of EL estimators have the same higher order mean square error. This finding has two substantive implications: first, that all the members of this MD subclass are third order efficient after the bias is removed; second, that third order efficiency is an inadequate criterion for prescribing which specific estimator should be used in applied work. If one insists on considering estimators that have the same bias as EL estimators, then another criterion must supplement third order efficiency.

This paper proposes such an additional criterion. For selecting from the class of third order efficient estimators, a researcher should consider the boundedness of the influence function of the MD estimator. There are two reasons why properties of the influence function should be considered when selecting from amongst MD estimators. First, the asymptotic expansions are polynomials in the influence functions of the estimators. If the influence function can become unbounded, then the ranking based on higher order comparisons can be misleading. Second, test statistics for overidentifying restrictions are likely to depend crucially on the boundedness of the influence function. For example, Imbens, Spady and Johnson (1998) analyze the properties of overidentifying restrictions tests and find that the Exponential Tilting, which is not third order efficient but whose influence function is bounded, delivers test statistics that are especially good in terms of size control even when compared to Empirical Likelihood, which is third order efficient but whose influence function is not bounded. Motivated by these reasons, I identify a subclass of third order efficient MD estimators whose influence function is bounded. Monte Carlo simulations show that members of this subclass deliver statistics that have strikingly good size control for testing hypotheses both on the parameters and on the overidentified structure of the model.

The plan of the paper is as follows. Section 2 defines moment conditions models and gives some examples from the economic literature. Section 3 reviews the results on the GMM estimation and testing framework. Section 4 presents the existing alternatives to GMM, while Section 5 defines the Minimum Divergence class of estimators and analyzes their properties. Section 6 establishes the Bayesian interpretation of MD techniques. Section 7 contains theoretical results on the higher order properties of MD estimators. Section 8 discusses the construction of third order efficient estimators whose influence function is bounded, while Section 9 presents Monte Carlo simulation results. Section 10 concludes.

2 The model and some examples

In this section we formally introduce models based on moment conditions and we provide some examples from the economic literature.

Let $\{w_i\}_{i=1}^n$ be a random sample in $\mathbb{R}^s$ drawn from an unknown probability distribution Q_o ; Θ is a compact subset of $\mathbb{R}^k$ and $q(w, \theta) : \mathbb{R}^s \times \Theta \to \mathbb{R}^m$, $m \geq k$, is a vector of known functions that relate the parameter vector θ to the random sample and it is referred to as the moment function. We consider the moment restrictions

$$E[q(w, \theta_o)] \equiv \int q(w, \theta_o)\mathrm{d}Q_o = 0 \qquad (1)$$

where $\theta_o \in \Theta$ is the true parameter value.

Often $\{w_i\}_{i=1}^n$ is partitioned as $\{x_i, y_i\}_{i=1}^n$ where $x_i \in R^d$ and $y_i \in R^p$, $d+p = s$. The partition is useful when we are interested in some aspects of the conditional distribution of y given x or, more generally, when y is a set of dependent variables to be determined at least partly on the basis of other variables x. For instance, the model specified in (1) is compatible with conditional restrictions of the form $E[\rho(y, \theta)|x]$ where $\rho(y, \theta) : R^p \to R^j$ is a known function. The conditional restrictions imply the unconditional moment restrictions $E[A(x)\rho(y, \theta))] = 0$, where $A(x)$ is a matrix of functions of the conditioning variables with j columns.

The econometric models given by equation (1) is extremely general and it is very common in many fields of economics.

Example 1 (Linear System of Equations): Let $y = (y_1, y_2)$ and $x = (x_1', z')'$ and consider the following system of two linear simultaneous equations:

$$\begin{aligned} y_1 &= y_2\beta + x_1'\gamma + u_1 & (2) \\ y_2 &= z'\delta + x_1'\phi + u_2 & (3) \end{aligned}$$

where (2) is the structural equation of interest, while (3) is the reduced form equation for y_2. The vectors of exogenous random variables x_1, and z have dimension $G_{x_1} \times 1$ and

$m \times 1$ respectively. One is interested in the estimation of the structural parameter β. In labor economics this approach is often taken to model the multitude of relations that characterize economic behavior in the labor market. In the classic example of Mincer's equation, (2) is the wage equation, where x_1 is a vector of personal characteristics and y_2 denotes schooling. In general $\text{Cov}(u_1, u_2) \neq 0$ and the regression of y_1 on y_2 will deliver inconsistent estimates of β. Using the exogeneity of z, β can be identified and estimated consistently by imposing (1) with $q(w, \theta) = (x_1' u_1, z' u_1)'$.

Moment conditions obtained from linear models similar to the previous example are linear in the parameters. However, nonlinearity of the moment function does not depend crucially on the linearity of the underlying model. Typically, this is the case for dynamic panel data models with fixed effects. Although the model is linear in the parameters, some of the moment conditions implied by usual assumptions on the error are nonlinear.

Example 2 (Dynamic Panel Data Models): Consider the following dynamic panel data model:

$$y_{it} = \theta y_{i,t-1} + \eta_i + u_{it}, \;\; i = 1, \ldots, n$$

where $(y_{i0}, y_{i1}, y_{i2}, \ldots, y_{iT})$ are random samples of n individual time series, η_i is an unobserved effect, T is small and n is large. Differencing to remove the fixed effect yields

$$\Delta y_{it} = \theta \Delta y_{i,t-1} + \Delta u_{it}$$

Nickell (1981) showed that fixed effect estimation in a short dynamic panel with individual specific effects (η_i) is inconsistent. However, the autoregressive parameter θ can be identified by exploiting the covariance structure of the initial observations. For $t = 3, \ldots, T$ and $s = 2, \ldots, t-1$ the model implies the following moment restrictions linear in θ

$$E\left[\left(\Delta y_{it} - \theta \Delta y_{i,t-1}\right) y_{i,t-s}\right] = 0, \;\; i = 1, \ldots, n$$

Ahn and Schmidt (1995) pointed out that the model also implies nonlinear restrictions. In particular, for any $t = 2, \ldots, T-1$, $E[u_{iT} \Delta u_{it}] = 0$ giving additional $T-2$ moment conditions that are nonlinear in θ.

Example 3 (Consumption CAPM): Consider the prototypical consumption based asset pricing model. The representative agent maximizes intertemporally time additive preferences. Let β denote the discount factor, $C_{t+\tau}$ the consumption level during period $t+\tau$, R_i the $L \times 1$ vector of gross returns on L assets, and ι_L a $L \times 1$ vector of ones. The Euler equation is given by (see Hansen and Singleton, 1982)

$$E\left(\beta(C_{t+1}/C_t)^{-\gamma} R_{t+1} - \iota_L \,|\mathcal{F}_t\right) = 0 \tag{4}$$

where $\mathcal{F}_t$ is the information set at time t. For every $Z_t \in \mathcal{F}_t$, the Euler equation implies (1) with

$$q(w,\theta) = (\beta(C_{t+1}/C_t)^{-\gamma} R_{t+1} - \iota_L) \otimes Z_t \tag{5}$$

Another application of moment conditions is in the context of optimal forecast.

Example 4 (Optimal Forecast): There exists an extensive literature of applied work in economics that tests for rational expectations using data on forecasts produced by economic agents. Properties of rational forecasts are intrinsically related to the loss function the economic agents use in producing the forecasts. Let $L(\widehat{y}_{t+h}(\mathbf{x},\theta), y_{t+h})$ be the function describing the loss associated with forecasting the value of the random variable y_{t+h} using the parametric model given by $\widehat{y}_{t+h}(\mathbf{x},\theta)$. The optimal forecast minimizes the expected loss

$$\min_\theta E[L(\hat{y}_{t+h}(\mathbf{x},\theta), y_{t+h})|\mathcal{F}_t]$$

Under regularity conditions that allow taking derivatives under the integral sign, the first order conditions take the form of the moment conditions in (1) with

$$q(w,\theta) = (\partial L(\hat{y}_{t+h}(\mathbf{x},\theta), y_{t+h})/\partial\theta) \otimes Z_t$$

for every $Z_t \in \mathcal{F}_t$. Elliott, Kumunjer, and Timmermann (2002) consider a class of loss functions indexed by a parameter that synthesizes the degree of asymmetry of the loss. Under the assumption of optimality of the forecasts, they exploit the orthogonality conditions to identify and estimate the asymmetry in the loss function.

The following notation will be kept for the remainder of the paper. For notational convenience the dependence of $q_i(\cdot,\theta)$ on w is often suppressed so that we write $q_i(\theta)$ for $q(w_i,\theta)$. When not otherwise mentioned, the index n denotes average over the sample, e.g. $q_n(\theta) = n^{-1}\sum_{i=1}^n q_i(\theta)$. The $m \times k$ Jacobian of $q_i(\theta)$ is denoted by $\nabla_\theta q_i(\theta) = \partial q_i(\theta)/\partial\theta$. The symbol $1_{(w \le x)}$ denotes the indicator function that takes the value 1 if $w \le x$ and 0 otherwise. The stochastic order symbols, $o_p(\cdot)$ and $O_p(\cdot)$, introduced by Mann and Wald (1943) are used to describe the asymptotic magnitude of statistical quantities. $\|A\|$ denotes the Euclidean norm $\sqrt{\text{Trace}(A'A)}$, that reduces to the absolute norm if A is a scalar. Given a matrix A, A^{-g} denotes the Moore-Penrose inverse of A. Finally, $\mathcal{S}(\theta_o, \epsilon)$ denotes an open sphere with center θ_o and radius $\epsilon > 0$.

3 The Generalized Method of Moments

This section serves mainly to restate the fundamental results concerning the Generalized Method of Moments estimation and establish the general estimation problem. The asymptotic properties of GMM are well known and proofs of the results given here are readily available in the literature for the *iid* case (see, Newey and McFadden, 1994), as well as for heterogeneous and dependent random variables (see, Gallant and White, 1988, and Potscher and Prucha, 1997).

Given a sample of observations $\{w_i\}_{i=1}^n$, empirical content is given to (1) by considering its sample counterpart $n^{-1}\sum_{i=1}^n q(w_i, \theta)$. When $m = k$, the model is said to be exactly identified and a consistent estimator of θ_o is the root of the m equations

$$n^{-1}\sum_{i=1}^{n} q(w_i, \theta) = 0 \tag{6}$$

When $m > k$, *i.e.* the number of equations is larger than the dimension of the parameter vector θ_o, the econometric model specified by (1) is overidentified and the existence of a unique solution to (6) is not guaranteed. The basic idea behind the GMM is to choose the parameter that sets the sample counterpart of the moment conditions close to zero, where closeness is measured as a quadratic form in a positive definite matrix. Let $\mathcal{W} = \{W_n\}$ be an $(m \times m)$ sequence of stochastic symmetric uniformly positive definite (u.p.d.) matrices. A GMM estimator associated with $\mathcal{W}$ is a sequence of solutions $\hat{\theta}$ to

$$\min_{\theta \in \Theta} Q_n(\theta, W_n) \tag{7}$$

where $Q_n(\theta, W_n) = q_n(\theta)' W_n q_n(\theta)$. The matrix W_n is referred to as distance or weighting matrix and, broadly speaking, it weighs the contribution of each average moment condition in pinning down the parameter estimate.

Under well known regularity conditions the solution of (7) is a consistent estimator of θ_o. Let W_o be the limit of the sequence $\mathcal{W}$, *i.e.* $W_n - W_o = o_p(1)$, and let $\hat{\theta}$ denote the solution to 7. The argmin $\hat{\theta}$ has a limiting normal distribution with

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, S_W)$$

where

$$S_W = \left(\Gamma_o' W_o \Gamma_o\right)^{-1} \Gamma_o' V_o^{-1} W_o V_o^{-1} \Gamma_o \left(\Gamma_o' W_o \Gamma_o\right)^{-1}$$

where $V_o = E[q_i(\theta_o) q_i(\theta_o)']$ and $\Gamma_o = E[\nabla_\theta q_i(\theta_o)]$. The minimization (7) defines a class of asymptotically consistent and normal estimators, because for any given uniformly positive definite sequence $\mathcal{W}$, the argmin $\hat{\theta}$ is consistent and asymptotically normal. The optimal GMM (OGMM) is the member of this class with the smallest asymptotic variance. Hansen (1982) shows that the OGMM is the argmin $\hat{\theta}$ associated to the sequence $\mathcal{W}$ with limit $W_o = V_o^{-1}$, in the sense that the resulting variance matrix, $S_o = (\Gamma_o' V_o^{-1} \Gamma_o)^{-1}$, satisfies $S_W - S_o \geq 0$ for every W_o. Bates and White (1993) analyze the properties of a generalization of GMM where the estimator is still defined in terms of the minimization problem (7), but various functional forms are considered for $Q_n(\theta, W_n)$. They find that the OGMM is optimal in this wider class under serial correlation and heterogeneity. The intuition for the above optimal weighting matrix is that weighting by the inverse of the variance of $\{q_i(\theta)\}_i^n$ avoids the possibility that one component with high variance can unduly dominate the minimand in (7). Chamberlain (1987) proves that the OGMM is not only the estimator with the smallest variance in the class defined by the optimization (7), but it is also optimal in the class of estimators that exploit the moment conditions given by (1). In other words, the OGMM reaches the semiparametric efficiency bound.

An obvious choice for W_n is given by

$$\hat{V}_n \equiv n^{-1} \sum_{i=1}^{n} q_i(\bar{\theta}) q_i(\bar{\theta})' \xrightarrow{p} V_o \tag{8}$$

where $\bar{\theta}$ is an initial consistent estimator of θ_o. In practice, $\bar{\theta}$ is obtained by solving the GMM minimization problem with an inefficient choice of W_n that does not depend on θ, for example the identity matrix of dimension m. Even if the asymptotic distribution of the two-step procedure that delivers the efficient GMM is equivalent to the infeasible optimal GMM estimator that uses V_o^{-1} as a weighting matrix, the finite sample properties are affected from the first step. Moreover, the two-step procedure introduces a degree of arbitrariness that is somewhat disturbing. Two researchers using the same data to estimate parameters through OGMM will obtain different estimates depending on the choice of the first-step. In addition, the two-step objective function is not invariant to reparametrization of the underlying model.

GMM provides a nice framework to conduct inference. Associated with the GMM estimators are tests for hypotheses about the parameter θ_o that can be expressed as a set of restrictions on the parameters, such as $H_o : \theta = \theta_o$ versus $H_1 : \theta \neq \theta_o$. Three classical test statistics are available: Wald, Lagrange multiplier and criterion based. The Wald test statistic is based on deviations of the unconstrained estimates from values consistent with the null hypothesis. Formally,

$$Wald(\hat{\theta}) = n \cdot (\hat{\theta} - \theta_o)\widehat{S}_n^{-1}(\hat{\theta} - \theta_o) \tag{9}$$

where $\hat{S}_n$ is any consistent estimator of S_o. The Lagrange multiplier statistic is based on deviations of the constrained estimates from values solving the unconstrained problem

$$LM(\hat{\theta}) = n \cdot q_n' \widehat{V}_n^{-1} \widehat{\Gamma}_n \widehat{S}_n \widehat{\Gamma}_n' \widehat{V}_n^{-1} q_n \tag{10}$$

where $\widehat{\Gamma}_n$ is any consistent estimator of Γ_o. Criterion based statistics are based on the distance of the GMM objective function when evaluated at the constrained θ_o and unconstrained $\hat{\theta}$ estimates

$$LR(\hat{\theta}) = -2n \cdot (Q_n(\theta_o) - Q_n(\hat{\theta})) \tag{11}$$

Under H_o, the three statistics have a χ^2_k asymptotic distribution. Valid hypothesis tests and confidence sets can be constructed in a straightforward way. To perform an asymptotically valid test of the hypothesis $\theta = \theta_o$, reject if the values of the test statistics exceeds the appropriate χ^2_{k-j} critical value. To construct an asymptotically valid confidence set, invert the test based on $Wald$, LM and LR. That is, for example, $\{\theta : W(\theta) \leq \chi^2_{k,l}\}$, $\{\theta : LM(\theta) \leq \chi^2_{k,l}\}$ and $\{\theta : LR(\theta) \leq \chi^2_{k,l}\}$ are asymptotic $100(1-l)\%$ confidence sets, where $\chi^2_{k,l}$ is the $100(1-l)\%$ critical value of a χ^2_k distribution.

Hansen (1982) also suggested a specification test that relies on the objective function being minimized. The test of overidentified restrictions tests whether the moment restrictions defining the model are satisfied. The test statistic is given by

$$J(\hat{\theta}) = n \cdot Q(\hat{\theta}, \hat{V}_n^{-1})$$

Under the null hypothesis that all the moments have expectation equal to zero at the true value of the parameter, the distribution converges to a χ^2 with degrees of freedom equal to the number of overidentified restrictions, $m-k$. To perform an asymptotically valid test of the hypothesis $E[q(w_i, \theta_o)] = 0$, reject if the values of the test statistics exceeds the appropriate χ^2_{m-k} critical value.

4 Existing Alternatives to GMM

In Monte Carlo simulations of model designs and sample sizes similar to those considered in real applications, evidence shows that GMM estimators are severely biased in finite samples (see Altonji and Segal (1996) among others). Given such bias, it is natural to expect that GMM's test statistics also have unsatisfying finite sample performance. This intuition is in line with the findings of Monte Carlo simulations that show that GMM based tests have empirical size that is hardly close to its nominal value. These findings have triggered research on refinement methods and alternative estimation techniques that may have better finite sample properties than GMM.

Since the seminal paper of Qin and Lawless (1994), the Empirical Likelihood (EL) estimator has received much attention as an alternative estimator for moment-condition-specified models. The EL estimator can be defined as the solution of a problem in which the empirical moments are set to zero by weighting the observations, that is

$$\hat{\theta} = \left\{ \arg\max_{\pi,\theta} \frac{1}{n} \sum_{i=1}^{n} \log \pi_i \mid s.t. \sum_{i=1}^{n} \pi_i q_i(\theta) = 0, \;\; \sum_{i=1}^{n} \pi_i = 1, \;\; \pi_i > 0 \right\} \tag{12}$$

This maximization can be interpreted as a constrained Maximum Likelihood (ML) procedure applied to joint estimation of θ and the parameters $\pi_1, \ldots, \pi_n$ of a multinomial distribution for n different types of data outcomes. As a ML estimator, EL inherits the first order asymptotic properties of ML, particularly asymptotic efficiency. Qin and Lawless (1994) showed that these properties are preserved when the underlying distribution of w is continuous.

The logarithm in the objective function does not play a fundamental role in obtaining efficient estimators. Kitamura and Stutzer (1997) suggested the Exponential Tilting (ET) estimator; it is defined similarly to the EL, save that the objective function is replaced by $\pi_i \log \pi_i$ giving

$$\hat{\theta} = \left\{ \arg\min_{\pi,\theta} \sum_{i=1}^{n} \pi_i \log \pi_i \mid s.t. \sum_{i=1}^{n} \pi_i q_i(\theta) = 0, \;\; \sum_{i=1}^{n} \pi_i = 1, \;\; \pi_i > 0 \right\} \tag{13}$$

An important feature of ET is usually singled out: $\sum_{i=1}^{n} \pi_i \log \pi_i$ is proportional to the Kullback-Leibler Information Criterion (KLIC) and (13) can be seen as minimizing the KLIC between the empirical distribution and the distribution implied by the constraints on $\{q(w_i, \theta)\}$.

A third estimator that has been considered as an alternative to GMM is the Continuous Updating Estimator (CUE). The CUE is obtained as

$$\hat{\theta} = \left\{ \arg\min_{\pi,\theta} n \sum_{i=1}^{n} \pi_i^2 \mid s.t. \sum_{i=1}^{n} \pi_i q_i(\theta) = 0, \; \sum_{i=1}^{n} \pi_i = 1 \right\} \tag{14}$$

Notice that here the positivity constraint on π_i is dropped because the objective function is defined on the whole real line. Strictly speaking, the CUE was first proposed by Hansen, Heaton, and Yaron (1996) who considered obtaining the GMM estimator without using a first step estimator of the variance matrix; that is, he considered the estimator that minimizes

$$\widetilde{Q}_n(\theta) = q_n(\theta)' \left[\frac{1}{n} \sum_{i=1}^{n} q_i(\theta) q_i(\theta)' \right]^{-g} q_n(\theta)$$

NS show that $\arg\min_{\theta\in\Theta} \widetilde{Q}_n(\theta)$ is numerically equivalent to the estimator obtained by solving (14).

The common feature of these estimation approaches is that they try to set the empirical moment conditions equal to zero by weighting the observations. EL, ET and CUE differ on the way the weighting scheme is found. In particular, while EL and ET are defined for $\pi_i > 0$, CUE is defined for negative values of the weights, allowing solutions that lie outside the convex hull of the data.

NS consider a generalization of EL, ET and CUE relying on a dual problem. They consider the following problem

$$\max_{\theta\in\Theta} \min_{\lambda\in\Lambda_n(\theta)} \frac{1}{n} \sum_{i=1}^{n} \psi(\lambda' q_i(\theta))$$

where $\psi(\cdot)$ is a convex function[1] defined on an interval $\mathcal{V}$ that contains zero and $\Lambda_n(\theta) =$

[1] In their specification they consider a problem definied as

$$\min_{\theta\in\Theta} \max_{\lambda\in\Lambda_n(\theta)} \frac{1}{n} \sum_{i=1}^{n} \rho(\lambda' q_i(\theta))$$

for a concave function $\rho(\cdot)$ defined on $\mathcal{V}$. Clearly the two problems coincide for $\psi(x) = -\rho(x)$. The formulation in terms of a convex function is kept here to stress the role of the convexity.

$\{\lambda|\lambda' q_i(\theta) \in \mathcal{V}, i = 1, \dots n\}$. NS show that the estimator of this problem, which they call Generalized Empirical Likelihood (GEL), is equivalent to EL for $\psi(x) = -\log(1-x)$, to ET for $\psi(x) = \exp(x) - 1$, and to CUE when $\psi(x) = x^2/2 - x$.

5 Minimum Divergence Estimators

The generalization of EL, ET and CUE estimators this paper considers is the class of Minimum Divergence (MD) estimators. The idea is to generalize the objective functions of EL, ET and CUE by considering the following problem

$$\hat{\theta} = \left\{ \arg\min_{\pi,\theta} \frac{1}{n} \sum_{i=1}^{n} \gamma(n\pi_i) \,|\, s.t. \sum_{i=1}^{n} \pi_i q(w_i, \theta) = 0,\ \sum_{i=1}^{n} \pi_i = 1,\, \pi_i \in (a_\gamma, b_\gamma) \right\} \qquad \text{(MD)}$$

where $\gamma(\cdot)$ is a divergence, weighting the distance between the π's and n^{-1}. Let $\gamma_r(\cdot)$ denotes the rth derivative of $\gamma(\cdot)$ and γ_r denotes the rth derivatives evaluates at 1, $\gamma_r \equiv \gamma_r(1)$. Throughout the paper $\gamma(\cdot)$ will denote a function that satisfies the following requirements:

Assumption 1 (γ): (i) $\gamma(\cdot)$ is a strictly convex function $\gamma : (a_\gamma, b_\gamma) \to [-\infty, +\infty]$, such that $a_\gamma < 1 < b_\gamma$; (ii) $\gamma(\cdot)$ is twice continuously differentiable on (a_γ, b_γ); (iii) the minimum of $\gamma(x)$ is 0, attained at $x = 1$; (iv) $\gamma_2 = 1$.

In many cases of interests the endpoints of the domain of $\gamma(\cdot)$ are given by a_γ= 0 and $b_\gamma = +\infty$, but in general the only requirement is that $a_\gamma < 1 < b_\gamma$. The assumption of strictly convexity of $\gamma(\cdot)$ over its domain could be relaxed at expenses of further complexity. Notice, however, that strict convexity is sufficient to guarantee that the problem as a unique solution $\hat{\pi}_i$. Suppose $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_n)$ and $\tilde{\pi} = (\tilde{\pi}_1, \tilde{\pi}_2, \dots, \tilde{\pi}_n)$ are both solution to MD. Then, for any $0 \leq \zeta \leq 1$, $\pi^\zeta = \zeta\hat{\pi} + (1-\zeta)\tilde{\pi}$ is a feasible solution. But if $\gamma(\cdot)$ is strictly convex, $\sum_i^n \gamma(n\, \pi_i^\zeta) < \zeta \sum_i^n \gamma(n\, \hat{\pi}_i) + (1-\zeta) \sum_i^n \gamma(n\, \tilde{\pi}_i)$, that is a contradiction, since by $\hat{\pi}$ and $\tilde{\pi}$ are both solutions. The condition on the the second derivative of $\gamma(\cdot)$ is imposed for convenience and it is not restrictive.

By setting $\gamma(x) = -\log x + x - 1$, $\gamma(x) = x \log x - x + 1$ and $\gamma(x) = x^2/2 - x$, one obtains MD problems that are equivalent to the EL, the ET and CUE respectively.

Considering this MD problem is interesting for two reasons. First, it clarifies the role played by assuming different objective functions $\gamma(x)$ in (MD) in pinning down optimal weights $\pi = (\pi_1, \dots, \pi_n)$ and the optimal θ . Second, it allows the problem to be linked to the underlying probabilistic model implied by the set of moment conditions considered.

The function $\sum_{i=1}^{n}\gamma(n\pi_i)/n$ is minimized over all probability allocations when all π_i equal n^{-1}. MD methods select, from all the π that are feasible, the weights $\widehat{\pi} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_n)$ that are closer to a weighting scheme that assigns n^{-1} to each observation in the sample. The location of the estimated parameter is implicitly identified by the shape of the divergence $\gamma(x)$. Intuitively, since under the model conditions $\pi_i \approx n^{-1}$ as $n \to \infty$ and $\gamma(1) = 0$, the shape of the divergence does not determine the asymptotic behavior of the estimator, but it does determine the finite sample location of the estimator of θ_o.

Probabilistic content to the MD methods is given by considering the collection of probability measures (p.m.) on the random variables w_i that satisfies the constraint on the moments for a given $\theta \in \Theta$. In the population, the problem can be reduced to that of selecting a p.m. that is as close as possible to Q_o in some meaningful sense. Formally, the stochastic model for the random vector $w = (w_1, w_2, ..., w_n)$ is defined as

$$\mathcal{G} = \bigcup_{\theta\in\Theta} \mathcal{G}(\theta)$$

$$\mathcal{G}(\theta) = \left\{ G : \int q(w,\theta)\mathrm{d}G = 0 \right\} \tag{15}$$

For a given γ define the following functional

$$I_\gamma(R, G) = \begin{cases} \int \gamma\left(\frac{\mathrm{d}Q}{\mathrm{d}G}\right)\mathrm{d}Q & if\ G \ll Q \\ +\infty & otherwise \end{cases} \tag{16}$$

The functional $I_\gamma(R, G)$, that is broadly speaking the population counterpart of (MD), can be interpreted as specifying a divergence function between two probability measures, R and G, and can be thought as generalizing the Kullback-Leibler Information Criterion (KLIC), that is obtained by setting $\gamma(x) = x \log x$. The population counterpart of the estimation problem defined by moment conditions can be cast in terms of finding some $G \in \mathcal{G}$ that minimizes the functional $I_\gamma(G, Q_o)$, formally $\inf_{G\in\mathcal{G}} I_\gamma(G, Q_o)$. If the model is correctly specified (i.e. $Q_o \in \mathcal{G}$) then clearly $Q_o = \inf_{G\in\mathcal{G}} I_\gamma(G, Q_o)$. Similarly, if $E_{Q_o}[q(w,\theta)] \neq 0$ for $\theta \neq \theta_o$, Q_o implicitly identifies θ_o. Note that under fairly weak conditions, when the model conditions are misspecified (i.e. $Q_o \notin \mathcal{G}$), the solutions to $Q_* = \inf_{G\in\mathcal{G}} I_\gamma(G, Q_o)$ can be interpreted as the pseudo-true probability measure, in the sense that it is the probability measure that satisfies the moment conditions and is the closest to the true distribution.

The discussion above makes clear the importance of studying the minimum divergence estimation in (MD): it is the sample counterpart of a population problem that solves for the probability that is closest to the true distribution of the data. In this counterpart, the constraint $\int q(w,\theta)dG = 0$ is substituted for by $\sum_i^n \pi_i q(w,\theta) = 0$, and the true distribution Q_o is substituted for by the empirical distribution function that assumes no ties.

This important feature of Minimum Divergence estimators allows us, in principle, to consider estimation and inference in misspecified models. However, the MD formulation is useful as long as a solution can be found by standard numerical methods. The next section takes up the issue of deriving first order conditions for the general MD problem and studies the relationship between the GEL class of NS and the MD class.

5.1 First Order Conditions and Duality

This section discusses the conditions under which the solution of a MD problem can be obtained by standard Lagrangian methods. Lagrangian methods are known to solve the ET, EL and CUE problems, but a general treatment has not yet been given in the literature. The Lagrangian of (MD) can be written as

$$\mathcal{L}(\theta,\pi,\lambda,\eta) = \frac{1}{n}\sum_{i=1}^{n}\gamma(n\pi_i) - \lambda'\sum_{i=1}^{n}\pi_i q_i(\theta) - \eta(\sum_{i=1}^{n}\pi_i - 1)$$

where $\lambda \in \mathbb{R}^m$ and $\eta \in \mathbb{R}$ are the Lagrange multipliers associated with the constraints. To further investigate the properties of the Lagrangian solution to the MD problem, some additional notation is required. Setting to zero the partial derivative of $\mathcal{L}(\theta,\pi,\lambda,\eta)$ with respect to π_i gives, for $i = 1,\dots,n$, the following

$$\gamma_1(n\pi_i) - \lambda' q_i(\theta) - \eta = 0 \tag{17}$$

Similarly, setting to zero the derivative of $\mathcal{L}(\theta,\pi,\lambda,\eta)$ with respect to θ and assuming that $q(\cdot,\theta)$ is differentiable on Θ, yields

$$\sum_{i=1}^{n}\pi_i \nabla_\theta q_i(\theta)'\lambda = 0 \tag{18}$$

The Lagrange multiplier η in (17) can be eliminated as follow. Multiplying (17) by π_i and summing over n gives

$$\sum_{i=1}^{n} \pi_i \gamma_1(n\pi_i) - \lambda' \sum_{i=1}^{n} \pi_i q_i(\theta) - \eta = 0$$

Using the constraint $\sum_i \pi_i q_i(\theta) = 0$, a solution must satisfy $\eta = \sum_{i=1}^{n} \pi_i \gamma_1(n\pi_i)$. Substituting this expression for η into (17) yields

$$\gamma_1(n\pi_i) - \sum_{i=1}^{n} \pi_i \gamma_1(n\pi_i) - \lambda' q_i(\theta) = 0$$

For any $c \in \mathbb{R}$, a solution to the previous expression is given by

$$\gamma_1(n\pi_i) = c + \lambda' q_i(\theta)$$

since $c + \lambda' q_i(\theta) - \sum_{i=1}^{n} \pi_i \left(c + \lambda' q_i(\theta)\right) - \lambda' q_i(\theta) = 0$ for any c. It is very convenient to set $c = 0$. Such a normalization allows us to consider the various estimators delivered by different choices of the divergence from a unified point of view. By the assumption of strict convexity and by two times continuously differentiability of $\gamma(x)$ on (a_γ, b_γ), it follows that $\gamma_1(\cdot)$ is continuously differentiable and by strict convexity, $\gamma_2(x) > 0$ for any $x \in (a_\gamma, b_\gamma)$. It follows that $\gamma_1(\cdot)$ is monotone on (a_γ, b_γ). Let $\mathcal{A} = \{y : y = \gamma_1(x),\ x \in (a_\gamma, b_\gamma)\}$. The optimal π can be found by inverting the function $\gamma_1(\cdot)$ whenever there exists a $\lambda' \in \mathbb{R}^m$ such that $\lambda' q_i(\theta) \in \mathcal{A}$, $i = 1, \ldots, n$, since in this case by the inverse function theorem

$$\pi_i = \frac{1}{n} \widetilde{\gamma}_1(\lambda' q_i(\theta)) \tag{19}$$

where here $\widetilde{\gamma}_1(\cdot)$ denotes the inverse function of $\gamma_1(\cdot)$. By substituting (19) into the constraint $\displaystyle\sum_{i=1}^{n} \pi_i q_i(\theta) = 0$ and into (18), the following first order conditions are obtained

$$\frac{1}{n} \sum_{i=1}^{n} \widetilde{\gamma}_1(\lambda' q_i(\theta)) q_i(\theta) = 0 \tag{20}$$

$$\frac{1}{n} \sum_{i=1}^{n} \widetilde{\gamma}_1(\lambda' q_i(\theta)) \nabla_\theta q_i(\theta)' \lambda = 0 \tag{21}$$

Remark 1: The shape of the set $\mathcal{A} = \{y : y = \gamma_1(x),\ x \in (a_\gamma, b_\gamma)\}$ determines the conditions under which the optimal solution (MD) is attained by Lagrange method. If, for a given sample, there not exist a $\theta \in \Theta$ and a $\lambda \in \mathbb{R}^m$ such that $\lambda' q_i(\theta) \in \mathcal{A}$ for $i = 1, ..., n$, the solution is not attained even if there exists a feasible solution $\hat{\theta}$ and $\hat{\pi}(\hat{\theta})$. When $\mathcal{A} = \{y : -\infty < y < +\infty\}$ the solution will be always attained (provided it exists). The form of $\mathcal{A}$ has also statistical implication, as discussed in Section 8.

Remark 2: The normalization $c = 0$ implies that when $\lambda' q_i(\theta) = 0$, $\pi_i = \frac{1}{n}\widetilde{\gamma}_1(0) = 1/n$.

Remark 3: A side effect of the elimination of the Lagrange multiplier associated with the constraint $\sum_{i=1}^n \pi_i = 1$ is that at the solution the optimal vector $\hat{\pi}$ does not satisfy the constraint, so that in general, $\sum_i^n \hat{\pi}_i \neq 1$. The optimal weight $\hat{\pi}$ satisfies the constraint if $\sum_{i=1}^{n} \hat{\pi}_i \gamma_1(n\hat{\pi}_i) = 0$. Since this is not generally the case, one needs to consider the normalized weights

$$\omega_i = \frac{\widetilde{\gamma}_1(\lambda' q_i(\theta))}{\sum_{i=1}^n \widetilde{\gamma}_1(\lambda' q_i(\theta))}$$

Clearly if $\pi_i = \frac{1}{n}\widetilde{\gamma}_1(\lambda' q_i(\theta))$ satisfies the first order conditions (20) and (21), the normalized weights $\{\omega_1, \omega_2, \ldots, \omega_n\}$ still solve (20) and (21) and, by construction, $\sum_{i=1}^{n} \omega_i = 1$.

Remark 4: The Lagrangian multiplier needs not to be eliminated. One can consider explicitly η. In that case the solution is attained by Lagrange methods if there exists $(\eta, \lambda')' \in \mathbb{R}^{m+1}$ such that $\eta + \lambda' q_i(\theta) \in \mathcal{A}$ for every $i = 1, \ldots, n$, and such that solves the first order conditions

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \widetilde{\gamma}_1(\eta + \lambda' q_i(\theta)) q_i(\theta) &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} \widetilde{\gamma}_1(\eta + \lambda' q_i(\theta)) \nabla_\theta q_i(\theta)' \lambda &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} \widetilde{\gamma}_1(\eta + \lambda' q_i(\theta)) &= 1
\end{aligned}$$

and in this case no normalization of the weights is required.

The first order conditions (20) and (21) reduce to the well known first order conditions for EL, ET and CUE.

Case 1 (Empirical Likelihood): *For the EL,* $\gamma_1(x) = -1/x + 1$*. The inverse of* $\gamma(\cdot)$ *is given by* $\widetilde{\gamma}_1(y) = 1/1 - y$ *and* $\mathcal{A} = \{y : -\infty < y < 1\}$*, it follows that, if there exists a*

$\lambda \in \mathbb{R}^m$ *such that* $\max_{i \leq n} \lambda' q_i(\theta) < 1$, *the optimal weights are given by*

$$\pi_i = (1 - \lambda' q_i(\theta))^{-1}/n$$

Notice that for EL $\sum_i^n \pi_i \gamma_1(n\pi_i) = \sum_i^n (1 - \lambda' q_i(\theta))^{-1}(-1 + \lambda' q_i(\theta))/n = 0$, *and hence the normalization of the weights is not necessary, since by construction* $\sum_i^n \pi_i = 1$.

Case 2 (Exponential Tilting): *The Exponential Tilting is obtained by setting* $\gamma(x) = x \log x - x + 1$ *and thus* $\gamma_1(x) = \log x$, $\mathcal{A} = \{y : -\infty < y < +\infty\}$, *and the optimal weights are given by*

$$\pi_i = \exp(\lambda' q_i(\theta))/n$$

The normalized weights given by

$$\omega_i = \frac{\exp(\lambda' q_i(\theta))}{\sum_{i=1}^n \exp(\lambda' q_i(\theta))}$$

satisfy the constraint $\sum_i^n \pi_i = 1$.

Case 3 (Continuous Updating): *For CUE,* $\gamma_1(x) = x - 1$, $\widetilde{\gamma}_1(y) = 1 + y$ *and* $\mathcal{A} = \{y : -\infty < y < 0\}$. *The optimal weights are given by*

$$\pi_i = (1 + \lambda' q_i(\theta))/n$$

and in this case too a normalization is required to satisfy the constraint.

A class of divergences that has received attention is the Cressie and Read (1984) (CR) power-divergence class given by

$$\gamma^{CR}(x) = \frac{x^{\alpha+1} - 1}{\alpha(1+\alpha)} - \frac{1}{a}x + \frac{1}{a};\ -\infty < \alpha < +\infty$$

The expression above is undefined for $a = -1$ and $\alpha = 0$, and in these cases the continuous limits

$$\lim_{\alpha \to -1} \gamma^{CR}(x) = -\log x + x - 1;\quad \lim_{\alpha \to 0} \gamma^{CR}(x) = x \log x - x + 1$$

are used. The limits above correspond to the divergences that define EL and ET, respectively. The divergence that defines the CUE is recovered by setting $\alpha = 1$. It should be pointed out that not all the members of the Read Cressie class of divergences are strictly convex. For $a \neq 0$, the first order conditions are given by

$$\frac{1}{n}\sum_{i=1}^{n}(1+\alpha\lambda' q_i(\theta))^{1/\alpha} q_i(\theta) = 0$$
$$\frac{1}{n}\sum_{i=1}^{n}(1+\alpha\lambda' q_i(\theta))^{1/\alpha} \nabla_\theta q_i(\theta) = 0$$

Despite the elegance of the derivation, the analysis based on the first order conditions have some undesirable features. In order to derive the first order conditions, an explicit formula for the inverse of the first derivative of $\gamma(\cdot)$ must exist. This, of course, is not the case generally. For example, consider the following divergence

$$\gamma(x) = \begin{cases} \frac{\overline{\alpha}(1-\alpha)(x-1)^2}{2[\alpha(x-1)+\overline{\alpha}](1-\alpha)} & x \neq 1 \\ 0 & x = 1 \end{cases}$$

for $\alpha \in [0,1]$ and $\overline{\alpha} = 1-\alpha$. This can be thought of as a generalization of the divergence that delivers the CUE, obtained by setting $\alpha = 0$. The derivatives of this divergence is given by

$$\gamma_1(x) = \frac{2\overline{\alpha}(x-1)}{\overline{\alpha}+\alpha(x-1)} - \frac{\alpha\overline{\alpha}(x-1)^2}{(\overline{\alpha}+\alpha(x-1))^2}$$

and clearly the inverse function is not explicitly available. Even when the inverse function is available, working with first order conditions has two considerable disadvantages. From a computational point of view, as pointed out by Imbens (2002) in the context of ET, calculating θ by solving the first order conditions by standard numerical methods can be problematic. From a statistical standpoint, investigating the asymptotic properties of $\widehat{\theta}$ by using standard estimating equation techniques leads to imposing conditions that are stronger than the ones needed to obtain consistency of the GMM estimator.

In this sense, the GEL representation of NS possesses both a computational and technical advantage. On the other hand, NS show that the GEL problem is equivalent to the Minimum Divergence framework only in a special case, i.e. when the divergence belongs to the Cressie-Read class. In an early version of their paper, NS conjecture that the Cressie-Read family may be the only family of divergencies admitting a GEL representation, undermining the usefulness of considering divergences outside the Cressie-Read class.

Fortunately, it turns out that it is possible to obtain a GEL representation of MD estimators that use divergences that do not belong to the Cressie-Read class. The following theorem establishes the equivalence between the MD problem and the GEL problem.

Theorem 5.1: *If the MD problem given in* (MD) *has an interior solution,* $q(w,\theta)$ *is differentiable in* $\theta \in \Theta$ *and* $\sum_i^n \widetilde{\gamma}_2(\lambda' q_i(\hat{\theta}))q_i(\hat{\theta})q_i(\hat{\theta})'$ *is non singular, then the first order conditions for MD coincide with the first order conditions of the following problem*

$$\hat{\theta}_{gel} = \max_{\theta\in\Theta} \min_{\lambda\in\Lambda_n(\theta)} \frac{1}{n}\sum_{i=1}^{n} \psi(\lambda' q_i(\theta))$$

where $\psi(\cdot)$ *is a strictly convex function defined on* $\mathcal{A}$ *given by*

$$\psi(x) = x\widetilde{\gamma}_1(x) - \gamma(\widetilde{\gamma}_1(x))$$

with $\psi_1(0) = \psi_2(0) = 1$.

As for all the results in the paper, the proof of Theorem 5.1 is given in the Appendix. Theorem 5.1 shows that given a strictly convex and twice continuously differentiable function $\gamma(\cdot)$, the MD problem that uses $\gamma(\cdot)$ delivers the same first order conditions as the GEL with an accurately chosen strictly convex function. A stronger version of Theorem 5.1 can be proved that does not rely on the first order conditions and simply establishes that any solution to the MD problem solves the GEL, avoiding the assumption of $q_i(\theta)$ being differentiable in θ. Since in this general case, the proof is more involved and since throughout this paper a differentiable moment function is assumed, the more general result is omitted and the reader is directed to Ragusa (2004) who explicitly considers duality with a nondifferentiable moment function.

Example 5 (Exponential Divergence): Consider the exponential divergence

$$\gamma(x) = e^x - ex$$

The domain of γ is $(-\infty, +\infty)$ and $\gamma_1(x) = e^x - e$. The inverse function of $\gamma_1(x)$ is given by $\widetilde{\gamma}_1(y) = \log(e+y)$, and $\mathcal{A} = \{y : -e < y < +\infty\}$. By using Theorem 5.1, the GEL problem that delivers first order conditions that are equivalent to MD is given by

$$\psi(y) = (e+y)\log(e+y) - ey - 1$$

and $\psi(x)$ is defined on $\mathcal{A}$.

Remark 5: A similar result holds when the Lagrange multiplier η is not substituted for γ_1 but it is instead is explicitly considered by slightly changing the assumption of Theorem 5.1. In particular, if $\sum_{i=1}^{n} \widetilde{\gamma}_2(\hat{\eta} + \hat{\lambda}' q_i(\hat{\theta}))q_i(\hat{\theta})q_i(\hat{\theta})'$ in non singular, the MD first order condition coincide with the first order conditions of the problem

$$\max_{\theta\in\Theta} \min_{\eta,\lambda\in\tilde{\Lambda}_n(\theta)} \frac{1}{n}\sum_{i=1}^{n}\left[\psi(\eta + \lambda' q_i(\theta)) - \eta\right]$$

where $\tilde{\Lambda}_n(\theta) = \{\eta, \lambda : \eta + \lambda' q_i(\theta) \in \mathcal{A}, i = 1, \dots n\}$, $\psi(x)$ is a strictly convex function defined on $\Lambda_n(\theta)$ and given by

$$\psi(\eta + \lambda' q_i(\theta)) = \left(\eta + \lambda' q_i(\theta)\right)\widetilde{\gamma}_1(\eta + \lambda' q_i(\theta)) - \gamma(\widetilde{\gamma}_1(\eta + \lambda' q_i(\theta)))$$

However, in many circumstances an expression for the inverse function of the derivative cannot be given explicitly, and the form function $\psi(\cdot)$ of Theorem 5.1 is unavailable. Although Theorem 5.1 allows to see MD as GEL, it does not say if given a strictly convex function $\psi(x)$ the GEL that uses $\psi(x)$ as objective function corresponds to a MD for a given divergence $\gamma(\cdot)$. Suppose $\psi(\cdot)$ satisfies the following assumptions:

Assumption 2 (ψ): (i) $\psi(x)$ is a strictly convex function $\psi : (a_\psi\ , b_\psi) \to [0, +\infty)$, $a_\psi < 0 < b_\psi$; (ii) $\psi(x)$ is twice continuously differentiable on (a_ψ, b_ψ); (iii) $\psi(0) = 0$.

Let $\mathcal{V} = \{y : y = \psi_1(x),\ x \in (a_\psi, b_\psi)\}$. By the assumption of strict convexity and by two times continuously differentiability of $\psi(x)$ on (a_γ, b_γ), it follows that $\psi_1(x)$ is continuously differentiable and by strict convexity, $\psi_2(x) > 0$ for any $x \in (a_\psi, b_\psi)$, and $\psi_1(x)$ is monotone on (a_ψ, b_ψ). By the inverse function theorem, $\widetilde{\psi}_1(y)$, the inverse function of $\psi_1(x)$, is well defined on $\mathcal{V}$.

Theorem 5.2: *Consider the GEL problem given by*

$$\max_{\theta\in\Theta} \min_{\lambda\in\Lambda_n(\theta)} \frac{1}{n}\sum_{i=1}^{n}\psi(\lambda' q_i(\theta)) \qquad \text{(GEL)}$$

where the function $\psi(\cdot)$ satisfies Assumption 2(ψ) and $\Lambda_n(\theta) = \{\lambda : \lambda' q_i(\theta) \in (a_\psi, b_\psi), i = 1, ..., n\}$.Then, if GEL has interior solution for θ and λ, and $q_i(\theta)$ is differentiable in

$\theta \in \Theta$, *there exists a strictly convex function* $\gamma(\cdot)$ *satisfying Assumption 1*(γ) *such that the first order conditions of the MD problem are equivalent to those of the GEL.*

The preceding result can be interpreted as the converse of Theorem 5.1 and it says that MD estimators can be built from the "bottom-up" by specifying $\psi(x)$ and then using this result to justify the problem in terms of divergence minimization. The duality extends also to the value of the objectives functions $\gamma(\cdot)$ and $\psi(\cdot)$, as the following result shows.

Theorem 5.3: Corollary 5.1: *Under Assumption 1*(γ),

$$\frac{1}{n}\sum_{i=1}^{n}\gamma(n\hat{\pi}_i) = -\frac{1}{n}\sum_{i=1}^{n}\psi(\hat{\lambda}'q_i(\hat{\theta}))$$

where $\hat{\pi}_i = \tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))/n$.

As shown in the next section, this result allows to characterize the asymptotic properties of estimators and test statistics in terms of the (MD) problem or equivalently in terms of the corresponding (GEL) problem.

5.2 First Order Asymptotic Properties

In this section we discuss the first order asymptotic properties of MD estimators. The following assumptions are needed to establish consistency of MD estimators.

Assumption A: *(i)* Θ is compact; *(ii)* θ_o is the only solution to $Eq(w_i,\theta) = 0$; *(iii)* $q(\cdot,\theta)$ is continuous for each $\theta \in \Theta$ with probability one; *(iv)* $E\left[\sup_{\theta\in\Theta}\|q(w_i\theta)\|^2\right] < \infty$; *(v)* $V_o \equiv E\left[q_i(\theta_o)q_i(\theta_o)'\right]$ is non singular.

Consistency of MD estimators can be proved under the same set of assumptions under which consistency of GMM is generally derived. Other works have assumed a slight stronger condition on the moment of $q(w,\theta_o)$ than the usual condition on the second moment. In particular, NS assume that $E(\sup_{\theta\in\Theta}\|q_i(\theta)\|^{\alpha}) < \infty$ for $\alpha > 2$. The results of Theorem 5.4 are derived under the assumption $\alpha = 2$.

Theorem 5.4: *Let Assumption A hold. Then (i) the solutions of the constrained optimization problem* (MD), $(\hat{\pi},\hat{\theta})$, *exist with probability approaching to one and (ii)* $\hat{\theta} \xrightarrow{p} \theta_o$; *(iii)* $\hat{\lambda} = O_p(n^{-1/2})$; *(iv)* $\max_{i\leq n}|\hat{\lambda}'q_i(\hat{\theta})| = o_p(1)$.

The following assumption is sufficient to show that $\hat{\theta}$ and $\hat{\lambda}$ are asymptotically normal.

Assumption B: *(i)* θ_o lies in the interior of Θ; *(ii)* $q(\cdot,\theta)$ is continuously differentiable on $\mathcal{S}(\theta_o,\epsilon)$, $\epsilon>0$; *(iii)* $E\left[\sup_{\theta\in\Theta}\|\nabla_\theta q(w_i,\theta)\|\right]<\infty$; *(iv)* $\Gamma_o\equiv E\left[\nabla_\theta q(w_i,\theta_o)\right]$ has full column rank.

Theorem 5.5: *Let Assumptions A-B hold. Then the sequence of solutions $\hat\theta$ and the vector of Lagrange multiplier $\hat\lambda$ are asymptotically normal and independent with*

$$\sqrt{n}\begin{pmatrix}\hat\theta-\theta_o\\ \hat\lambda\end{pmatrix}\xrightarrow{d}\mathrm{N}\left(0,\begin{pmatrix}S_o & 0\\ 0 & P_o\end{pmatrix}\right)$$

where $S_o=(\Gamma_o'V_o^{-1}\Gamma_o)$ and $P_o=V_o^{-1}(I_m-\Gamma_oS_o\Gamma_o'V_o^{-1})$.

Theorem 5.5 makes clear that MD and OGMM estimators are first order equivalent. That is, they are both asymptotically normal and they share the same asymptotic variance S_o. A notable difference between the class of GMM estimators and the class of MD estimators is the following. The GMM class is indexed by $\mathcal{W}$ and only the estimator associated with the sequence $V_n^{-1}\xrightarrow{p}V_o^{-1}$ is efficient. In the MD case, the class of estimators is indexed by the strictly convex function $\gamma(\cdot)$, but for each choice of $\gamma(\cdot)$ the resulting MD estimator is efficient.

MD estimators obtain estimates of the parameter θ_o by selecting a probability compatible with the moment condition as close as possible to the estimated probability of the data. It is not surprising that the (normalized) weights $\{\omega_i\}$ represent the probability structure implied by the model. Let $Q_o(w)=\int \mathbb{1}_{(w_i\le w)}\mathrm{d}Q_o$ and $\sigma_w=Q_o(w)(1-Q_o(w))$. Under the conditions of Theorem 5.5, the empirical distribution function given by $\mathbb{Q}_n(w)=n^{-1}\sum_i^n\mathbb{1}_{(w_i\le w)}$ converges pointwise to the p.d.f. of $\{w_i\}$ and has limiting normal distribution described by

$$\sqrt{n}(\mathbb{Q}_n(w)-Q_o(w))\xrightarrow{d}\mathrm{N}(0,\sigma_w)$$

Let $\mathbb{M}_n(w)=\sum_i^n\mathbb{1}_{(w_i\le w)}\hat\omega_i$, where $\widehat\omega=(\widehat\omega_1,\widehat\omega_2,\ldots,\widehat\omega_n)$ are the normalized MD weights.

Theorem 5.6: *Suppose Assumptions A-B hold. Then $\mathbb{M}_n(w)\xrightarrow{p}Q_o(w)$ point-wise and $\mathbb{M}_n(w)$ has limiting normal distribution given by*

$$\sqrt{n}(\mathbb{M}_n(w)-Q_o(w))\xrightarrow{d}N(0,\sigma_w-q(w)'V_o^{-1}q(w))$$

where $q(w)=\int\mathbb{1}_{(w_i\le w)}q_i(w_i,\theta_o)\mathrm{d}Q_o$

Comparing the variance of the asymptotic distributions of $\mathbb{Q}_n(w)$ and $\mathbb{M}_n(w)$ yields that $\mathbb{M}_n(w)$ is asymptotically more efficient than the empirical distribution function. This is easily proved by noticing that $q(w)'V_o^{-1}q(w) > 0$, by positive definiteness of V_o^{-1}. The intuition behind this result is that the estimated c.d.f. based on $\mathbb{M}_n(w)$ is asymptotically more efficient than the empirical distribution function because it incorporates the information in $Eq_i(w_i, \theta_o) = 0$.

Theorem 5.6 has an important implication. Sample counterparts of population moments can be weighted by weights $\{\omega_i\}$ that are proportional to efficient estimates of the pdf of w. For example, the following expression gives a consistent variance estimator

$$\hat{S}_\omega = \left(\hat{\Gamma}_\omega' \hat{V}_\omega^{-1} \hat{\Gamma}_\omega\right)^{-1}$$

where $\hat{\Gamma}_\omega = \left(\sum_{i=1}^n \hat{\omega}_i \nabla_\theta q_i(\hat{\theta})\right)$ and $\hat{V}_\omega = \sum_{i=1}^n \hat{\omega}_i q(\hat{\theta}) q_i(\hat{\theta})'$. Consistency follows by noting that, by Taylor's expansion, $\widehat{\omega}_i = 1/n + o_p(n^{-1})$, and by triangular inequality

$$\begin{aligned}\left\|\hat{V}_\omega - V_o\right\| &\leq \left\|\sum_{i=1}^n (\hat{\omega}_i - n^{-1}) q_i(\hat{\theta}) q_i(\hat{\theta})'\right\| + o_p(1) \\ &\leq \max_{\hat{\omega}_i, i \leq n} n \left|\hat{\omega}_i - n^{-1}\right| \cdot O_p(1) = o_p(1)\end{aligned}$$

and similarly $\left\|\hat{\Gamma}_\omega - \Gamma_o\right\| = o_p(1)$. A robust estimator of V_o is proposed by Imbens, Spady, and Johnson (1998) and it is given by

$$\hat{V}_R = \sum_{i=1}^n \hat{\omega}_i q_i(\hat{\theta}) q_i(\hat{\theta})' \left(n \sum_{i=1}^n \hat{\omega}_i^2 q_i(\hat{\theta}) q_i(\hat{\theta})'\right)^{-1} \sum_{i=1}^n \hat{\omega}_i q_i(\hat{\theta}) q_i(\hat{\theta})' \tag{22}$$

In the MD framework, inference can be conducted by using the usual test statistics briefly discussed in Section 3. Suppose one wants to test an hypothesis concerning a subvector of θ_o. For instance, let $\theta = (\theta_1', \theta_2')'$ where $\theta_1 \in \Theta_1 \subset \mathbb{R}^d$ and $\theta_2 \in \Theta_2 \subset \mathbb{R}^{k-d}$ and consider testing the null hypothesis $H_o : \theta_1 = \theta_{1,o}$ versus $H_1 : \theta_1 \neq \theta_{1,o}$. The constrained MD estimation problem is defined as

$$\begin{pmatrix} \hat{\pi}_r \\ \hat{\theta}_r \end{pmatrix} = \arg \min_{\theta_1 = \theta_{1,o}, \theta_2 \in \Theta_2, \pi} \left\{ \frac{1}{n} \sum_{i=1}^n \gamma(n\pi_i) \middle| \sum_{i=1}^n \pi_i q_i(\theta) = 0; \sum_{i=1}^n \pi_i = 1; \pi \in (a_\gamma, b_\gamma) \right\} \tag{CMD}$$

Let $\hat{\Gamma}_{\omega^r}$ be a consistent estimator of $E[\nabla_{\theta_1} q_i(\theta_o)]$, where $\nabla_{\theta_1} q_i(\theta_o)$ denotes the $m \times d$ matrix collecting the derivatives with respect to the restricted component of θ, and

$\hat{S}_{\omega^r} = \hat{\Gamma}'_{\omega^r}\hat{V}_{\omega}\hat{\Gamma}'_{\omega^r}$. Then one can use the unrestrcted MD estimator $\hat{\theta}_1$ and the restricted the (restricted) weights $\hat{\omega}$ to construct a Wald's test using the variance $\hat{S}^{-1}_{\omega^r}$

$$Wald_{\omega}(\hat{\theta}) = n \cdot (\hat{\theta}_1 - \theta_{1,o})'\hat{S}^{-1}_{\omega^r}(\hat{\theta}_1 - \theta_{1,o}) \xrightarrow{d} \chi^2_d$$

or a Lagrange Multiplier test

$$LM_{\omega^r}(\hat{\theta}) = nq(\hat{\theta}_r)'\hat{V}^{-1}_{\omega^r}\hat{\Gamma}_{\omega^r}\hat{S}_{\omega^r}\hat{\Gamma}_{\omega^r}\hat{V}^{-1}_{\omega^r}q(\hat{\theta}_r) \xrightarrow{d} \chi^2_d \tag{23}$$

A criterion based test can be constructed by using the objective function of the MD problem. An advantage of the LR test based on the objective function of the MD estimator is that no estimation of the variance matrix is necessary to obtain a χ^2 calibration.

Theorem 5.7: *Let Assumptions A-B and Assumption 1(γ) hold. Then*

$$LR_{\omega} = 2\left\{\sum_{i=1}^{n}[\gamma(n\hat{\pi}^r_i) - \gamma(n\hat{\pi}_i)]\right\} \xrightarrow{d} \chi^2_d$$

where $\hat{\pi}_i = \widetilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))/n$ are the unconstrained weights and $\hat{\pi}^r = \widetilde{\gamma}_1(\hat{\lambda}'_r q_i(\hat{\theta}_r))/n$ are the constrained weights obtained as solution of CMD.

Notice that the Likelihood Ratio statistics of Theorem 5.7 is expressed in terms of the un-normalized weights $\hat{\pi}_i$, but it could be also defined in terms of the normalized weights $\hat{\omega}_i$. When a close form for the expression of the divergence is not available, test statistics can be based on the dual function $\psi(x)$, by considering the constrained GEL problem

$$\max_{\theta_1=\theta_{1,o},\theta_2\in\Theta_2}\ \min_{\lambda\in\Lambda_n(\theta_{1,o},\theta_2)} \frac{1}{n}\sum_{i=1}^{n}\psi(\lambda' q_i(\theta)) \tag{CGEL}$$

where $\psi(x)$ satisfies Assumption 2(ψ).

Theorem 5.8: *Let Assumptions A-B and Assumption 2(ψ) hold. Then*

$$LR_{\omega} = -2\left\{\sum_{i=1}^{n}\psi(\hat{\lambda}'_r q_i(\hat{\theta}_r)) - \sum_{i=1}^{n}\psi(\hat{\lambda}' q_i(\hat{\theta}))\right\} \xrightarrow{d} \chi^2_d$$

where $(\hat{\lambda}', \hat{\theta}')'$ and $(\hat{\lambda}'_r, \hat{\theta}'_r)'$ are the unconstrained and constrained GEL estimates of λ and θ_o.

The Lagrange multiplier itself can be used to construct test statistics. The intuition is that if H_o holds, the constrained Lagrange multiplier converges to zero, while if the null hypothesis does not hold the Lagrange multiplier converges to a value other than zero. Let $\hat{\lambda}_r$ denote the constrained Lagrange multiplier associated with the problem CMD or to problem CGEL. The following result holds under either Assumption $1(\gamma)$ if the Lagrange Multiplier associated with (CMD) is considered or under Assumption $2(\psi)$ if the Lagrange multiplier associated with (CGEL) is considered.

Theorem 5.9: *Let Assumptions A-B hold. Then under* $H_o : \theta_1 = \theta$

$$\begin{aligned} \hat{\lambda}_r^{'} \hat{P}_{\omega^r}{}^{-g} \hat{\lambda}_r &\xrightarrow{d} \chi^2_d \\ \hat{\lambda}_r' \hat{V}_{\omega^r}^{-1} \hat{\lambda}_r &\xrightarrow{d} \chi^2_d \end{aligned}$$

where $\hat{P}_{\omega^r}(\hat{\theta}^r)$ *and* $\hat{V}_{\omega^r}(\hat{\theta}^r)$ *are, under* H_o, *consistent estimators of* P_o *and* V_o *respectively.*

A special case is given when one wishes to test hypotheses concerning the full vector θ_o. Consider the null hypothesis $H_o : \theta = \theta_o$ versus $H_1 : \theta \neq \theta_o$. In this special case the MD problem reduces to

$$\hat{\pi}^r = \arg\min_{\pi} \left\{ \frac{1}{n} \sum_{i=1}^{n} \gamma(n\pi_i) \left| \sum_{i=1}^{n} \pi_i q(w_i, \theta_o) = 0; \sum_{i=1}^{n} \pi_i = 1; \pi \in (a_\gamma, b_\gamma) \right. \right\} \tag{24}$$

and hence only optimization with respect to π is required. Similarly, for the GEL, the problem reduce to

$$\min_{\lambda \in \Lambda_n(\theta_o)} \frac{1}{n} \sum_{i=1}^{n} \psi(\lambda' q_i(\theta_o))$$

As $\psi(x)$ is strictly convex by Assumption $2.(\psi)$, this is particularly convenient and the numerical convergence to the solution is rapid.

MD techniques provide, as GMM does, a framework for testing the null hypothesis of overidentification. Similar tests have been proposed by Qin and Lawless (1994) for EL, by Kitamura and Stutzer (1997) for ET and by NS for GEL. If $\hat{\lambda}$ and $\hat{\theta}$ denote the GEL estimators and $\hat{\omega}_i$ denotes the normalized MD weight, we have that

$$2 \left[\sum_{i=1}^{n} \gamma(n\hat{\pi}_i) \right] \xrightarrow{d} \chi^2_{m-k}$$

and

$$-2 \left[\sum_{i=1}^{n} \psi(\widehat{\lambda}' q_i(\widehat{\theta})) \right] \xrightarrow{d} \chi^2_{m-k}$$

Imbens, Spady and Johnson (1998) consider test of overidentification based on the Lagrange multiplier obtained from the ET and CUE problems. These test can be extended to all MD [GEL] estimators.

Theorem 5.10: *Let Assumptions A-B and Assumption 1(γ) [Assumption 2(ψ)] hold. Then,*

$$\hat{\lambda}'\hat{P}_{\omega}{}^{-g}\hat{\lambda} \xrightarrow{d} \chi^2_{m-k}$$
$$\hat{\lambda}\hat{V}_{\omega}{}^{-1}\hat{\lambda} \xrightarrow{d} \chi^2_{m-k}$$

where $\hat{P}_\omega$ and $\hat{V}_\omega$ are consistent estimates of P_o and V_o.

6 Bayesian Connections

In the previous sections it was shown that, under weak conditions on the moment function $q(w,\theta)$, MD methods deliver consistent and asymptotically normal estimators, regardless of the specific divergence $\gamma(\cdot)$ chosen. An important issue is whether the choice of the divergence can be given a probabilistic interpretation. This section shows that the weighting implied by a divergence can be obtained by imposing a prior on the space of distributions that support the sample moment conditions. The prior is derived by the Maximum Entropy principle with respect to a reference distribution. For a fixed θ, different assumptions on the reference distribution deliver weighting schemes that are equivalent to those obtained by using specific divergences.

Given the sample of observations $w^n = (w_1, w_2, ..., w_n\}$, define

$$q_n(w^n, \theta; \nu) = \frac{1}{n}\sum_{i=1}^{n} q(w_i, \theta)\nu_i$$

Let $\tilde{p}(\nu^n)$ denote a reference prior on $\nu^n = (\nu_1, \nu_2, ..., \nu_n)$

$$\tilde{p}(\nu^n) = \prod_{i=1}^{n} \tilde{p}(\nu_i)$$

The support of $\tilde{p}(\nu^n)$ can be either continous or discrete. We consider finding a $p(\nu^n)$ under which the sample moment conditions are satisfied in expectation. The tool to find such a distribution is the Maximum Entropy principle introduced by Jaynes (1957, 1968). $p(\nu^n)$ is chosen in such a way to have the maximum entropy with respect to the reference prior $\tilde{p}(\nu^n)$. Consider the following set of probability distributions on ν^n

$$\mathcal{P}^b_\theta = \left\{ p(\nu^n) | \int q_n(w^n, \theta; \nu) p(\nu^n) d\nu^n = 0 \right\}$$

The maximum entropy $p(\nu^n)$ is obtained by solving the following problem

$$\min_{p(\nu^n|\theta) \in \mathcal{P}^b_\theta} \int \log \left(\frac{p(\nu^n)}{\tilde{p}(\nu^n)} \right) p(\nu^n) \mathrm{d}v^n \tag{25}$$

where the integral over the vector ν^n is a shorthand notation for multiple integral.

Theorem 6.1: *The solution to the problem (25) is given by*

$$p^*_n(\nu^n|w^n, \theta) = \frac{\exp\left[\frac{1}{n}\sum_i^n \tilde{\lambda}' q(w_i, \theta)\nu_i\right] \tilde{p}(\nu^n)}{\int \exp[\frac{1}{n}\sum_i^n \tilde{\lambda}' q(w_i, \theta)] \tilde{p}(\nu^n) \mathrm{d}\nu^n} \tag{26}$$

where $\tilde{\lambda}$ is the minimand of

$$\min_{\lambda \in \tilde{\Lambda}_n(\theta)} \log \int \exp[\frac{1}{n}\sum_i^n \tilde{\lambda}' q(w_i, \theta)] \tilde{p}(\nu^n) \mathrm{d}\nu^n]$$

where $\tilde{\Lambda}_n(\theta) = \{\lambda : \log \int \exp[\frac{1}{n}\sum_i^n \tilde{\lambda}' q(w_i, \theta)] \tilde{p}(\nu^n) \mathrm{d}\nu^n < \infty\}$.

Under $p^*_n(\nu^n|w^n, \theta)$ the expected value of the sample moment conditions is zero. It also holds that $\{\nu_1, \nu_2, ..., \nu_n\}$ are independent as the following result shows.

Theorem 6.2: *Under the $p^*(\nu^n|w^n, \theta)$, $\{\nu_1, \nu_2, ..., \nu_n\}$ are independent, each with distribution*

$$p^*_n(\nu_i|w^n, \theta) = \exp[\tau_i \nu_i - \varphi(\tau_i)] \tilde{p}(\nu_i|\theta)$$

where $\tau_i = \frac{1}{n}\sum_i^n \tilde{\lambda}' q(w_i, \theta)$ and

$$\varphi(\tau) = \log \int \exp(\tau\nu) \tilde{p}(\nu) d\nu$$

and

$$\tilde{\lambda} = \arg \min_{\lambda \in \tilde{\Lambda}_n(\theta)} \sum_{i=1}^n \varphi(\tau_i)$$

The term $\varphi(x)$ in Theorem 6.2 is the logarithm of the moment generating function of ν. A property of $\varphi(\cdot)$ is that its derivative is given by

$$\varphi_1(\tau_i) = \int \nu_i \left[\exp(\tau_i \nu_i) - \varphi(\tau_i)\right] \tilde{p}(\nu_i) \mathrm{d}\nu_i$$

Now consider the random measure induced by ν^n

$$\mu_n = \sum_{i=1}^{n} \nu_i \delta(w_i)$$

where $\delta(w_i)$ denotes the Kronecker delta. Taking the expected value of μ_n with respect to $p^*(\nu^n|w^n, \theta)$ yields

$$\begin{aligned} G_n(w^n, \theta) &= \sum_{i=1}^{n} \int \nu_i \exp[\tau_i \nu_i - \varphi(\tau_i)] \tilde{p}(\nu_i) d\nu_i \delta(w_i) \\ &= \sum_{i=1}^{n} \varphi_1(\tau_i) \delta(w_i) \end{aligned}$$

As $n \to \infty$, it can be shown that $G_n(w_n, \theta)$ converges in probability to the solution of the following minimum divergence problem

$$\int \gamma_\varphi(\frac{dG}{dP_o}) dP_o$$
$$st. \int q(w, \theta) dG = 0$$

where $\gamma_\varphi(\cdot)$ is a divergence that satisfies Assumption $1(\gamma)$ and whose functional form depends on the reference distribution $\tilde{p}(\nu)$. Now we study the form of $G_n(w^n, \theta)$ for specific reference distribution $\tilde{p}(\nu^n)$. For a given $\tilde{p}(\nu^n)$ the function $\varphi_1(\cdot)$ gives the weights of EL, ET and CUE and other MD estimators. In practice, we need to carefully choose a reference distribution $\tilde{p}(\nu)$ such that its logarithmic moment generating function coincides with the weights of the MD estimators.

Theorem 6.3 (EL): *If the reference prior is given by $\tilde{p}(\nu^n) = \prod_{i=1}^{n} e^{-\nu_i}$, that is $\{\nu_1, \nu_2, ..., \nu_i\}$ are independently and identically distributed as Exponential with mean 1, then*

$$G(w^n, \theta) = \frac{1}{n} \sum_{i=1}^{n} (1 - \hat{\lambda}' q(w_i, \theta))^{-1} \delta(w_i)$$

and

$$\hat{\lambda} = \arg \min_{\lambda \in \Lambda_n(\theta)} \sum_{i=1}^{n} -\log(1 - \lambda' q(w_i, \theta))$$

and $\Lambda_n(\theta) = \{\lambda : \lambda' q(w_i, \theta) < -1, i = 1, ..., n\}$.

Theorem 6.4 (ET): *If the reference prior is given by* $\tilde{p}(\nu^n) = \prod_{i=1}^{n} \nu_i e^{-\nu_i}$*, that is* $\{\nu_1, \nu_2, ..., \nu_n\}$ *are independently and identically distributed as Poisson with parameter* 1*, then*

$$G(w^n, \theta) = \frac{1}{n} \sum_{i=1}^{n} \exp(\hat{\lambda}' q(w_i, \theta)) \delta(w_i)$$

and

$$\hat{\lambda} = \arg \min_{\lambda \in \Lambda_n(\theta)} \sum_{i=1}^{n} \exp(\lambda' q(w_i, \theta))$$

and $\Lambda_n(\theta) = \{\lambda : -\infty < \lambda < +\infty\}$.

Theorem 6.5 (CUE): *If the reference prior is given by* $\tilde{p}(\nu^n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\nu_i^2}$*, that is* $\{\nu_1, \nu_2, ..., \nu_n\}$ *are independently and identically distributed as* $Normal$ *with mean* 1 *and variance* 1*, then*

$$G(w^n, \theta) = \frac{1}{n} \sum_{i=1}^{n} (1 + \hat{\lambda}' q(w_i, \theta))$$

and

$$\hat{\lambda} = \arg \min_{\lambda \in \Lambda_n(\theta)} \sum_{i=1}^{n} \left\{ \lambda' q(w_i, \theta) + [\lambda' q(w_i, \theta)]^2 / 2 \right\}$$

and $\Lambda_n(\theta) = \{\lambda : -\infty < \lambda < +\infty\}$.

Notice that the set $\Lambda_n(\theta)$ corresponds to the domain of the moment generating function of the reference distribution considered in each case.

The results presented in this section allow to define a correspondence between a reference distribution and the weights of EL, ET, CUE, and possibly of other members of the MD class. As in Kim (2002), one may also interpret $G(w^n, \theta)$ as a limited information measure. Using standard Bayesian arguments, the limited information posterior can be based on the limited information likelihood. Let $p(\theta)$ be a prior density of θ. Then a posterior can be derived from Bayes' rule

$$p(\theta|w^n) = p(w^n)^{-1}\{p(\theta)\prod_{i=1}^{n} g_n(w_i,\theta)\}$$

where $g_n(w_i,\theta) = \varphi_1(\tau_i)$ and $p(w^n) = \int_\Theta p(\theta)g(w^n,\theta)d\theta$. If the reference distribution is an Exponential with mean 1, the limited information posterior is given by

$$p(\theta|w^n) \propto p(\theta)\prod_{i=1}^{n}(1-\hat{\lambda}'q(w_i,\theta))^{-1} \tag{27}$$

where

$$\hat{\lambda} = \arg\min_{\lambda\in\Lambda_n(\theta)}\sum_{i=1}^{n} -\log(1-\lambda' q(w_i,\theta))$$

If the reference distribution is a Poisson with parameter 1, the limited information posterior is given by

$$p(\theta|w^n) \propto p(\theta)\prod_{i=1}^{n}\exp(\hat{\lambda}'q(w_i,\theta)) \tag{28}$$

where

$$\hat{\lambda} = \arg\min_{\lambda}\sum_{i=1}^{n}\exp(\lambda' q(w_i,\theta))$$

Limited information posteriors coincide with the posteriors one would obtain by replacing the likelihood with the product of the weights obtained from Minimim Divergence techniques. The posterior involving the weights of EL is analyzed by Lazar (2003). In a very interesting paper, Schennach (2005) studies the Exponential Tilted Empirical Likelihood, that is a posterior likelihood that arises from assuming a prior on the space of distributions and corresponds to the posterior given in (28). In general, as long as one is able to find distributions with moment generating functions that correspond to the weights of the MD, one can always specify the limited information posterior as

$$p(\theta|w^n) \propto p(\theta)\prod_{i=1}^{n}\tilde{\gamma}_1(\hat{\lambda}'q(w_i,\theta))$$

where λ is obtained by solving

$$\hat{\lambda} = \arg\min_{\lambda}\sum_{i=1}^{n}\psi(\lambda' q_i(\theta))$$

and $\psi(\cdot)$ can be obtained by duality from $\gamma(\cdot)$. The analysis of this section suggests the intriguing possibility of carrying nonparametric Bayesian inference in model specified by moment conditions that are overidentified.

7 Higher Order Expansions

In this section we explore the higher order properties of MD estimators. The analysis is similar under some aspects to that in NS, but it also differs in many regards. Importantly, the emphasis is different. While they focus on the relation between GEL and GMM estimators, we examine the higher order properties of members of the MD family of estimators.

We look for an expansion of $\hat{\theta}$ of the following form

$$(\hat{\theta} - \theta_o) = u_n + b_n + r_n + O_p(n^{-2}) \tag{29}$$

where $u_n = O_p(n^{-1/2})$, $b_n = O_p(n^{-1})$ and $r_n = O_p(n^{-3/2})$. The terms in the expansion are tractable, in the sense that they are expressed as sums and products of sample averages. Similar expansions have been carried out in the context of instrumental variables models by Nagar (1959) and Hausman (2001) and Hausman, Hahn, and Kuersteiner (2001), among others.

Definition 7.1 (Higher Order Bias): *If an estimator $\hat{\theta}$ of θ_o admits an expansion as in (29), its $O(n^{-1})$ bias is given by*

$$B_{-1}(\hat{\theta}) = E[u_n] + E[b_n]$$

However, to obtain an expression for the $O(n^{-1})$ bias an expansion of order $n^{-3/2}$ is sufficient

$$(\hat{\theta} - \theta_o) = u_n + b_n + O_p(n^{-3/2})$$

Definition 7.2 (Higher Order Mean Square Error): *If an estimator $\hat{\theta}$ of θ_o admits an expansion as in* (29)*, the $O(n^{-2})$ MSE is given by*

$$\mathcal{M}_{-2}(\hat{\theta}) = E(u_n u_n' + u_n r_n' + r_n u_n' + b_n b_n')$$

A few remarks are worth making with respect to the higher order expansions. The higher order bias and MSE obtained by taking expectation of the corresponding terms in the expansion (29) are equivalent to the bias and MSE obtained through a valid $o(n^{-1})$ Edgeworth expansion of $\sqrt{n}(\hat{\theta}-\theta_o)$, if the last term in the expansion is appropriately bounded. The bias of order $O(n^{-1})$ and MSE of order $O(n^{-2})$ are defined as expectations of terms that are bounded in probability. Even if the remainders are bounded in probability they can still diverge in expectation. As pointed out by Srinivasan (1970), it is possible that an estimator posseses a valid asymptotic expansion, yet it does not have finite sample moments. In this sense higher order comparisons of estimators could be misleading. This is important in the context of MD estimators. For instance, in a linear simultaneous equations setting, Kunitomo and Matsushita (2003) show that the Empirical Likelihood estimator does not have finite moments.

7.1 Asymptotic Bias

Asymptotic expansions require that additional moments of the underlying distribution of the data exist. Given the nonlinearity of the estimating equation defining the MD estimators, the terms in the expansions (29) are not simple functions of w and smoothness assumptions must imposed on $q(w,\theta)$.

The following notation is used. The partial derivatives with respect to θ_j, $j=1,\dots,k$ are denoted by $q_i^j(\theta)=(\partial/\partial\theta_j)q_i(\theta)$ and $q_o^j=E[q_i^j(\theta_o)]$. The second derivatives with respect to θ_j and θ_r, $j,r=1,\dots,k$ are $q_i^{jr}(\theta)=(\partial^2/\partial\theta_j\partial\theta_r)q_i(\theta)$ and $q_o^{jr}=q_i^{jr}(\theta_o)$. The higher order derivatives are defined accordingly.

Assumption C: There is a $\epsilon>0$ and $\mathcal{B}(w_i)$, $E[\mathcal{B}(w_i)^5]<\infty$ such that for any $\theta\in\mathcal{S}(\theta_o,\epsilon)$ and any $j,r,s=1,\dots,k$: *i)* $\sup_{\theta\in\Theta}\|q_i(\theta)\|\leq\mathcal{B}(w_i)$; *ii)* $q_i^{jrs}(\theta)$ exists; *iii)*$\|q_i^j(\theta)-q_o^j\|\leq\mathcal{B}(w_i)$; *iv)* $\|q_i^{jr}(\theta)-q_o^{jr}\|\leq\mathcal{B}(w_i)$; *v)* $\|q_i^{jrs}(\theta)-q_o^{jrs}\|\leq\mathcal{B}(w_i)\|\theta-\theta_o\|$; *v)* $\gamma(\cdot)$ is four times differentiable in a neighborhood of 1.

We first provide the $O_p(n^{-3/2})$ expansion of $(\hat{\theta}-\theta_o)$ and of $\hat{\lambda}$. Let $B_o=S_o\Gamma_o'V_o^{-1}$, $u_n=B_o\sum_i^n q_i(\theta_o)/n$, $l_n=P_o\sum_i^n q_i(\theta_o)/n$. The vectors ∇_1 and ∇_2 are of size $k\times 1$ and $m\times 1$ and their expression is given in the Appendix.

Theorem 7.1: *Suppose Assumption A-C hold. Then the MD estimators and the associated Lagrange multiplier admit $O_p(n^{-3/2})$ expansion*

$$(\hat{\theta}-\theta_o)=u_n+b_n^{\theta}+O_p(n^{-3/2})$$

and

$$\hat{\lambda} = l_n + b_n^{\lambda} + O_p(n^{-3/2})$$

where

$$\begin{aligned} b_n^{\theta} &= -B_o\Gamma_n u_n + S_o\Gamma_n' u_n - B_o V_n l_n + \frac{1}{2}S_o\nabla_1 - \frac{1}{2}B_o\nabla_2 \\ b_n^{\lambda} &= P_o\Gamma_n u_n + B_o'\Gamma_n' l_n + P_o V_n l_n - \frac{1}{2}B_o'\nabla_1 - \frac{1}{2}P_o\nabla_2 \end{aligned}$$

There are minor differences between the result of Theorem 7.1 and the expansion given in NS. First, they derive the asymptotic bias from the $O_p(n^{-2})$ expansion and thus they have to make stronger assumptions on the moments of the derivatives of the moment functions. Here, we follow Rilestone, Ullah and Srivastava (1996) and derive the bias from the $O_p(n^{-3/2})$ expansion, avoiding making assumptions on fourth derivatives of the moment function. Second, NS consider the expansion of the vector $((\hat{\theta} - \theta_o)', \hat{\lambda})$, while Theorem 7.1 gives an explicit expansion for $(\hat{\theta} - \theta_o)$ and $\hat{\lambda}$.

The bias up to order $O_p(n^{-1})$ of MD estimators is obtained by taking the expectation of the first term in the expansion for $(\hat{\theta} - \theta_o)$. Let a be the $m \times 1$ vector whose element j is given by

$$a_j = \text{Trace}\left\{S_o E\left[\partial^2 q_{ij}(\theta_o)/\partial\theta\partial\theta'\right]\right\}/2$$

where $q_{ij}(\theta_o)$ denotes the jth element of $q_i(\theta_o)$.

Theorem 7.2: *Suppose Assumption A-C hold. Then the asymptotic bias up to order $O_p(n^{-1})$ for a MD estimator of θ_o is given by*

$$B_{-1}(\hat{\theta}) = n^{-1}\left\{b_1 + (1 - \frac{\gamma_3}{2})b_2)\right\} \tag{30}$$

where $b_1 = B_o\left\{E\left[\nabla_\theta q_i(\theta_o)B_o q_i(\theta_o)\right] - a\right\}$ and $b_2 = B_o E\left[q_i(\theta_o)q_i(\theta_o)'P_o q_i(\theta_o)\right]$.

The formula for the bias of MD estimators given in Theorem 7.2 is analogous to that of NS for GEL estimators. There, the bias involves a parameter that depends on the function $\psi(\cdot)$ that characterizes the GEL estimators; here it depends on the third derivatives of $\gamma(\cdot)$. Using the result in Theorem 5.1, it follows that, under Assumption C, $\psi_3(0) = \gamma_3(1)$. In the Cressie Read family, the only estimator with $\gamma_3 = 2$ is EL.

The bias depends also on the curvature of the model through the term $a(\theta_o)$. For highly nonlinear models the bias induced by this term can be relatively large. When

$q(w, \theta)$ has non zero generalized third moments, only MD estimators with $\gamma_3 = 2$ get rid of the bias induced by the asymmetries of the moment functions.

The bias corrected estimator can in theory be obtained by looking at the sample counterpart of the expressions involved in the $O(n^{-1})$ bias formula given in Theorem 7.2. If $\hat{\theta}$ is the original MD estimator, $b_1(\theta_o)$ and $b_2(\theta_o)$ can be estimated by

$$\begin{aligned} \hat{b}_1 &= n^{-1}\hat{B}_n \left(\sum_{i=1}^{n} \nabla_\theta q_i(\hat{\theta})\hat{B}_n q_i(\hat{\theta}) - \hat{a}(\hat{\theta}) \right) \\ \hat{b}_2 &= n^{-1}\hat{B}_n \sum_{i=1}^{n} q_i(\hat{\theta})q_i(\hat{\theta})'\hat{P}_n q_i(\hat{\theta}) \end{aligned}$$

where $\hat{B}_n$, $\hat{S}_n$ and $\hat{P}_n$ are sample counterparts of B_o, S_o and P_o, respectively. As pointed out by NS, the sample terms involved in the expression above can be weighted by the efficient estimated probabilities given in Section 5. The assumptions that need to hold in order to derive the $O(n^{-3/2})$ expansion of $(\hat{\theta} - \theta_o)$ are also sufficient for $\hat{b}_1$ and $\hat{b}_2$ to be consistent estimator of $b_1(\theta_o)$, $b_2(\theta_o)$. It follows that the bias corrected MD estimator defined as

$$\hat{\theta}_{bc} = \hat{\theta} - n^{-1}\left\{ \hat{b}_1 + (1 - \frac{\gamma_3}{2})\hat{b}_2 \right\} \tag{31}$$

is unbiased of order $O(n^{-1})$. The formula of the bias correction simplifies when $\gamma_3 = 2$, because one needs not to estimate the term $\hat{b}_2$.

From an applied perspective the bias correction can be a difficult exercise, but nevertheless it is a feasible strategy. The critical point is rather to assess if the bias correction can lead to substantial improvements. For example, Hahn, Hausman and Kuersteiner (2002) present Monte Carlo simulations of the Nagar's bias adjusted IV estimator that show that the bias correction may be ineffective over many points in the parameter space considered. On the other hand, Rilstone, Srivastava, and Ullah (1996) apply bias correction to nonlinear logistic regressions and show through Monte Carlo that in these situations the bias correction can lead to substantial improvements.

An interesting result is that in the important special case of nonlinear instrumental variables models, the first order conditions of MD can be slightly modified in order to deliver estimators that have smaller bias than the original MD.

Let consider

$$q_i(w_i, \theta) = z_i g(x_i, \theta) \tag{32}$$

where $g(x_i, \theta) : \mathrm{X} \times \mathbb{R}^k \to \mathbb{R}$ and $\{z_i\}$ is a $m \times 1$ vector of random variables such that

$Ez_i g(x_i, \theta_o) = 0$. Let $G_i(\theta) = \nabla_\theta g(x_i, \theta)$.

Assumption IV: (i) $\{x_i, z_i'\}$ are iid random variables; (ii) θ_o lies in the interior of Θ; (iii) $E(z_i z_i')$ has full column rank; (iv) $E\|z_i\|^2 \leq \infty$; (v) $g_i(\theta)$ is continuous for each $\theta \in \mathcal{S}(\theta_o, \epsilon)$; (vi) $E[\sup_{\theta\in\Theta} \|g_i(\theta)\|^2] < \infty$ (vii) $E[\sup_{\theta\in\Theta} \|G_i(\theta)\|] < \infty$; (viii) $\sigma_{gG} = E[g_i(\theta_o)G_i(\theta_o)|z_i]$;(iii) $\sigma_g^2 = E[g_i(\theta_o)^2|z_i]$ (ix) $\sigma_g^3 = E[g(x_i, \theta_o)^3|z_i]$; (ix) Assumption C-D holds with $q(w, \theta)$ replaced by $g(x, \theta)$.

Under Assumption IV, the MD estimator is consistent and asymptotically normal as it can be easily seen by comparing the conditions given for the general case. The bias for this model is given by the following result.

Theorem 7.3: *Under Assumption IV the bias of the MD estimator is given by*

$$B_{-1}(\theta_o) = \tilde{S}_o \sigma_{gG}/\sigma_g^2 - \tilde{B}_o \tilde{a} + (1 - \frac{\gamma_3}{2})\sigma_g^3 \sigma_z^3 \tag{33}$$

where $\tilde{a}$ is the $k \times 1$ vector whose element j is given by

$$\tilde{a}_j = Trace(\tilde{S}_o E(\partial/\partial\theta\partial\theta')g_i(\theta_o)z_{ij})$$

and $\sigma_z^3 = E(z_i z_i' \tilde{P}_o z_i)$.

The expression for the bias in (33) specializes immediately to the bias of the homoschedastic linear IV as given by $B_{-1}(\hat{\theta}) = -\tilde{S}_o E(x_i \varepsilon_i|z_i)/\sigma^2$, that, as mentioned in NS, is the bias of the Limited Informtion Maximum Likelihood estimator. Now consider the estimator that solves the following equations

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^n \tilde{\gamma}_1(\lambda' z_i g_i(\theta)) z_i g(x_i, \theta) & (34) \\ 0 &= \frac{1}{n}\sum_{i=1}^n \tilde{\gamma}_1(\kappa\lambda' z_i g_i(\theta)) G_i(\theta)' z_i' \lambda' & (35) \end{aligned}$$

for some $\kappa > 0$. For $\kappa = 1$, these estimating equations are equivalent to those of MD estimators given in (20) and (21).

Theorem 7.4: *Let $\hat{\theta}$ and $\hat{\lambda}$ be the solution of the estimating equations in* (34) *and* (35). *Then $\hat{\theta} = \theta_o + o_p(1)$ and has asymptotic bias given by*

$$B_{-1}(\hat{\theta}) = \kappa_\theta \tilde{S}_o \sigma_{gG}/\sigma_g^2 - \tilde{B}_o \tilde{a} + (1 - \frac{\gamma_3}{2})\sigma_g^3 \sigma_z^3$$

where $\kappa_\theta = \kappa(m-k) - (m-k-1)$.

The above result shows that the first term of the bias can be eliminated by setting $\kappa = (m-k-1)/(m-k)$, so that the asymptotic bias of the estimator solving (34) and (35) reduces to $\tilde{B}_o\tilde{a} + (1 - \frac{\gamma_3}{2})\sigma_g^3\sigma_z^3$. If $\gamma_3 = 2$ the bias reduces to $\tilde{B}_o\tilde{a}(\theta_o)$ and it vanishes when the model is linear, since then $\tilde{a} = 0$.

It is difficult to express the estimating equations as first order conditions of an optimization problem in the MD framework. If one considers the GEL representation, equations (34) and (35) can be obtained by considering the following nested optimization problem

$$\lambda(\theta) = \arg\min_{\lambda \in \Lambda(\theta)} \frac{1}{n}\sum_{i=1}^{n} \psi(\lambda q_i(\theta))$$

and

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^{n} \psi(\kappa\lambda(\theta) q_i(\theta))$$

It is easy to verify that this nested problem gives first order conditions that are equivalent to (34) and (35).

7.2 Mean Square Error

Adapting the argument in Pfanzagl and Wefelmeyer (1979), NS show that the $O(n^{-1})$ bias corrected EL estimator is third order efficient, in the sense that it has the lowest $O(n^{-2})$ MSE among all the bias corrected estimators based on the same set of moment conditions. The higher order efficiency of EL only holds among bias corrected estimators. If the bias corrections are dropped, then EL may not have the smallest MSE.

In many applications, however, the bias term $b_2(\theta_o)$ can be large and hence it is interesting to consider MD estimators with $\gamma_3 = 2$. Even if a direct comparison of higher order MSE of MD estimators is difficult in the general case, it turns out that if one restricts attention to the subclass of MD estimators with $\gamma_3 = 2$ some interesting results can be given.

Comparing the higher order MSE of MD estimators with $\gamma_3 = 2$ amounts to verify whether other members of this class share the same higher order efficiency. Let $\hat{\theta}_{el}$ denote the EL estimator and $\tilde{\theta}_{md}$ any other MD estimators and let $B_{-1}(\hat{\theta}_{el})$ and $B_{-1}(\hat{\theta}_{md})$ denote the $O(n^{-1})$ bias of EL and MD respectively.

Theorem 7.5: *If the estimator $\hat{\theta}$ admits a $O_p(n^{-2})$ expansion, then*

$$\begin{aligned} & \mathcal{M}_{-2}(\hat{\theta}_{el} - B_{-1}(\hat{\theta}_{el})) - \mathcal{M}_{-2}(\tilde{\theta}_{md} - B_{-1}(\hat{\theta}_{md})) \\ = \; & \mathcal{M}_{-2}(\hat{\theta}_{el}) - \mathcal{M}_{-2}(\hat{\theta}_{md}) - B_{-1}(\hat{\theta}_{el})B_{-1}(\hat{\theta}_{el})' + B_{-1}(\hat{\theta}_{md})B_{-1}(\hat{\theta}_{md}) \end{aligned}$$

If MD estimators with $\gamma_3 = 2$ are considered, Theorem 7.5 yields that

$$\mathcal{M}_{-2}(\hat{\theta}_{el} - B_{-1}(\hat{\theta}_{el})) - \mathcal{M}_{-2}(\tilde{\theta}_{md} - B_{-1}(\hat{\theta}_{md})) = \mathcal{M}_{-2}(\hat{\theta}_{el}) - \mathcal{M}_{-2}(\hat{\theta}_{md})$$

since in this case $B_{-1}(\hat{\theta}_{el}) = B_{-1}(\hat{\theta}_{md})$. It follows that a bias corrected MD estimator with $\gamma_3 = 2$ has the same higher order efficiency of EL if the uncorrected estimator has the same $O(n^{-2})$ MSE of EL, that is when

$$\mathcal{M}_{-2}(\hat{\theta}_{el}) - \mathcal{M}_{-2}(\tilde{\theta}_{md}) = 0$$

Considering the difference in higher order MSE simplifies the calculations and allows to give a general results about efficiency. The following assumptions are needed to obtain a valid expansion of order $O(n^{-2})$.

Assumption 3: Assumption D: There is an $\epsilon > 0$ and $\mathcal{B}(w_i)$, $E[\mathcal{B}(w_i)^6] < \infty$ such that for any $j, r, s, h = 1, \ldots, k$: *i)* $\sup_{\theta \in \mathcal{S}(\theta_o, \epsilon)} \|q_i(\theta)\| \leq \mathcal{B}(w_i)$; *ii)* $q_i^{jrs}(\theta)$ exists on $\mathcal{S}(\theta_o, \epsilon)$; *iii)* $\sup_{\theta \in \mathcal{S}(\theta_o, \epsilon)} \|q_i^j(\theta) - q_o^j\| \leq \mathcal{B}(w_i)$; *iv)* $\sup_{\theta \in \mathcal{S}(\theta_o, \epsilon)} \|q_i^{jr}(\theta) - q_o^{jr}\| \leq \mathcal{B}(w_i)$; *v)* $\sup_{\theta \in \mathcal{S}(\theta_o, \epsilon)} \|q_i^{jrs}(\theta) - q_o^{jrs}\| \leq \mathcal{B}(w_i)$; *vi)* $\|q_i^{jrsh}(\theta) - q_o^{jrsh}\| \leq \mathcal{B}(w_i)\|\theta - \theta_o\|$ for any $\theta \in \mathcal{S}(\theta_o, \epsilon)$; *v)* $\gamma(\cdot)$ is five times continuously differentiable in a neighborhood of 1.

Theorem 7.6: *Suppose Assumption A-D hold. Then MD estimators and the associated Lagrange multipliers admit $O(n^{-2})$ expansion of the form*

$$(\hat{\theta} - \theta_o) = u_n + b_n^{\theta} + r_n^{\theta} + O_p(n^{-2})$$

Further, if $\bar{\theta}_{md}$ is an MD estimator with $\gamma_3 = 2$, then

$$\begin{aligned} \mathcal{M}_{-2}(\hat{\theta}_{el}) - \mathcal{M}_{-2}(\bar{\theta}_{md}) \; = \; & \left(1 - \frac{\bar{\gamma}_4}{6}\right) \sum_{j=1}^{m} \sum_{r=1}^{m} S_o q_{jr}^4(\theta_o) E\left[l_{n,j} l_{n,r} l_n u_n'\right] \\ & + \left(1 - \frac{\bar{\gamma}_4}{6}\right) \sum_{j=1}^{m} \sum_{r=1}^{m} E\left[l_{n,j} l_{n,r} u_n l_n'\right] S_o' q_{jr}^4(\theta_o)' \end{aligned}$$

where $q^4_{jr}(\theta_o) = E\left[q_{n,j}(\theta_o)q_{n,r}(\theta_o)q_i(\theta_o)q_i(\theta_o)'\right]$ *and* $\bar{\gamma}_4$ *is the fourth derivative evaluated at 1 of the divergence from which* $\bar{\theta}_{md}$ *is obtained.*

A consequence of Theorem 7.6 is that in general the $O(n^{-2})$ MSE of EL is different from that of other MD estimators unless the MD considered is obtained from a divergence with $\tilde{\gamma}_4 = 6$ (for EL, $\gamma_4 = 6$). In this EL and MD are equivalent up to $O(n^{-2})$ and they have the same asymptotic variance to that order. It turns out that MD estimators that are equivalent to EL up to order $n^{-3/2}$ also have the same higher order MSE.

Theorem 7.7: *Suppose Assumptions A-D hold. If* $\bar{\theta}_{md}$ *is an MD estimator with* $\gamma_3 = 2$, *then*

$$MSE(\hat{\theta}_{el}) - MSE(\bar{\theta}_{md}) = o(n^{-2})$$

This is a very interesting result that has two substantive implications: first, that all the members of this MD subclass with $\gamma_3 = 2$ are third order efficient after the bias is removed; second, that third order efficiency is an inadequate criterion for prescribing which specific estimator should be used in applied work. If one insists on considering estimators that have the same bias as EL estimators, then another criterion must supplement third order efficiency.

It can also be verified that the when $\gamma_3 = 2$ the estimator that solves the first order conditions (34)-(35) is third order efficient having the same $O(n^{-2})$ MSE than EL.

8 Robust Higher Order Efficient Estimators

In the previous section it was shown that any MD estimator with $\gamma_3 = 2$ has the same $O(n^{-2})$ MSE. This section discusses how third order efficiency may be complemented by another property that can imrpove the finite sample performance of MD estimators.

Recall from Section 5 that the Minimum Divergence problem cannot always be solved by Lagrange Multiplier methods. A requirement that was imposed is that there exist $\hat{\lambda} \in \mathbb{R}^m$ and $\hat{\theta} \in \Theta$ such that

$$\begin{aligned} \hat{\lambda}' q_i(\hat{\theta}) &\in \mathcal{A}, \, i = 1, \ldots, n \\ \mathcal{A} &= \{y : y = \gamma_1(x), x \in (a_\gamma, b_\gamma)\} \end{aligned}$$

and such that

$$\begin{aligned}\sum_{i=1}^{n}\tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))q_i(\hat{\theta}) &= 0\\ \sum_{i=1}^{n}\tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))\nabla_\theta q_i(\hat{\theta})'\hat{\lambda} &= 0\end{aligned}$$

Features of the set $\mathcal{A}$ have statistical implications. MD estimators that are defined from divergences that imply a $\mathcal{A} = \{y : -\infty < y < +\infty\}$ are robust relatively to MD estimators defined from divergences that imply a set A with at least a finite endpoint, because in this second case the Influence Function (IF) of the estimator can become unbounded even when $q(w, \theta)$ is bounded.

Hampel, Ronchetti, Rousseeuw, and Stahel (1986) show that the influence function of an estimator obtained as solution to the estimating equation

$$\sum_{i=1}^{n} s(z_i, \theta) = 0$$

for a specified function $s(\cdot)$ is proportional to the estimating equation, so that

$$IF(z, \theta, \lambda) = -E\left[\frac{\partial s(z, \theta)}{\partial \theta'}\right]^{-1} s(z, \theta)$$

Euristically, the Influence Function measures the asymptotic bias caused by fractional data contamination. It is known that an estimator θ whose influence function is unbounded may have an unbounded asymptotic bias under single point contamination.

When evaluated at the true parameter values $\theta = \theta_o$ and $\lambda = 0$, the IF of any MD estimator is then given by

$$IF(w, \theta_o, 0) = -\begin{bmatrix} B_o q_n(\theta_o)\\ P_o q_n(\theta_o)\end{bmatrix}$$

However, when evaluated at $\lambda = \varepsilon$, the IF of MD is proportional to the weights and hence can become unbounded if the weights are not defined for every value of λ. Hence, MD estimators with $\mathcal{A} = \{y : -\infty < y < +\infty\}$ are preferable because their influence function is less sensitive to deviations of λ from its asymptotic limit. For ET and CUE $\mathcal{A} = \{y : -\infty < y < +\infty\}$ and the IF will be bounded in λ, while for EL $\mathcal{A} = \{y : -\infty < y < -1\}$ and the IF is unbounded in λ. Considering MD estimators with bounded IF have two main advantages. First, the asymptotic expansions are polynomial in the influence function. If the IF can become unbounded for relatively small deviations of λ from its limit, the higher order ranking of estimators can be entirely misleading. Second, test statistics based on MD estimators with bounded influence function should have better size properties. This intuition is indeed confirmed by the finding in Imbens, Spady, and Johnson (1998). They show that test statistics based on ET tend to be superior to the same statistics based on the the third order efficient EL. Their findings support the view that the influence function could be important in determining the small sample behavior of estimators and related test statistics.

Another reason to consider MD estimators whose divergence implies $\mathcal{A} = \{y : -\infty < y < +\infty\}$ is related to the existence of the asymptotic variance of MD estimators in the presence of global misspecification. When the moment function is unbounded in w, that is $\inf_{\theta\in\Theta} \sup_w \|q(w,\theta)\| = +\infty$, Schennach (2003) shows that ET asymptotic behavior is robust to misspecification, while EL does not have finite asymptotic variance.

It is very interesting to consider estimators that combine the superior higher order behavior of EL with the properties of having a bounded IF in λ. Since we know from Theorem 7.7 that all MD estimators with $\gamma_3 = 2$ have the same higher order efficiency as EL, the task is to find a divergence $\gamma(\cdot)$ with $\gamma_3 = 2$ and such that $\mathcal{A} = \{y : -\infty < y < +\infty\}$. It is difficult to derive a divergence with a closed form solution that satisfies the above conditions. However, we can study the MD estimators obtained as solution of the dual problem

$$\max_{\theta\in\Theta} \min_{\lambda\in\Lambda_n(\theta)} \frac{1}{n}\sum_{i=1}^{n} \psi(\lambda' q_i(\theta)) \tag{36}$$

where $\psi(\cdot)$ satisfies Assumption 2(ψ) and $\Lambda_n(\theta) = \{\lambda : \lambda' q_i(\theta) \in \mathcal{V},\ i = 1, \ldots, n\}$. By Theorem 5.2, the estimator associated with problem (36) corresponds to a MD estimator defined from a strictly convex divergence. Constructing a MD estimator with bounded IF in λ is equivalent to find a strictly convex function satisfying Assumption 2(ψ) and such that the endpoints of $\mathcal{V} = \{y : y = \psi_1(x), x \in (a_\psi, b_\psi)\}$ are infinite.

Since ET has bounded IF, we consider modifying the ET objective function as in $\psi(v) = \exp(h(v))$, where $h(\cdot) : \mathbb{R} \to \mathbb{R}$ is three times continuously differentiable on $\mathbb{R}$ and such that $h(0) = 0$. The first derivative is given by $\psi_1(v) = \exp(h(v))h_1(v)$. In order to have bounded influence function the function $h(\cdot)$ must satisfy

$$\{y : y = h_1(x),\ x \in \mathbb{R}\} = \mathbb{R}$$

In order for the estimator defined in (36) to have the same bias of EL, it must hold that

$$\left.\frac{\partial^2 \exp(h(\upsilon))h_1(\upsilon)}{\partial^2 \upsilon}\right|_{\upsilon=0} = 2 \tag{37}$$

Assumption $2(\psi)$ also requires that $h_1(0) = 1$ and $h_2(0) = 1$. By expanding the derivative in (37), it follows that the function $h(\cdot)$ must solve the following local differential equation

$$3h''(0) + h'''(0) = 1$$

A function that satisfies the following restriction is given by

$$h(\upsilon) = \frac{1}{2}(e^{\upsilon} - e^{-\upsilon})$$

It is easy to see that $h(\upsilon)$ is continuously differentiable, $h(0) = 0$ and $h_1(0) = 1$, $h_2(0) = 0$ and $h_3(0) = 1$. The function $\frac{1}{2}(e^{\upsilon} - e^{-\upsilon})$ is usually referred to as hyperbolic sine and denoted as $\sinh(\upsilon)$.We name the problem in (36) with $\psi(x) = [\exp(\sinh(x)) - 1]$ as Hyperbolic Tilting (HT) by analogy with the Exponential Tilting from which it originates.

Definition 8.1 (Hyperbolic Tilting): *The Hyperbolic Tilting (HT) estimator is defined as the solution of the following problem*

$$\max_{\theta\in\Theta} \min_{\lambda} \frac{1}{n}\sum_{i=1}^{n}[\exp(\sinh(\lambda' q_i(\theta))) - 1] \tag{38}$$

Since the divergence that corresponds to $\psi(x) = [\exp(\sinh(x)) - 1]$ is such that $\gamma_3 = 2$, the HT estimator has the same $O(n^{-2})$ MSE of EL. Differently from EL, the HT estimator has bounded influence function.

Clearly, HT is not the only estimator that has bounded influence function and is third order efficient. Many other estimators could be given by solving the local differential equation above and making sure that the resulting function has derivative with unbounded domain. We focus on the HT because from Imbens, Spady, and Johnson (1998), we know that ET has nice finite sample properties in terms of size of resulting statistics for testing overidentified restrictions and restrictions on the paramaters. The choice between EL and HT should be based on the comparison of the actual asymptotic performance of estimators and test statistics and their performance in simulations, which is what the remainder of this paper is devoted to.

9 Numerical Examples

This section provides some simulation evidence on the performance of MD estimators. To explore the perfomance of the estimators and their associated test statistics we consider the experimental design of Hall and Horowitz (1996). This experiment has also been considered by other authors. Imbens, Spady, and Johnson (1998) explore the performances of statistics designed to test the null hypothesis of overidentification. They consider only the ET estimator and the CUE. Kitamura (2001) considers the power of the likelihood ratio test based on EL.

In each Monte Carlo replication a vector (w_1, w_2) is drawn from a bivariate normal distribution with zero correlation coefficient, both means equal to zero and variances equal to 0.16. Given these random variables, we consider the following moment conditions

$$E[r(w_1, w_2, \theta)(1, w_2)'] = 0 \tag{39}$$

where

$$r(w_1, w_2, \theta) = \exp(-0.72 - \theta(w_1 + w_2) + 3w_2) - 1$$

Given the $m = 2$ restrictions in (39), the objective is to estimate the $k = 1$ parameter θ. The true value of the parameter is $\theta_o = 3$, since

$$E[\exp(-0.72 - 3w_1) - 1] = 0$$

and

$$E[\exp(-0.72 - 3w_1)w_2 - w_2] = 0$$

We do not consider more complicated models to avoid computational problems that may arise from the nonlinear nature of the problem considered. For the same reason, the CUE is not considered because it shows a very erratic behavior, as already reported by Kitamura (2001) and Imbens, Spady, and Johnson (1998), which could bias the ranking of estimators and test statistics.

9.1 Estimators

Table 1 reports the basic statistics of the estimated parameters for sample sizes $n = 50, 100, 200$. The statistics reported are the mean, the median, the square-root Mean Square Error (RMSE), the Mean Absolute Error (MAE) and the Interquartile Range

(IQR). The MD estimators considered are EL, ET and HT. For comparison purposes, we also report the GMM estimation results.

For both sample sizes considered in the experiment, the GMM estimator has larger bias and larger RMSE than all the three MD estimators considered. GMM also performs worse than ET, EL and HT in terms of MAE and IQR. The performance of the MD estimators is very similar for $n = 200$. In contrast, at the smaller sample size EL is the preferred estimator based on the statistics given in Table 1. As predicted in Section 7, the bias and the RMSE of HT are very close to that of EL, while ET shows larger bias and larger RMSE.

Table 1 consistent with the prediction that differences between MD estimators are of smaller order than the difference between GMM and MD. Although only this simple experiment is reported, these qualitative findings are robust to other settings.

9.2 Overidentification Tests

In this section we analyze the performance of statistics based on MD estimators for testing the null hypothesis of overidentification. First, we consider the criterion based statistics for EL, ET and HT whose expressions are given, respectively, by

$$\begin{aligned}
l_{el} &= -2n\sum_{i=1}^{n}\log[1-\hat{\lambda}' q_i(\hat{\theta})] \\
l_{et} &= 2n\left(\frac{1}{n}\sum_{i=1}^{n}\exp[\hat{\lambda}' q_i(\hat{\theta})]-1\right) \\
l_{ht} &= 2n\left(\frac{1}{n}\sum_{i=1}^{n}\exp[\sinh(\hat{\lambda}' q_i(\hat{\theta}))]-1\right)
\end{aligned}$$

Second, we examine the performance of two LM statistics for each MD estimator considered. Both these statistics use the estimated Lagrange multiplier, but they differ for the choice of the variance. LM_u^R, $u = el, et, ht$ denotes the LM statistic constructed using the robust variance estimate discussed in Section 5

$$\hat{V}_R = \sum_{i=1}^{n}\hat{\omega}_i q_i(\hat{\theta})q_i(\hat{\theta})'\left(n\sum_{i=1}^{n}\hat{\omega}_i^2 q_i(\hat{\theta})q_i(\hat{\theta})'\right)^{-1}\sum_{i=1}^{n}\hat{\omega}_i q_i(\hat{\theta})q_i(\hat{\theta})' \qquad (40)$$

that uses the estimated probabilities obtained from EL, ET and HT. LM_u^{ω} $u = el, et, ht$ denotes the statistics constructed using the simpler variance estimate

$$\hat{V}_{\omega} = \sum_{i=1}^{n} \hat{\omega}_i q_i(\hat{\theta}) q_i(\hat{\theta})'$$

Finally we consider average moment tests: the J-test based on the GMM estimator $\hat{\theta}_{gmm}$

$$J_{gmm} = n \cdot q_n(\hat{\theta}_{gmm})' \left[\frac{1}{n} \sum_{i=1}^{n} q_i(\hat{\theta}_{gmm}) q_i(\hat{\theta}_{gmm})' \right]^{-1} q_n(\hat{\theta}_{gmm})$$

and its modified versions

$$J_u^R = n \cdot q_n(\hat{\theta}_{md}) \widehat{V}_R(\hat{\theta}_{md})^{-1} q_n(\hat{\theta}_{md})$$

where $\widehat{V}_R(\hat{\theta}_{md})$ is the robust variance in (40) based on MD estimators ($u = el, et, ht$).

Results of Monte Carlo simulations are given in Table 2. In general, the LM multiplier version of the test is superior to both the LR type tests and average moment tests. The LR type tests have better size than either J_{gmm} and its modified versions that use the MD estimators. The ranking of the different LM type tests depends on both the sample size and the nominal size (α) considered. The empirical size of EL based LM statistics tends to be closer to the nominal size for large value of α. When $n = 100$, the LM_{el}^{ω} is superior to all the other form of the tests whenever $\alpha \leq 0.15$. When $\alpha > 0.15$, the distribution of LM_{ht}^{R} shows very close agreement with the reference distribution. When $n = 200$, the other version of the EL based LM test is superior to the other tests for $\alpha > 0.2$, but similarly to the $n = 100$ case, the LM_{ht}^{R} test is again superior when $\alpha < 0.2$. Imbens, Spady, and Johnson (1998) also consider an average moment test based on the CUE. In their simulations, this test shows very good properties. A comparison of Table 2 with their Table II, though, shows that the Lagrange multiplier test based on the Hyperbolic Tilting has also better size than the CUE based test.

10 Conclusion

The first important result of this paper is to provide a unifying framework for the several techniques that recently have been proposed in the literature as alternatives to the traditional Generalized Method of Moments. Different techniques are presented and compared as members of a general class of Minimum Divergence (MD) estimators, and an interesting relationship between this class and the generalized empirical likelihood (GEL) class of Newey and Smith (2004) is found: every MD estimator has a GEL representation and, conversely, every GEL estimator can be represented as an MD estimator. Previously, this relationship was only known for the Cressie-Read family of divergences, whereas here it is shown to hold for every strictly convex divergence.

The properties of the members of the MD class are analyzed with the goal of identifying an optimality criterion that guides the choice of an estimator from this class. While consistency and first-order efficiency are traditionally proposed as optimality criteria, they are not sufficient when comparing estimators that share the same first-order asymptotic variance, as do the GMM and MD estimators. A key finding of this paper is that all MD estimators sharing the asymptotic bias of Empirical Likelihood (EL) have the same higher order Mean Square Error, implying that even third order efficiency is not a compelling criterion.

I propose the boundedness of the influence function of the MD estimator as the criterion for selecting an estimator from the class of third order efficient estimators. Despite desirable higher order properties, EL does not have a bounded influence function. Monte Carlo simulations show that considering MD estimators that are both third order efficient and have a bounded influence function delivers statistics that have strikingly good size control for testing hypotheses. In particular, Lagrange Multiplier test based on the newly proposed Hyperbolic Tilting estimator, perform much better than the standard tests and test based both on EL, ET and CUE.

Finally, this paper suggests a couple of issues that would benefit from future research. From the present analysis, one may question what other possible characteristics may be useful in distinguishing between divergences: in particular, what might we learn from the direct analysis of the Edgeworth expansions of the test statistics? In addition, can the Bayesian perspective on weighting schemes (implied by different choices of divergences) give more guidance on selecting a "good" MD estimator for some particular or general applications? The framework built here is hopefully a useful basis from which to research these and other topics.

References

Ahn, S. C., and P. Schmidt (1995): "A separability result for gmm estimation, with applications to gls prediction and conditional moment tests," *Econometric Reviews*, 14(1), 19–34.

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

Bates, C. E., and H. White (1993): "Determination of estimators with minimum asymptotic covariance matrices," *Econometric Theory*, 9, 633–648.

Bekker, P. A. (1994): "Alternative approximations to the distributions of instrumental variable estimators," *Econometrica*, 62(3), 657–81.

Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34(3), 305–34.

Corcoran, S. (1998): "Bartlett adjustment of empirical discrepancy statistics," *Biometrika*, 85, 967–972.

Cressie, N., and T. Read (1984): "Multinomial goodness-of-fit tests," *J. R. Statist. Soc. B*, 46(3), 440–464.

Csiszar, I. (1975): "I-divergence geometry of probability distribution and minimization problems," *The Annals of Probability*, 3(1), 146–158.

Elliott, G., I. Kumunjer, and A. Timmermann (2002): "Estimating loss function parameters," mimeo, Univerity of California, San Diego.

Gallant, R. A., and H. White (1988): *a unified theory of estimation and inference for nonlinear dynamic models*. Basil Blackwell, New York.

Hall, P., and J. L. Horowitz (1996): "Bootstrap critical values for tests based on generalized-method-of-moments estimators," *Econometrica*, 64(4), 891–916.

Hampel, F., E. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986): *Robust Statistics: the Approach based on Influence Functions*. Wiley.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50(4), 1029–54.

HANSEN, L. P., J. HEATON, AND A. YARON (1996): "Finite-sample properties of some alternative gmm estimators," *Journal of Business and Economic Statistics*, 14(3), 262–80.

HANSEN, L. P., AND K. J. SINGLETON (1982): "Generalized instrumental variables estimation of nonlinear rational expectations models," *Econometrica*, 50(5), 1269–86.

HAUSMAN, JERRY A. AND HAHN, J. (2001): "Bias Corrected Instrumental Variables Estimation for Dynamic Panel Models with Fixed Effects," .

HAUSMAN, J. A., J. HAHN, AND G. KUERSTEINER (2001): "Higher Order MSE of Jacknife 2SLS," .

HAYASHI, F. (2000): *Econometrics*. Princeton University Press.

IMBENS, G. W. (1997): "One-step estimators for over-identified generalized method of moments models," *Review of Economic Studies*, 64(3), 359–83.

——— (2002): "Generalized method of moments and empirical likelihood," *Journal of Business and Economic Statistics*, 20(4), 493–506.

IMBENS, G. W., R. H. SPADY, AND P. JOHNSON (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66(2), 333–57.

JAYNES, E. (1957): "Information Theory and Statistical Mechanics," *Phis. Rev.*, 106, 620–30.

——— (1968): "Prior Probability," *IEEE Transactions on System, Science and Cybernetics*, SSC-4, 227–41.

JUDGE, G., AND R. MITTELHAMMER (2001): "Empirical Evidence concerning the finite sample performance of EL-type structural equation estimators," .

KIM, J. Y. (2002): "Limited Information Likelihood adn Bayesian Analysis," *Journal of Econometrics*, 107, 175–193.

KITAMURA, Y. (2001): "Asymptotic optimality of empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661–1672.

KITAMURA, Y., AND M. STUTZER (1997): "An information-theoretic alternative to generalized method of moments estimation," *Econometrica*, 65(4), 861–74.

Kleibergen, F. (2002): "Pivotal statistics for testing structural parameters in instrumental variables regression," *Econometrica*, 70(5), 1781–1803.

Kunitomo, N., and Y. Matsushita (2003): "Finite Sample Distributions of the Empirical Likelihood Estimator and the GMM Estimator," University of Tokyo.

Lazar, N. (2003): "Bayesian Empirical Likelihood," *Biometrika*, 90, 319–26.

Mann, H. B., and A. Wald (1943): "On the Statistical Treatment of Linear Stochastic Equations," *Econometrica*, 11, 173–220.

Moreira, M. J. (2003): "A conditional likelihood ratio test for structural models," *Econometrica*, 71(4), 1027–48.

Nagar, A. L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27(4), 575–595.

Newey, W. K., and D. McFadden (1994): "estimation and inference in large samples," in *handbook of econometrics*, ed. by R. Engle, and D. McFadden, pp. 2113–2245, Amsterdam. North-Holland.

Newey, W. K., and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72(1), 219–55.

Nickell, S. J. (1981): "Biases in dynamic models with fixed effects," *Econometrica*, 49(6), 1417–26.

Owen, A. (1990): "Empirical Likelihood Ratio Confidence Intervals," *The Annals of Statistics*, 18(1).

Pfanzagl, J., and W. Wefelmeyer (1979): "A third-order optimum property of the maximum likelihood estimator," *Journal of Multivariate Analysis*, 8, 1–29.

Potscher, B. M., and I. R. Prucha (1997): *Dynamic nonlinear econometric models.* Springer, New York.

Qin, J., and J. Lawless (1994): "Empirical likelihood and general estimating equations," *Annals of Statistics*, 22, 300–325.

Ragusa, G. (2004): "Testing for Dynamic Stability of Misspecified Moment Conditions Models," mimeo.

Rilstone, P., V. K. Srivastava, and A. Ullah (1996): “The Second-Order Bias and Mean Squared Error of Nonlinear Estimators,” *Journal of Econometrics*, 75(2), 369–95.

Schennach, S. C. (2003): “Exponentially Tilted Empirical Likelihood,” .

——— (2005): “Bayesian Exponentially-Tilted Empirical Likelihood,” *Biometrika*, 92(1).

Srinivasan, T. N. (1970): “Approximations to finite sample moments of estimators whose exact sampling distributions are unknown,” *Econometrica*, 38(3), 533–41.

Staiger, D., and J. H. Stock (1997): “Instrumental variables regression with weak instruments,” *Econometrica*, 65(3), 557–86.

Table 1: Properties of GMM and MD Estimators, 5,000 Replications

Estimator	*Mean*	*Median*	*RMSE*	*MAE*	*IQR*
		$k = 1$, $m = 2$, $n = 200$, $\theta_o = 3$			
EL	3.032	3.038	0.204	0.161	0.270
ET	3.039	3.048	0.207	0.163	0.273
HT	3.033	3.039	0.204	0.161	0.271
GMM	2.963	3.009	0.523	0.235	0.281
		$k = 1$, $m = 2$, $n = 100$, $\theta_o = 3$			
EL	3.068	3.040	0.302	0.234	0.392
ET	3.087	3.054	0.313	0.241	0.400
HT	3.072	3.043	0.304	0.237	0.397
GMM	2.966	2.967	0.763	0.379	0.4251
		$k = 1$, $m = 2$, $n = 50$, $\theta_o = 3$			
EL	3.119	3.078	0.460	0.340	0.550
ET	3.154	3.104	0.514	0.366	0.565
HT	3.128	3.083	0.474	0.350	0.567
GMM	2.844	3.021	1.019	0.558	0.594

Table 2: Size of Over-Idenfication Tests, 5,000 Replications, $m = 2$, $k = 1$

Nominal Size	l_{EL}	l_{ET}	l_{HT}	LM_1^{el}	LM_R^{el}	LM_1^{et}	LM_R^{et}	LM_1^{ht}	LM_R^{ht}	J_R^{el}	J_R^{et}	J_R^{ht}	J_{gmm}
$n = 50$													
0.700	0.424	0.415	0.421	0.427	0.402	0.382	0.414	0.384	0.406	0.425	0.428	0.428	0.463
0.750	0.374	0.364	0.371	0.383	0.353	0.336	0.369	0.329	0.352	0.380	0.385	0.387	0.421
0.800	0.328	0.316	0.321	0.334	0.306	0.284	0.317	0.275	0.294	0.334	0.341	0.344	0.371
0.850	0.280	0.265	0.268	0.284	0.255	0.239	0.254	0.223	0.219	0.287	0.297	0.303	0.327
0.900	0.218	0.203	0.203	0.226	0.200	0.191	0.181	0.164	0.147	0.241	0.253	0.262	0.276
0.950	0.146	0.141	0.135	0.155	0.149	0.133	0.104	0.109	0.074	0.181	0.191	0.203	0.210
0.975	0.106	0.100	0.095	0.114	0.110	0.097	0.064	0.079	0.038	0.143	0.156	0.167	0.173
0.990	0.066	0.068	0.061	0.076	0.081	0.072	0.035	0.052	0.015	0.111	0.124	0.138	0.141
$n = 100$													
0.700	0.387	0.383	0.387	0.390	0.358	0.359	0.383	0.368	0.371	0.388	0.392	0.392	0.410
0.750	0.341	0.333	0.340	0.347	0.306	0.307	0.335	0.317	0.319	0.342	0.348	0.347	0.360
0.800	0.292	0.282	0.289	0.296	0.260	0.260	0.280	0.261	0.264	0.293	0.299	0.302	0.309
0.850	0.233	0.230	0.227	0.238	0.213	0.215	0.218	0.198	0.200	0.251	0.256	0.261	0.264
0.900	0.180	0.176	0.173	0.180	0.168	0.166	0.153	0.146	0.134	0.201	0.208	0.216	0.214
0.950	0.118	0.117	0.112	0.122	0.109	0.118	0.088	0.091	0.064	0.149	0.157	0.168	0.163
0.975	0.080	0.080	0.074	0.081	0.081	0.086	0.053	0.063	0.032	0.114	0.122	0.132	0.129
0.990	0.045	0.051	0.045	0.051	0.058	0.057	0.026	0.040	0.012	0.085	0.091	0.101	0.102
$n = 200$													
0.700	0.358	0.351	0.358	0.357	0.330	0.338	0.352	0.342	0.340	0.360	0.363	0.363	0.373
0.750	0.306	0.303	0.304	0.310	0.278	0.290	0.300	0.287	0.283	0.315	0.319	0.319	0.321
0.800	0.260	0.252	0.260	0.256	0.227	0.237	0.245	0.236	0.233	0.264	0.270	0.270	0.271
0.850	0.203	0.201	0.200	0.205	0.180	0.190	0.189	0.179	0.169	0.218	0.227	0.228	0.221
0.900	0.153	0.148	0.147	0.154	0.132	0.143	0.129	0.128	0.110	0.165	0.176	0.182	0.169
0.950	0.086	0.089	0.083	0.092	0.092	0.091	0.068	0.065	0.057	0.115	0.123	0.132	0.116
0.975	0.055	0.054	0.052	0.059	0.066	0.061	0.035	0.042	0.027	0.082	0.086	0.097	0.084
0.990	0.028	0.035	0.029	0.033	0.045	0.040	0.017	0.023	0.010	0.056	0.058	0.066	0.063

A Mathematical Appendix

Proof to Theorem 5.1

Since, by assumption, the solutions of the MD problem are interior we have that $\hat{\lambda} q(w, \hat{\theta}) \in \mathcal{A}$, where

$$\mathcal{A} = \{y : y = \gamma_1(x), x \in (a_\gamma, b_\gamma)$$

and hence $\hat{\pi}_i = \tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta}))/n$, and if $q_i(\theta)$ is differentiable in θ, $\hat{\lambda}$ and $\hat{\theta}$ solves the following first order conditions

$$\sum_{i=1}^{n} \hat{\pi}_i q_i(\hat{\theta}) = 0; \quad \sum_{i=1}^{n} \hat{\pi}_i \nabla_\theta q_i(\hat{\theta}) = 0$$

Consider the following GEL problem

$$\max_{\theta} \left[\min_{\lambda \in \Lambda_n(\theta)} \frac{1}{n} \sum_{i=1}^{n} (\lambda' q_i(\theta)) \tilde{\gamma}_1(\lambda' q_i(\theta)) - \gamma(\tilde{\gamma}_1(\lambda' q_i(\theta))) \right]$$

where $\Lambda_n(\theta) = \{\lambda | \lambda' q_i(\theta) \in \mathcal{A}, i = 1, ..., n\}$. First of all, notice that

$$\psi(x) = x\tilde{\gamma}_1(x) - \gamma(\tilde{\gamma}_1(x))$$

is well defined on $\mathcal{A}$. By the Inverse Function Theorem and strict convexity of $\gamma(\cdot)$ on (a_γ, b_γ),

$$\frac{\partial \psi_2(x)}{\partial x} = \frac{1}{\gamma_2(\tilde{\gamma}_1(x))} > 0$$

for $x \in \mathcal{A}$, and hence $\psi(\cdot)$ is strictly convex on $x \in \mathcal{A}$. The first order conditions for λ are given by

$$\frac{1}{n} \sum_{i=1}^{n} \tilde{\gamma}_1(\lambda' q_i(\theta)) q_i(\hat{\theta}) = 0 \tag{41}$$

that corresponds the first order conditions for λ in the MD problem. Since

$$\sum_{i=1}^{n} q_i(\hat{\theta}) q_i(\hat{\theta}) \tilde{\gamma}_2(\hat{\lambda}' q_i(\hat{\theta}))$$

is non singular by assumption then there is a neighborhood of $\hat{\theta}$ where $\hat{\lambda}(\theta)$ that solves (41) exists and it is continuously differentiable in a neighborhood of θ. By the envelope theorem, the first order conditions for θ are then given by

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_1(\lambda' q(w_i,\theta))\nabla_\theta q_i(\theta)'\lambda = 0$$

and the result follows.

Proof to Theorem 5.2

Let $\mathcal{V} = \{y : y = \psi_1(x), x \in (a_\psi, b_\psi)\}$ and consider the function

$$\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$$

that is well defined on $\mathcal{V}$. Strict convexity follows as in the proof of Theorem 5.1 by noting that strict convexity of $\psi(\cdot)$ implies

$$\frac{\partial\gamma_2(x)}{\partial x} = \frac{1}{\psi_2(\tilde{\psi}_1(x))} > 0$$

By Assumption, there exist $\hat{\lambda} \in \mathbb{R}^m$ and $\theta \in \Theta$ such that $\{\hat{\lambda}' q_i(\hat{\theta}) \in \mathcal{V}, i = 1, ..., n\}$ and satisfy the first order conditions

$$\begin{aligned}\sum_{i=1}^{n}\psi_1(\hat{\lambda}' q(\hat{\theta}))q(\hat{\theta}) &= 0\\ \sum_{i=1}^{n}\psi_1(\hat{\lambda}' q(\hat{\theta}))\nabla_\theta q(\hat{\theta})'\hat{\lambda} &= 0\end{aligned}$$

The MD problem

$$\begin{aligned}\min_{\pi,\theta}\frac{1}{n}\sum_{i=1}^{n}&\left\{n\pi_i\tilde{\psi}_1(n\pi_i) - \psi(\tilde{\psi}_1(n\pi_i))\right\}\\ s.t.\ \sum_{i=1}^{n}&\pi_i q_i(\theta) = 0;\quad \sum_{i=1}^{n}\pi_i = 1\end{aligned}$$

has first order conditions given by

$$\begin{aligned}\tilde{\psi}_1(n\pi_i) &= \eta + \lambda' q_i(\theta)\\ \sum_{i=1}^{n}\pi_i\nabla_\theta q_i(\theta) &= 0\end{aligned}$$

Notice that $\gamma_1 = \tilde{\psi}_1(1) = 0$, since by assumption $\psi_1(0) = 1$ and $\gamma_2 = \tilde{\psi}_2(1) = 1$ since

$$\gamma_2 = \tilde{\psi}_2(1) = \frac{1}{\psi_2(\tilde{\psi}_1(1))} = \frac{1}{\psi_2(0)} = 1$$

Setting the Lagrange multiplier $\eta = 0$, by hypothesis there exists $\hat{\lambda} \in \mathbb{R}^m$ and $\theta \in \Theta$ such that $\lambda' q_i(\theta) \in \mathcal{V}$, $i = 1, ..., n$ and hence by inverting $\tilde{\psi}_1(\cdot)$

$$\pi_i = \frac{1}{n}\psi_1(\lambda' q_i(\theta))$$

and the conclusion follows.

A.1 Proof to Corollary 5.1

The reults follows from the definition of $\psi(\cdot)$,

$$\psi(\hat{\lambda}' q_i(\hat{\theta})) = \hat{\lambda}' q_i(\hat{\theta})\tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta})) - \gamma(\tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta})))$$

Summing over $i = 1, 2, ..., n$ and using $\sum_i^n \tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta}))q_i(\hat{\theta}) = 0$ yields

$$\frac{1}{n}\sum_{i=1}^{n} \psi(\hat{\lambda}' q_i(\hat{\theta})) = -\frac{1}{n}\sum_{i=1}^{n} \gamma(\tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta})))$$

as required.

Proof to Theorem 5.3

The consistency follows from using the duality results established in the paper and using the result of NS for GEL. To relax their assumption $E[\sup_{\theta\in\Theta} \|q(w, \theta)\|^{\alpha}] = 0$, $\alpha > 2$ required by their Lemma A1, it is sufficient to show that Assumption A.(iv) implies $\max_{\theta\in\Theta, i\leq n} \|q(w_i, \theta)\| = o_p(n^{1/2})$. This can be shown along the lines of Owen (1990). Since $\sup_{\theta\in\Theta} E(\|q(w, \theta)\|^2) \leq \infty$ for $\theta \in \Theta$ implies that $\sum_{i=1}^{\infty} P(\sup_{\theta\in\Theta} \|q(w_i, \theta)\|^2 > n) < \infty$ and hence that $\sum_{i=1}^{\infty} P(\sup_{\theta\in\Theta} \|q(w_i, \theta)\| > n^{1/2}) < \infty$. By applications of the Borel Cantelli Lemma, $\left\{\sup_{\theta\in\Theta} \|q(w_n, \theta)\| > n^{1/2}\right\}$ finitely often with probability 1. Also, for any $A > 0$, $\left\{\sup_{\theta\in\Theta} \|q(w_i, \theta)\| > An^{1/2}\right\}$ finitely often and hence

$$\lim_{n\to\infty} \{ \sup_{1\leq i\leq n} \sup_{\theta\in\Theta} \|q(w_i, \theta)\| \} n^{-1/2} \leq A$$

holds with probability 1. The probability 1 applies simultaneously over any countable set of values A so

$$\sup_{1\leq i\leq n}\sup_{\theta\in\Theta}\|q(w_i,\theta)\| = o(n^{1/2})$$

Let $\tilde{\Lambda}_n(\theta)=\{\lambda : \|\lambda\| < n^{-\zeta}\}$. By Cauchy-Swartz

$$\sup_{\theta\in\Theta,\lambda\in\tilde{\Lambda}_n,1\leq i\leq n}|\lambda' q_i(\theta)| \leq n^{-\zeta}\sup_{\theta\in\Theta,1\leq i\leq n}\|q_i(\theta)\| = o_p(n^{-\zeta+1/2})$$

and $\sup_{\theta\in\Theta,\lambda\in\tilde{\Lambda}_n,1\leq i\leq n}|\lambda' q_i(\theta)| = o_p(1)$ for $\zeta \geq 1/2$. Theorem 3.1 of NS then holds by replacing their Assumption 1.(d) with $E[\sup_{\theta\in\Theta}\|q(w,\theta)\|^{\alpha}]$, $\alpha = 2$.

Proof to Theorem 5.4

The first order conditions are

$$\begin{aligned}\sum_{i=1}^{n}\tilde{\gamma}_1(\gamma_1+\hat{\lambda}' q(w_i,\hat{\theta}))q(w_i,\hat{\theta}) &= 0\\ \sum_{i=1}^{n}\tilde{\gamma}_1(\gamma_1+\hat{\lambda}' q(w_i,\hat{\theta}))\nabla_\theta q(w_i,\hat{\theta})'\lambda &= 0\end{aligned}$$

Applying a law of large number for stationary and ergodic sequences we have that

$$\hat{\Gamma}(\theta_o) \equiv n^{-1}\sum_{i}^{n}\nabla_\theta q_i(\theta_o)\xrightarrow{p} -\Gamma_o;\ \hat{V}(\theta_o)\equiv n^{-1}\sum_{i}^{n}q_i(\theta_o)q_i(\theta_o)'\xrightarrow{p} V_o$$

Using $\max_i|\lambda' q(w_i,\theta)|\xrightarrow{p} 0$, expanding around $\hat{\lambda}=0$ and $\hat{\theta}=\theta_o$ as in Newey and Smith (2004) and noting that from the normalizations imposed on γ it follows that $d\tilde{\gamma}_1(x)/dx|_{x=\gamma_1}=1$, we have

$$\sqrt{n}\begin{pmatrix}\hat{\theta}-\theta_o\\ \hat{\lambda}\end{pmatrix} = \sqrt{n}\begin{pmatrix}-S_o & B_o\\ B_o' & P_o\end{pmatrix}\begin{pmatrix}0\\ -\frac{1}{n}\sum_{i=1}^{n}q(\theta_o)\end{pmatrix}+o_p(1)$$

where S_o, B_o and P_o are given by $S_o=(\Gamma_o' V_o^{-1}\Gamma_o)^{-1}$, $P_o=(V_o^{-1}-V_o^{-1}\Gamma_o S_o\Gamma_o' V_o^{-1})$ and $B_o=S_o\Gamma_o' V_o^{-1}$. The conclusion follows by application of the CLT for stationary ergodic sequences.

Proof to Theorem 5.5

Imbens (1997) gives a proof for the EL case. Here we extend the result to the MD class of estimators. Let $\hat{\omega}_i(\hat{\lambda},\hat{\theta}) = \tilde{\gamma}_1(\hat{\lambda}' q(w_i,\hat{\theta}))/\sum_i^n \tilde{\gamma}_1(\hat{\lambda}' q(w_i,\hat{\theta}))$. Taylor expansion of $\hat{\omega}_i(\hat{\lambda},\hat{\theta})$ around $\hat{\lambda}=0$ and using $\max_{i\leq n} |\hat{\lambda}' q(w_i,\hat{\theta})| \xrightarrow{p} 0$ gives, after some manipulation

$$\begin{aligned}
\hat{\omega}_i(\hat{\lambda},\hat{\theta}) &= \frac{1}{n} + \frac{1}{n}(1-\gamma_3/2)\hat{\lambda}' q_i(\hat{\theta}) q_i(\hat{\theta})'\hat{\lambda} + o_p(n^{-1}) \\
&= \frac{1}{n} + o_p(n^{-1})
\end{aligned}$$

Then, $\mathbb{M}_n(w) = \mathbb{Q}_n(w) + o_p(1)$. Thus, by the Glivenko-Cantelli, we have, pointwise, that

$$\mathbb{M}_n(w) - Q_o(w) \xrightarrow{p} 0$$

Let $\beta = (\mathbb{F}(w),\hat{\theta})'$, $f_i(\beta) = (\mathbb{M}_i(w) - Q_o(w), q_i(\theta)')'$ and $\nabla_\beta f_i(\beta) = \partial f_i/\partial\beta$. The variance of β under the model (1) is given by

$$E\left(\nabla_\beta f_i(\beta_o)\right)' E\left(f_i(\beta_o) f_i(\beta_o)'\right)^{-1} E\left(\nabla_\beta f_i(\beta_o)\right)$$

Then, for $\tilde{Q}_o = Q_o(w)(1-Q_o(w))$ and $q_i^w(\theta) = q_i(\theta)\mathbb{1}_{\{w_i\leq w\}}$ and $q_o(w) = \int q_i(\theta_o)\mathbb{1}_{\{w_i\leq w\}} \mathrm{d}Q_o$, we have

$$E\left(f_i(\beta_o) f_i(\beta_o)'\right)^{-1} = \begin{pmatrix} \triangle_o(w) & \Pi_o(w) \\ \Pi_o(w) & \Xi(w) \end{pmatrix}$$

where $\Delta_o(w) = \left(\tilde{Q}_o - q_o(w) V_o^{-1} q_o(w)\right)^{-1}$, $\Xi_o = V_o^{-1} - \Delta(w) V_o^{-1} q_o(w) q_o(w)' V_o^{-1}$ and $\Pi = -\Delta_o(w) q_o(w)$. The variance of $\mathbb{M}_n(w)$ is then given by $\Delta_o(w)^{-1} = \tilde{Q}_o - q_o(w) V_o^{-1} q_o(w)$.

Proof to Theorem 5.6

Consider first the case in which the null hypothesis concerns a subvector $d\times 1$ of θ_o, say $\theta_{1,o}$. Under the null hypothesis $H_o : \theta_1 = \theta_{1,o}$, the unrestricted subvector, $\hat{\theta}_2$ and $\hat{\lambda}_r$ solve the following first order conditions

$$\begin{aligned}
\sum_{i=1}^n \tilde{\gamma}_1(\hat{\lambda}_r' q(w_i,\hat{\theta})) q(w_i,\hat{\theta}) &= 0 \\
\sum_{i=1}^n \tilde{\gamma}_1(\hat{\lambda}_r' q(w_i,\hat{\theta})) \nabla_\theta q(w_i,\hat{\theta})'\hat{\lambda}_r &= 0
\end{aligned}$$

where here $\hat{\theta} = (\theta_{1,o}, \hat{\theta}_2)$. By an expansion, as in the proof of Theorem 4.4, we have that

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_2 \\ \hat{\lambda}_r \end{pmatrix} = \sqrt{n} \begin{pmatrix} -\tilde{S}_o & \tilde{B}_o \\ \tilde{B}_o' & \tilde{P}_o \end{pmatrix} \begin{pmatrix} 0 \\ \frac{1}{n} \sum_i^n q(w_i, \theta_o) \end{pmatrix} + o_p(1)$$

where $\tilde{S}_o = (\tilde{\Gamma}_o' V_o^{-1} \tilde{\Gamma}_o)^{-1}$, $\tilde{P}_o = (V_o^{-1} - V_o^{-1} \tilde{\Gamma}_o S_o \tilde{\Gamma}_o' V_o^{-1})$ and $\tilde{B}_o = S_o \tilde{\Gamma}_o' V_o^{-1}$ for $\tilde{\Gamma}_o = E[\nabla_{\theta_r} q(w_i, \theta_o)]$. By the CLT, it follows that under the null hypothesis $\hat{\lambda}_r \sim \mathrm{N}(0, \tilde{P}_o)$ where $\tilde{P}_o$ is idempotent with rank d. Then the statistic $n\hat{\lambda}_r' \widehat{\tilde{P}}^{-g} \hat{\lambda}_r$, where $\widehat{\tilde{P}} \xrightarrow{p} \tilde{P}_o$, is asymptotically distributed as χ^2 with d degrees of freedom. The second assertion follows from observing that $\sqrt{n}\hat{\lambda}_r = \tilde{P}_o \sum_{i=1}^n q(w_i, \theta_o)/\sqrt{n} + o_p(1)$ and that $\tilde{P}_o V_o^{-1} \tilde{P}_o = \tilde{P}_o \tilde{P}_o^- \tilde{P}_o = V_o^{-1}$. Hence,

$$n\hat{\lambda}_r' \tilde{P}_o^- \hat{\lambda}_r = \left(\sum_{i=1}^n q(w_i, \theta_o)/\sqrt{n} \right) \tilde{P}_o \tilde{P}_o^- \tilde{P}_o \left(\sum_{i=1}^n q(w_i, \theta_o)/\sqrt{n} \right) + o_p(1) \sim \chi_d^2$$

The claim then is proved by noting that the result holds with V_o^{-1} replaced by a consistent estimator $\hat{V}^{-1}$.

A.2 Proof to Theorem 5.7

By Assumption 1(γ), $\gamma_1 = 0$ and $\gamma_2 = 1$. By Taylor expansion of

$$LR_\omega = -2\frac{1}{n} \sum_{i=1}^n \left[\gamma(n\tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta}))) - \gamma(n\tilde{\gamma}_1(\hat{\lambda}_r' q_i(\hat{\theta}_r))) \right]$$

gives, buy using the fact that at the solutions $\gamma(\tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta}))) = -\psi(\hat{\lambda}' q_i(\hat{\theta}))$

$$-2 \sum_{i=1}^n \left[\gamma(\tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta}))) - \gamma(\tilde{\gamma}_1(\hat{\lambda}_r' q_i(\hat{\theta}_r))) \right] =$$

$$\hat{\lambda}' q_n(\hat{\theta}) + \sum_{i=1}^n \psi_2(\bar{\lambda}' q_i(\bar{\theta})) \hat{\lambda}' q_i(\hat{\theta}) q_i(\hat{\theta})' \hat{\lambda}$$

$$-\hat{\lambda}_r' q_n(\hat{\theta}_r) - \sum_{i=1}^n \psi_2(\bar{\lambda}_r' q_i(\bar{\theta}_r)) \hat{\lambda}_r' q_i(\hat{\theta}_r) q_i(\hat{\theta}_r)' \hat{\lambda}_r'$$

where $\bar{\lambda}' q_i(\bar{\theta})$ lies on the segment joining $\hat{\lambda}' q_i(\hat{\theta})$ to 0 and $\bar{\lambda}_r' q_i(\bar{\theta}_r)$ lies on the segment joining $\hat{\lambda}_r' q_i(\hat{\theta}_r)$ to 0. By Theorem 5.3, $\max_i |\bar{\lambda}' q_i(\bar{\theta})| = o_p(1)$ and $\max_i |\bar{\lambda}_r' q_i(\bar{\theta}_r)| = o_p(1)$ and hence $\psi_2(\bar{\lambda}' q_i(\bar{\theta})) = 1 + o_p(1)$ and $\psi_2(\bar{\lambda}_r' q_i(\bar{\theta}_r)) = 1 + o_p(1)$. Also, under the null hypothesys, $\hat{\lambda}' q_n(\hat{\theta}) = O_p(n^{-1})$ and $\hat{\lambda}_r' q_n(\hat{\theta}_r) = O_p(n^{-1})$. Thus,

$$2\sum_{i=1}^{n}\left[\gamma(\tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta}))) - \gamma(\tilde{\gamma}_1(\hat{\lambda}_r' q_i(\hat{\theta}_r)))\right] =$$

$$nq_n(\theta_o)P_o q_n(\theta_o) - nq_n(\theta_o)\tilde{P}_o q_n(\theta_o) + o_p(1)$$

and the result follows by noting that P_o has rank $m-k$ and $\tilde{P}_o$ has rank $m-k-d$.

A.3 Proof to Theorem 5.8

It can be proved along the lines of Theorem 5.7.

A.4 Proof to Theorem 5.9

It can be easily proved along the lines of Theorem 5.6.

A.5 Proof to Theorem 6.1

Let $\sum_{i=1}^{n} q(w_i,\theta)\nu_i = \xi'\nu$ then the MaxEnt problem is given by

$$\min \int \log\left(\frac{p(\nu|\theta)}{\tilde{p}(\nu|\theta)}\right) p(\nu|\theta)\mathrm{d}\upsilon$$

subject to

$$\mathcal{P}_\theta^b = \left\{ p(\nu|\theta) | \int (\xi'\nu)p(\nu|\theta)d\nu = 0 \right\}$$

The solution to the following probloem is well known to have the following Gibbs canonical density (see for a proof Theorem 3.1 of Csiszar(1975) and Kitamura and Stutzer(1997) for a discussion):

$$\frac{p(\nu|\theta)}{\tilde{p}(\nu|\theta)} = \frac{\exp[\lambda'\xi'\nu]}{\int \exp[\lambda'\xi'\nu]\tilde{p}(\nu|\theta)d\nu} \tag{42}$$

where

$$\lambda = \arg\min_{\lambda} \int \exp\{\lambda'\xi'\nu\}\tilde{p}(\nu|\theta)d\nu$$

A.6 Proof to Theorem 6.2

The MaxEnt solution solution (26) can be rewritten as

$$p(\nu|\theta) = \exp\left[\sum_{i=1}^{n} \tilde{\lambda}' q(w_i,\theta)\nu_i - \log \int \exp\left(\sum_{i=1}^{n} \tilde{\lambda}' q(w_i,\theta)\nu_i\right) \tilde{p}(\nu|\theta)\mathrm{d}\nu\right] \tilde{p}(\nu|\theta) \tag{43}$$

By the independence of $\{\nu_1, \nu_2, ..., \nu_n\}$ under $\tilde{p}(\nu|\theta)$ and using the properties of the logarithm and of the exponential, we get that for $\tau_i = \tilde{\lambda}' q(w_i, \theta)$

$$\begin{aligned} \log \int \exp(\sum_{i=1}^{n} \tau_i \nu_i) \tilde{p}(\nu|\theta)\mathrm{d}\nu &= \log \prod_{i=1}^{n} \int \exp(\tau_i \nu_i) \tilde{p}(\nu_i|\theta)\mathrm{d}\nu \\ &= \sum_{i=1}^{n} \log \int \exp(\tau_i \nu_i) \tilde{p}(\nu_i|\theta)\mathrm{d}\nu \end{aligned}$$

Using the expression above, (43) can be rewritten as

$$p(\nu|\theta) = \exp\left[\sum_{i=1}^{n} [\tau_i \nu_i - \varphi(\tau_i)]\right] \tilde{p}(\nu|\theta) \tag{44}$$

and $\tilde{\lambda}$ is now given by

$$\tilde{\lambda} = \arg \min_{\lambda \in \tilde{\Lambda}_n(\theta)} \sum_{i=1}^{n} \log \int \exp\left(\lambda' q(w_i,\theta)\nu_i\right) \tilde{p}(\nu_i|\theta)\mathrm{d}\nu$$

Independence follows from (44), since

$$\begin{aligned} p(\nu|\theta) &= \exp\left[\sum_{i=1}^{n} [\tau_i \nu_i - \varphi(\tau_i)]\right] \tilde{p}(\nu|\theta) \\ &= \prod_{i=1}^{n} \exp[\tau_i \nu_i - \varphi(\tau_i)] \tilde{p}(\nu_i|\theta) \end{aligned}$$

A.7 Proof to Theorems 6.3,6.4 and 6.5

We need to show that the moment generating function of Exponential, Poisson and Normal corresponds to the objective functions in the Theorems.

(1) The moment generating function of the Exponential distribution is given by $\ell(y) = \int_0^{+\infty} e^{-x} e^{yx} \mathrm{d}x$ and hence in this case $\varphi(y) = \log \int_0^{+\infty} e^{-x} e^{yx} \mathrm{d}x = -\log(1-y)$ and $\varphi_1(y) = (1-y)^{-1}$.

(2) For the Poisson, the moment generating function is given by $\varphi(y) = \exp(\exp(y) - 1)$ and hence $\varphi(y) = \exp(y) - 1$ and $\varphi_1(y) = \exp(y)$.

(3) The Moment generating function fo a normal is given by $\varphi(y) = \exp(y + y^2/2)\}$ and $\varphi(y) = y + y^2/2$ and $\varphi_1(y) = 1 + y$.

B Asymptotic Expansions

This Appendix provides proofs for the theorems concerning higher order properties given in Section 7.

Let $m_i(\tau) = m(w_i, \tau)$ denotes a vector valued function $m_i(\tau) : \Theta_\tau \to \mathbb{R}^g$ and let $m_n(\tau) = \frac{1}{n}\sum_i^n m_i(\tau)$. For Lemma B.2 and Lemma B.3 below, it is assumed that the $\hat{\tau}$ is $\sqrt{n}$-consistent for τ_o solving the unbiased estimating equation $\sqrt{n}m_n(\hat{\tau}) = 0$, at least with probability tending to one.

The Jacobian of $m_n(\tau)$ is denoted by $J_n(\tau)$ and $Q_n(\tau) = J_n(\tau)^{-1}$. The higher order derivatives of $m_n(\tau)$ are arranged recursively into matrices. $H_n(\tau)$, the matrix collecting the second derivatives of $m_n(\tau)$ is of dimension $g \times g^2$. $D_n(\tau)$, the matrix collecting the third derivatives of $m_n(\tau)$ is of dimension $g \times g^3$. The arrangement of the elements $(\partial/\partial\tau_j\partial\tau_r)m_T(\tau)$ for $j, r = 1, \dots, g$, into $H_n(\tau)$ is as follow:

$$H_n(\tau) = \begin{pmatrix} (\partial^2/\partial\tau_1\partial\tau')m_n(\tau) & \cdots & (\partial^2/\partial\tau_g\partial\tau')m_n(\tau) \end{pmatrix}$$

where $(\partial^2/\partial\tau_j\partial\tau')m_n(\tau)$ is a $g \times g$ matrix. The arrangement of the third derivatives into $D_n(\tau)$ follows the same pattern

$$D_n(\tau) = \begin{pmatrix} (\partial^3/\partial\tau_1\partial\tau_1\partial\tau')m_n(\tau) & \cdots & (\partial^3/\partial\tau_g\partial\tau_g\partial\tau')m_n(\tau) \end{pmatrix}$$

where $(\partial^3/\partial\tau_j\partial\tau_r\partial\tau')m_n(\tau)$ is a $g \times g$ matrix. This specification of the higher order derivatives is very convenient because it allows expressing Taylor's expansions as tensor products. Very similar notation is used by Rilstone, Srivastava, and Ullah (1996) and indeed Lemma B.2 and Lemma B.3 are adaptation of their results.

Given two matrices $A \in R^{m\times k}$and $B \in R^{g\times j}$, the Kronecker product, $A \otimes B$, is defined as the $(m \cdot g) \times (k \cdot j)$ matrix whose elements are given by $[a_{ij}B]_{ij}$. The vector e_j denote the $j - th$ unitary vecor of dimension $m \times 1$ or of dimension $k \times 1$, depending on the contest. If x is $g \times 1$ vector, $[x]_u$ denotes the u-th element of x. Similarly, $[x]_{1,\dots,k}$ denotes the first k elements of x.

B.1 Lemmas

Lemma B.1: *Suppose J_n is bounded and uniformly positive definite matrix such that $J_n \xrightarrow{p} J_o$. Suppose there exists a matrix $Z_n = O_p(n^{-1/2})$ such that $J_n = J_o - Z_n$, then the following expansions hold for $Q_n = J_n^{-1}$:*

$$\begin{aligned} Q_n &= Q_o + O_p(n^{-1/2}) \\ Q_n &= Q_o - Q_o Z_n Q_o + O_p(n^{-1}) \\ Q_n &= Q_o - Q_o Z_n Q_o + Q_o Z_n Q_o Z_n Q_o + O_p(n^{-3/2}) \end{aligned}$$

where $Q_o = \text{plim}_{n\to\infty} Q_n$.

Lemma B.2: *Suppose (i) $\parallel \hat{\tau} - \tau_o \parallel = O_p(n^{-1/2})$; (ii) $Q_o(\tau) = E[\nabla_\tau m_i(\tau)] < \infty$, for any $\tau \in \mathcal{S}(\tau_o, \delta)$, $\delta > 0$, exists; (iii) $Q_o m_n(\tau_o) = O_p(n^{-1/2})$; (iv) for $j, r = 1, \ldots, g$ $E\left[\left(\frac{\partial m_i(\tau_o)}{\partial \tau_j \partial \tau_r}\right)^2\right]$; (iv) there exists $\bar{\mathcal{B}}(w_i)$, $E[\bar{\mathcal{B}}(w_i)] < \infty$ and $\delta > 0$ such that for $j, r = 1, \ldots, g$,*

$$|(\partial/\partial\tau_j\partial\tau_r)m_i(\tau) - (\partial/\partial\tau_j\partial\tau_r)m_i(\tau_o)| \leq \bar{\mathcal{B}}_i(w) \parallel \tau - \tau_o \parallel$$

for any $\tau \in \mathcal{S}(\tau_o, \delta)$.

Then

$$(\hat{\tau} - \tau_o) = f_n + b_n + O_p(n^{-3/2})$$

where

$$f_n = -Q_o m_n(\tau)$$

and

$$b_n = Q_o Z_n(\tau_o) f_n(\tau_o) - \frac{1}{2} H_o(f_n(\tau_o) \otimes f_n(\tau_o))$$

Lemma B.3: *Suppose (i) $\parallel \hat{\tau} - \tau_o \parallel = O_p(n^{-1/2})$; (ii) $Q_o(\tau) = E[\nabla_\tau m_i(\tau)] < \infty$, for any $\tau \in \mathcal{S}(\tau_o, \delta)$, $\delta > 0$, exists; (iii) $Q_o m_n(\tau_o) = O_p(n^{-1/2})$; (iv) for $j, r, p = 1, \ldots, g$ $E\left[\left(\frac{\partial m_i(\tau_o)}{\partial \tau_j \partial \tau_r \partial \tau_p}\right)^2\right] \leq \infty$; (vi) there exists $\bar{\mathcal{C}}(w_i)$, $E[\bar{\mathcal{C}}(w_i)] < \infty$ and $\delta > 0$ such that for $j, r, p = 1, \ldots, g$,*

$$|(\partial/\partial\tau_j\partial\tau_r\partial\tau_l)m_i(\tau) - (\partial/\partial\tau_j\partial\tau_r\partial\tau_l)m_i(\tau_o)| \leq \bar{\mathcal{C}}_i(w) \parallel \tau - \tau_o \parallel$$

, $E[\bar{\mathcal{C}}_i(w)] < \infty$

Then

$$(\hat{\tau} - \tau_0) \;=\; f_n + b_n + r_n + O_p(n^{-2})$$

where the expressions for $f_n(\tau_o)$ *and* $b_n(\tau_o)$ *are the same of those in Lemma B.2 and*

$$\begin{aligned} r_n &= -\frac{1}{2} Q_o H_o \left\{ (f_n + b_n) \otimes (a_n + b_n) \right\} \\ &\quad + \frac{1}{6} D_o \left\{ f_n \otimes f_n \otimes f_n \right\} \end{aligned}$$

B.2 Proofs of Lemma

Proof of Lemma B1:

By assumption $Q_o Z_n = O_p(n^{-1/2})$. Rewrite Q_n as $Q_n = (J_o + Z_n)$. Multiplying and dividing by $(Q_o - Q_o Z_n Q_o)$ yields

$$\begin{aligned} Q_n &= (J_o + Z_n)^{-1} \\ &= (Q_o - Q_o Z_n Q_o)(Q_o - Q_o Z_n Q_o)^{-1}(I + Q_o Z_n)^{-1} \\ &= (Q_o - Q_o Z_n Q_o)(I - Z_n Q_o Z_n Q_o)^{-1} \end{aligned}$$

Notice that

$$(I - Z_n Q_o Z_n Q_o)^{-1} \;=\; (I + Z_n Q_o Z_n Q_o) + O_p(n^{-3/2})$$

Thus,

$$\begin{aligned} Q_n &= (Q_o - Q_o Z_n Q_o) \left[I + Z_n Q_o Z_n Q_o + O_p(n^{-3/2}) \right] \\ &= Q_o - Q_o Z_n Q_o + Q_o Z_n Q_o Z_n Q_o + O_p(n^{-3/2}) \end{aligned}$$

The result follows by noting that last two terms are of the required order, that is $O_p(n^{-1/2})$ and $O_p(n^{-1})$ respectively.

Proof to Lemma B2

By assumption $\hat{\tau}$ is a consistent root of $m_n(\tau)$, that is

$$m_n(\hat{\tau})/\sqrt{n} = 0$$

at least with probability tending to one. Taking a mean value expansion around τ_o gives

$$m_n(\tau_o)/\sqrt{n} + J_n(\tau_o)(\hat{\tau} - \tau_o)/\sqrt{n} + \frac{1}{2}H_n(\bar{\tau})[(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)]/\sqrt{n} = 0$$

where $\bar{\tau}$ lies between $\hat{\tau}$ and τ_o and it is allowed to differ between rows of $H_n(\cdot)$. Solving for $(\hat{\tau} - \tau_o)$ yields

$$(\hat{\tau} - \tau_o) = -Q_n(\tau_o)m_n(\tau_o) - \frac{1}{2}Q_n(\tau_o)H_n(\bar{\tau})[(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)]$$

Adding and subtracting $\frac{1}{2}Q_n(\tau_o)H_o[(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)]$ gives

$$\begin{aligned}(\hat{\tau} - \tau_o) \quad = \quad & -Q_n(\tau_o)m_n(\tau_o) - \frac{1}{2}Q_n(\tau_o)H_o[(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)] \\ & +\frac{1}{2}Q_n(\tau_o)\left\{H_o - H_n(\bar{\tau})\right\}[(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)]\end{aligned}$$

By the assumption, the Jacobian is bounded

$$\|J_n(\tau_o) - J_o\| = O_p(n^{-1/2})$$

It follows that the results of Lemma B.1 can be applied to $J_n(\tau_o)$ with $Z_n(\tau_o) = J_n(\tau_o) - J_o$. Substituting the approximation for $Q_n(\tau_o)$ from the Lemma B.1 up to order n^{-1} in the first term, up to order $n^{-1/2}$ in the second and third terms and substituting $(\hat{\tau} - \tau_o) = -Q_o m_n(\tau_o) + O_p(n^{-1})$ in the terms involved in the Kronecker products, gives

$$\begin{aligned}(\hat{\tau} - \tau_o) \quad = \quad & -\left\{Q_o - Q_o Z_n(\tau_o) Q_o + O_p(n^{-1})\right\} m_n(\tau_o) \\ & -\frac{1}{2}\left\{Q_o + O_p(n^{-1/2})\right\} H_o \left\{Q_o m_n(\tau_o) \otimes Q_o m_n(\tau_o)\right\} \\ & +\frac{1}{2}\left\{Q_o + O_p(n^{-1/2})\right\}\left\{H_o - H_n(\bar{\tau})\right\}\left\{Q_o m_n(\tau_o) \otimes Q_{o,m} m_n(\tau_o)\right\}\end{aligned} \qquad (45)$$

It can be shown that the elements of $\{H_o - H_n(\bar{\tau})\}$ are bounded in probability of order $n^{-1/2}$. The (j, r) element of $\{H_o - H_n(\tau)\}$ satisfies the following inequality

$$\begin{aligned}
&\left\| \frac{\partial^2 m_n(\bar{\tau})}{\partial\tau_j \partial\tau_r} - E\left[\frac{\partial^2 m_n(\tau_o)}{\partial\tau_j \partial\tau_r}\right] \right\| \\
&\qquad \leq \left\| \frac{\partial^2 m_n(\bar{\tau})}{\partial\tau_j \partial\tau_r} - \frac{\partial^2 m_n(\tau_o)}{\partial\tau_j \partial\tau_r} \right\| + \left\| \frac{\partial^2 m_n(\tau_o)}{\partial\tau_j \partial\tau_r} - E\left[\frac{\partial^2 m_n(\tau_o)}{\partial\tau_j \partial\tau_r}\right] \right\| \qquad (46)
\end{aligned}$$

The first term after the inequality is bounded by

$$\begin{aligned}
\left\| \frac{\partial^2 m_n(\bar{\tau})}{\partial\tau_j \partial\tau_r} - \frac{\partial^2 m_n(\tau_o)}{\partial\tau_j \partial\tau_r} \right\| &\leq \frac{1}{n} \sum_{i=1}^{n} \left\| \frac{\partial^2 m_i(\bar{\tau})}{\partial\tau_j \partial\tau_r} - \frac{\partial^2 m_i(\tau_o)}{\partial\tau_j \partial\tau_r} \right\| \\
&\leq \left[\frac{1}{n} \sum_{t=1}^{n} \bar{\mathcal{B}}_i(w)\right] \|\bar{\tau} - \tau_o\|
\end{aligned}$$

By *(iii)* and the law of large numbers, $\sum_i^n \bar{\mathcal{B}}_i(w)/n = O_p(1)$. Since $\bar{\tau} \xrightarrow{p} \tau_o$ by *(i)* $\|\bar{\tau} - \tau_o\| = O_p(n^{-1/2})$ and thus $\left[\frac{1}{n}\sum_{t=1}^{n} \bar{\mathcal{B}}_i(w)\right] \|\bar{\tau} - \tau_o\| = O_p(n^{-1/2})$, giving that

$$\left\| \frac{\partial^2 m_n(\bar{\tau})}{\partial\tau_j \partial\tau_r} - \frac{\partial^2 m_n(\tau_o)}{\partial\tau_j \partial\tau_r} \right\| = O_p(n^{-1/2})$$

Under *(iv)* and by applications of the Central Limit Theorem gives that

$$\left\| \frac{\partial^2 m_n(\tau_o)}{\partial\tau_j \partial\tau_r} - E\left[\frac{\partial^2 m_n(\tau_o)}{\partial\tau_j \partial\tau_r}\right] \right\| = O_p(n^{-1/2})$$

and hence, $H_o - H_n(\bar{\tau}) = O_p(n^{-1/2})$. By observing that the Kronecker products in the second and third terms of (45) are are bounded in probability of order n^{-1} it follows that

$$O_p(n^{-1/2}) \cdot \{H_o - H_n(\bar{\tau})\} \{Q_o m_n(\tau_o) \otimes Q_o m_n(\tau_o)\} = o_p(n^{-3/2})$$

and similarly

$$O_p(n^{-1/2}) \cdot H_o \{Q_o m_n(\tau_o) \otimes Q_o m_n(\tau_o)\} = O_p(n^{-3/2})$$

Since $O_p(n^{-1})m_n(\tau_o) = O_p(n^{-3/2})$, collecting terms and dropping terms of order $n^{-\zeta}$, $\zeta \geq 3/2$, gives

$$\begin{aligned}(\hat{\tau} - \tau_o) &= -Q_o(\tau_o)m_n(\tau_o) + Q_oZ_n(\tau_o)Q_om_n(\tau_o) \\ &\quad -\frac{1}{2}Q_oH_o\left\{Q_om_n(\tau_o) \otimes Q_om_n(\tau_o)\right\} + O_p(n^{-3/2})\end{aligned}$$

as required.

Proof to Lemma B.3

The prove is very similar to that of Lemma B.2. By third order Taylor expansion with Lagrange remainder of $m_n(\hat{\tau})/\sqrt{n} = 0$ around τ_o and by solving for $\hat{\tau} - \tau_o$, we obtain

$$\begin{aligned}(\hat{\tau} - \tau_o) &= -Q_n(\tau_o)m_n(\tau_o) - \frac{1}{2}Q_n(\tau_o)H_n(\tau_o)\left\{(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)\right\} \\ &\quad -\frac{1}{6}Q_n(\tau_o)D_n(\bar{\tau})\left\{(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)\right\}\end{aligned}$$

where $\bar{\tau}$ lies between $\hat{\tau}$ and τ_o and it is allowed to differ across different rows of $D_n(\cdot)$. Adding and subtracting the last term with $D_n(\bar{\tau})$ replaced by D_o, the above expression can be rewritten as

$$\begin{aligned}(\hat{\tau} - \tau_o) &= -Q_n(\tau_o)m_n(\tau_o) - \frac{1}{2}Q_n(\tau_o)\left(H_o - H_n(\tau_o)\right)\left\{(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)\right\} \\ &\quad -\frac{1}{6}Q_n(\tau_o)D_o\left\{(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)\right\} \\ &\quad -\frac{1}{6}Q_n(\tau_o)\left[D_n(\bar{\tau}) - D_o\right]\left\{(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)\right\}\end{aligned}$$

where $H_o - H_n(\tau_o)$ is $O_p(n^{-1/2})$ in virtue of *(iv)* by application of the CLT to its elements. Also, $(\hat{\tau} - \tau_o) = O_p(n^{-1/2})$ and the Kronecker products in the second and third term is $O_n(n^{-3/2})$. The term involving the difference between the matrix of third derivatives and its expectation is bounded in probability. Considering the generic element of $[D_n(\bar{\tau}) - D_o]$ gives that

$$\begin{aligned}
&\left\| \frac{\partial^3 m_n(\bar{\tau})}{\partial\tau_j \partial\tau_r \partial\tau_l} - E\frac{\partial^3 m_n(\tau_o)}{\partial\tau_j \partial\tau_r \partial\tau_l} \right\| \\
&\qquad \leq \left\| \frac{\partial^3 m_n(\bar{\tau})}{\partial\tau_j \partial\tau_r \partial\tau_l} - \frac{\partial^3 m_n(\tau_o)}{\partial\tau_j \partial\tau_r \partial\tau_l} \right\| + \left\| \frac{\partial^3 m_n(\tau_o)}{\partial\tau_j \partial\tau_r \partial\tau_l} - E\left[\frac{\partial^3 m_n(\tau_o)}{\partial\tau_j \partial\tau_r \partial\tau_l} \right] \right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left\| \frac{\partial^3 m_i(\bar{\tau})}{\partial\tau_j \partial\tau_r \partial\tau_l} - \frac{\partial^3 m_i(\tau_o)}{\partial\tau_j \partial\tau_r \partial\tau_l} \right\| + \frac{1}{n}\sum_{i=1}^{n} \left\| \frac{\partial^3}{\partial\tau_j \partial\tau_r \partial\tau_l} m_i(\tau_o) - E\frac{\partial^3}{\partial\tau_j \partial\tau_r \partial\tau_l} m_i(\tau_o) \right\|
\end{aligned}$$

The first term after the inequality above is, by assumption, bounded above by $\|\bar{\tau} - \tau_o\| \frac{1}{n}\sum_i^n \bar{\mathcal{C}}_i(w)$. Since $\bar{\tau} \xrightarrow{p} \tau_o$, normality of $\sqrt{n}(\hat{\tau} - \tau_o)$ and $E\bar{\mathcal{C}}_i(w) \leq \infty$, implies that $\|\bar{\tau} - \tau_o\| \frac{1}{n}\sum_i^n \bar{\mathcal{C}}_i(w) = O_p(n^{-1/2})$. The second term is $O_p(n^{-1/2})$ by *(v)* and the Central Limit Theorem. Substituting the expansions of $Q_m(\tau_o)$ given in Lemma B.1 and collecting terms of similar order, we obtain

$$\begin{aligned}
(\hat{\tau} - \tau_o) \quad = \quad & -[Q_o - Q_o Z_n(\tau_o) Q_o + Q_o Z_n(\tau_o) Q_o V_n(\tau_o) Q_o] m_n(\tau_o) \\
& -\frac{1}{2}\left[Q_o - Q_o Z_n(\tau_o) Q_o\right] H_o \left\{(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)\right\} \\
& -\frac{1}{2} Q_o [H_o - H_n(\tau_o)] \left\{(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)\right\} \\
& -\frac{1}{6} Q_o D_o \left\{(\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o) \otimes (\hat{\tau} - \tau_o)\right\} \\
& +\underbrace{O_p(n^{-3/2}) m_n(\tau_o)}_{O_p(n^{-2})} + \underbrace{O_p(n^{-1})[H_o - H_n(\tau_o)] \left\{(\hat{\tau} - \tau_0) \otimes (\hat{\tau} - \tau_0)\right\}}_{O_p(n^{-1}) O_p(n^{-1/2}) O_p(n^{-1}) = o_p(n^{-2})} \\
& +\underbrace{O_p(n^{-1/2}) D_{o,m} \left\{(\hat{\tau} - \tau_0) \otimes (\hat{\tau} - \tau_0) \otimes (\hat{\tau} - \tau_0)\right\}}_{O_p(n^{-1/2}) \cdot O_p(n^{-1}) \cdot O_p(n^{-3/2}) = O_p(n^{-2})} \\
& +\underbrace{\frac{1}{6} Q_o \left(D_n(\bar{\tau}) - D_o\right) \left\{(\hat{\tau} - \tau_0) \otimes (\hat{\tau} - \tau_0) \otimes (\hat{\tau} - \tau_0)\right\}}_{O_p(1) \cdot O_p(n^{-1/2}) O_p(n^{-3/2}) = O_p(n^{-2})}
\end{aligned}$$

Substituting $(\tau - \tau_o) = f_n(\tau_o) + O_p(n^{-1/2})$ in the second and third summand and $(\hat{\tau} - \tau_o) = f_n(\tau_o) + A_n(\tau_o) + O_p(n^{-3/2})$ in the fourth term, and noting that

$$Q_o Z_n(\tau_o) Q_o \left\{[f_n(\tau_o) \otimes A_n(\tau_o)] + [f_n(\tau_o) \otimes A_n(\tau_o)]\right\} = o_p(n^{-2})$$

and dropping terms of order lower than $O_p(n^{-\zeta})$ $\zeta \geq 2$ gives

$$\begin{aligned}(\hat{\tau}-\tau_o) \;=\;& -Q_o m_n(\tau_o) + Q_o Z_n(\tau_o) Q_o m_n(\tau_o) \\ & -Q_o Z_n(\tau_o) Q_o V_n(\tau_o) Q_o m_n(\tau_o) \\ & -\frac{1}{2} Q_o H_o \left\{[f_n(\tau_o)\otimes A_n(\tau_o)] + [f_n(\tau_o)\otimes A_n(\tau_o)]\right\} \\ & -\frac{1}{2} Q_o [H_o - H_n(\tau_o)] \left\{f_n(\tau_o)\otimes f_n(\tau_o)\right\} \\ & +O_p(n^{-2})\end{aligned}$$

as required.

B.2.1 Proof to Theorem 7.1

Let $g = k+m$ and $h = k+1$. Let $\hat{\tau}$ denote the $(k+m)\times 1$ vector that stacks the MD estimator of θ_o and λ, that is $\hat{\tau} = (\hat{\theta}', \hat{\lambda}')'$. With probability approaching to one, $\hat{\tau}$ solves the first order conditions of MD given by

$$\begin{aligned}[m_n(\hat{\tau})]_{1,k} &\equiv \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_i \nabla_\theta q(w,\hat{\theta})'\hat{\lambda} = 0 \\ [m_n(\hat{\tau})]_{h,g} &\equiv \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_i q(w,\hat{\theta}) = 0\end{aligned}$$

where here and throughout the proof $\hat{\pi}_i = \frac{1}{n}\tilde{\gamma}_1(\gamma_1 + \hat{\lambda}' q(w_i,\hat{\theta}))$ and even if the arguments are dropped for notational reasons, $\hat{\pi}_i$ must be interpreted as function of $\hat{\theta}$ and $\hat{\lambda}$. Similarly, $\bar{\pi}_i = \frac{1}{n}\tilde{\gamma}_1(\gamma_1 + \bar{\lambda}' q(w_i,\bar{\theta}))$. Form Theorem ?.??, it follows that $(\hat{\tau}-\tau_o) = -Q_o m_n(\tau_o) + o_p(1)$, where

$$Q_o = \left[\begin{array}{cc} -S_o & B_o \\ B_o' & P_o \end{array}\right]$$

To be apply to apply Lemma B.1, required by Lemma B.2, is sufficient to show that $\|J_n(\tau_o) - J_o\|$ is of order $n^{-1/2}$. Note that

$$J_n(\tau_o) = \left[\begin{array}{cc} 0 & \frac{1}{n}\sum_i^n \nabla_\theta q_i(\theta_o)' \\ \frac{1}{n}\sum_i^n \nabla_\theta q(w,\theta_o) & \frac{1}{n}\sum_i^n q_i(\theta_o) q_i(\theta_o)' \end{array}\right]$$

By Assumption C, the elements of $\frac{1}{n}\sum_i^n \nabla_\theta q_i(\theta_o)$ and $\frac{1}{n}\sum_i^n q_i(\theta_o) q_i(\theta_o)'$ obeys the CLT and hence $\|J_n(\tau_o) - J_o\| = O_p(n^{-1/2})$. To apply Lemma B.2, it is sufficient that $\|$

$H_n(\bar{\tau}) - H_o \parallel$ be of order $n^{-1/2}$. For $j = 1, \dots k+m$, and letting $\nu_i(\tau) = \gamma_1 + \lambda' q_i(\theta)$ and $b_i(\tau) = \partial \nu_i(\tau)/\partial\tau$

$$\begin{aligned} \frac{\partial^2 m_i(\tau)}{\partial\tau_j \partial\tau} &= \tilde{\gamma}_1(\nu_i(\tau))\partial^2 b_i(\tau)/\partial\tau_j\partial\tau + \tilde{\gamma}_3(\nu_i(\tau))b_i(\tau)_j b_i(\tau)b_i(\tau)' \\ &+\tilde{\gamma}_2(\nu_i(\tau))\left\{\partial[b_i(\tau)b_i(\tau)']/\partial\tau_j + b_i(\tau)_j\partial b_i(\tau)/\partial\tau\right\} \end{aligned}$$

where $\tilde{\gamma}_2(x) = 1/\gamma_2(\tilde{\gamma}_1(x))$ and $\tilde{\gamma}_3(x) = -\gamma_3(\tilde{\gamma}_1(x))\tilde{\gamma}_2(x)/[\gamma_2(\tilde{\gamma}_1(x))]^2$ and $\gamma_3(x) = \partial^3\gamma(x)/\partial^3 x$. By assumption, $\tilde{\gamma}_j(x)$, $j = 2, 3$ are compositions of continuously differentiable and function in a neighborhood of zero and thus for $\bar{\tau} \in \mathcal{S}(\tau_o, \delta)$

$$|\tilde{\gamma}_1(\nu_i(\bar{\tau}))| \leq C\|\bar{\tau}\|\|q(w_i, \bar{\theta})\| \leq C\mathcal{B}_i(w)\|\bar{\tau} - \tau_o\|$$

and similarly

$$\tilde{\gamma}_j(\nu_i(\bar{\tau})) \leq C\mathcal{B}_i(w)\|\bar{\tau} - \tau_o\|$$

Since the expressions for the second derivatives involves at maximum combinations of three functions $b(\cdot)$, the elements of $\|H_m(\tau) - H_m(\tau_o)\|$ are bounded in a neighborhood by $C\frac{1}{n}\sum_i^n \mathcal{B}_i^4(w)\|\bar{\theta} - \theta_o\|$ and by Assumption C, $E[\mathcal{B}_i(w)^4] \leq \infty$, $\|H_n(\tau) - H_o\| = O_p(n^{-1/2})$. Thus, Lemma B.2 applies with $\bar{\mathcal{B}}(w_i) = \mathcal{B}(w_i)^4$.

The terms of order $n^{-1/2}$ are given by

$$Q_o q_n = \begin{pmatrix} -B_o q_n \\ -P_o q_n \end{pmatrix} = \begin{pmatrix} u_n \\ l_n \end{pmatrix} \tag{47}$$

The terms of order n^{-1} are given by

$$b_n = Q_o Z_n Q_o q_n - Q_o H_o \left\{\begin{pmatrix} u_n \\ l_n \end{pmatrix} \otimes \begin{pmatrix} u_n \\ l_n \end{pmatrix}\right\}/2$$

where

$$Z_n = \begin{bmatrix} 0 & \Gamma_o' - \Gamma_n' \\ \Gamma_o - \Gamma_n & V_o - V_n \end{bmatrix}$$

It follows that using (47) we obtain

$$\begin{aligned} Q_o Z_n Q_o q_n &= \left\{ \begin{bmatrix} I_k & 0 \\ 0 & I_m \end{bmatrix} - \begin{bmatrix} B_o \Gamma_n & -S_o \Gamma_n' + B_o V_n \\ P_o \Gamma_n & B_o' \Gamma_n' + P_o V_n \end{bmatrix} \right\} \begin{pmatrix} u_n \\ l_n \end{pmatrix} \\ &= \begin{bmatrix} u_n - B_o \Gamma_n u_n + S_o \Gamma_n' u_n - B_o V_n l_n \\ l_n - P_o \Gamma_n u_n - B_o' \Gamma_n' l_n - P_o V_n l_n \end{bmatrix} \end{aligned}$$

Define $\beta_j^{sh} \equiv \; E\left[(\partial/\partial s_j \partial h)[m_n(\tau_o)]_{1,k}\right]$ and $\mu_j^{sh} \equiv \; E\left[(\partial/\partial s_j \partial h)[m_n(\tau_o)]_{h,g}\right]$ for $s, h = \{\theta, \lambda\}$. The term

$$Q_o H_o \left\{ f(\tau_o) \otimes f(\tau_o) \right\} = \begin{pmatrix} -S_o \nabla_1 + B_o \nabla_2 \\ B_o' \nabla_1 + P_o \nabla_2 \end{pmatrix}$$

where

$$\begin{aligned} \nabla_1 &= \sum_{j=1}^{k} \beta_j^{\theta\theta} u_{n,j} u_n + \sum_{j=1}^{k} \beta_j^{\theta\lambda} u_{n,j} l_n \\ & \quad \sum_{j=1}^{m} \beta_j^{\lambda\theta} l_{n,j} u_n + \sum_{j=1}^{m} \beta_j^{\lambda\theta} l_{n,j} l_n \\ \nabla_2 &= \sum_{j=1}^{k} \mu_j^{\theta\theta} u_n u_{n,j} + \sum_{j=1}^{k} \mu_j^{\theta\lambda} u_{n,j} l_n \\ & \quad \sum_{j=1}^{m} \mu_j^{\lambda\theta} l_{n,j} u_n + \sum_{j=1}^{m} \mu_j^{\lambda\lambda} l_{n,j} l_n \end{aligned}$$

and $l_{n,j}$ and $u_{n,j}$ denote respectively the j-esime elements of l_n and u_n respectively. Thus, the first k elements of the expansion for $(\hat{\tau} - \tau_o)$ are given by

$$\begin{aligned} (\hat{\theta} - \theta_o) \quad = u_n - & \quad B_o \Gamma_n u_n + S_o \Gamma_n' u_n - B_o V_n l_n \\ & + \frac{1}{2} S_o \nabla_1 - \frac{1}{2} B_o \nabla_2 \end{aligned}$$

The expression above gives the first conclusion for $f_n^{\theta} = u_n$ and

$$b_n^{\theta} = -B_o \Gamma_n u_n + S_o \Gamma_n' u_n - B_o V_n l_n + \frac{1}{2} S_o \nabla_1 - \frac{1}{2} B_o \nabla_2$$

Similarly, taking the last m elements yields the expression up to order $n^{-3/2}$ for the

Lagrange multiplier

$$\begin{aligned}\hat{\lambda} \ &= l_n - \quad P_o\Gamma_n u_n + B_o'\Gamma_n' l_n - P_o V_n l_n \\ &\qquad -\frac{1}{2}B_o'\nabla_1 - \frac{1}{2}P_o\nabla_2\end{aligned}$$

Giving for $f_n^\lambda = l_n$ and

$$b_n^\lambda = P_o\Gamma_n u_n + B_o'\Gamma_n' l_n + P_o V_n l_n - \frac{1}{2}B_o'\nabla_1 - \frac{1}{2}P_o\nabla_2$$

the conclusion of the theorem.

B.2.2 Proof to Theorem 7.2

The bias up to order n^{-1} of $\hat{\theta}$ is given by the expectation of the terms of order n^{-1} in the expansion of $\hat{\theta}$. Dropping terms of zero expectation, the higher order bias is given by

$$\begin{aligned}E\left[b_n^\theta\right] \ &= \ -E\left[B_o\Gamma_n u_n\right] + E\left[S_o\Gamma_n' u_n\right] - E\left[B_o V_n l_n\right] \\ &\quad +S_o\nabla_1/2 - B_o\nabla_2/2\end{aligned}$$

We analyze the expectation of the terms involved in $E(b_n^\theta)$:
(i) $E\left[B_o\Gamma_n u_n\right]$

$$\begin{aligned}E\left[B_o\Gamma_n u_n\right] \ &= \ -B_o\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[\nabla_\theta q_i(\theta_o)B_o q_j(\theta_o)\right]/n^2 \\ &= \ -B_o E\left[\nabla_\theta q_t(\theta)B_o q_t(\theta_o)\right]/n\end{aligned}$$

(ii) $E\left[B_o V_n l_n\right]$

$$\begin{aligned}E\left(B_o V_n l_n\right) \ &= \ -B_o\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[q_i(\theta_o)q_i(\theta_o)P_o q_j(\theta_o)\right]/n^2 \\ &= \ -B_o E\left[q_i(\theta_o)q_i(\theta_o)P_o q_i(\theta_o)\right]/n\end{aligned}$$

(iii) $E\left[S_o \Gamma_n l_n\right]$

$$\begin{aligned} E\left(S_o \Gamma_n l_n\right) &= - \quad S_o \sum_{i=1}^{n} \sum_{j=1}^{n} E\left[\nabla_\theta q_i(\theta_o) P_o q_j(\theta_o)\right]/n^2 \\ &= - \quad S_o E\left[\nabla_\theta q_i(\theta) P_o q_i(\theta_o)\right]/n \end{aligned}$$

(iv) $E\left[S_o \nabla_1\right]$

$$\begin{aligned} E\left(\nabla_1\right) &= \sum_{j=1}^{k} \beta_j^{\theta\lambda} E(u_{n,j} l_n) + \sum_{j=1}^{m} \beta_j^{\lambda\theta} E(l_{n,j} u_n) \\ & \quad + \sum_{j=1}^{m} \beta_j^{\lambda\lambda} E(l_{n,j} l_n) \end{aligned}$$

It is easy to show that the expectations involving products of the influence function of λ and θ vanish:

$$\begin{aligned} E\left[u_{n,j} l_n\right] &= E\left[l_n u_n'\right] e_j \\ &= E\left[P_o \sum_{i=1}^{n} q_i(\theta_o) \sum_{i=1}^{n} q_i'(\theta_o) B_o'\right] e_j \\ &= P_o E[q_i(\theta_o) q_i(\theta_o)'] B_o' e_j \\ &= P_o V_o B_o' e_j = 0 \end{aligned}$$

and similarly

$$\begin{aligned} E\left[l_{n,j} u_n\right] &= E(u_n l_n') e_j \\ &= B_o E[q_i(\theta_o) q_i(\theta_o)'] P_o e_j \\ &= 0 \end{aligned}$$

Thus, $E(\nabla_1) = \sum_{j=1}^{m} \beta_j^{\lambda\lambda} E(l_{n,j} l_n)$. By $P_o V_o P_o = P_o$, it follows that $E[l_{n,j} l_n] = P_o e_j / n$. The terms $\beta_j^{\lambda\lambda}$ is given by

$$\beta_j^{\lambda\lambda} = E\left[\nabla_\theta q_t(\theta_o)' e_j q_i(\theta_o)' + q_{t,j} \nabla_\theta q_t(\theta_o)'\right]$$

and by symmetry of P_o follows that

$$E[\nabla_1] = 2E\left[\nabla_\theta q_i(\theta_o)' P_o q_i(\theta_o)\right]$$

yielding

$$E[S_o\nabla_1] = 2S_oE\left[\nabla_\theta q_i(\theta_o)' P_o q_i(\theta_o)\right] \tag{48}$$

(v) $E[B_o\nabla_2]$

$$\begin{aligned} E[\nabla_2] &= \sum_{j=1}^{k}\mu_j^{\theta\theta}E[u_{n,j}u_n] + \sum_{j=1}^{k}\mu_j^{\theta\lambda}E[u_{n,j}l_n] \\ &\quad + \sum_{j=1}^{m}\mu_j^{\lambda\theta}E[l_{n,j}u_n] + \sum_{j=1}^{m}\mu_j^{\lambda\lambda}E[l_{n,j}l_n] \end{aligned}$$

By the same arguments of point *(iv)*, $E[l_{n,j}u_n] = E[u_{n,j}l_n] = 0$, and hence

$$E(\nabla_2) = \sum_{j=1}^{k}\mu_j^{\theta\theta}E[u_{n,j}u_n] + \sum_{j=1}^{m}\mu_j^{\lambda\lambda}E[l_{n,j}l_n]$$

where

$$\mu_j^{\theta\theta} = E\left[\frac{\partial q_i(\theta_o)}{\partial\theta_j\partial\theta}\right];\ \ \mu_j^{\lambda\lambda} = \gamma_3E[q_{i,j}(\theta_o)q_i(\theta_o)q_i(\theta_o)]$$

Since $E(u_{n,j}u_n) = B_oV_oB_o'e_j/n$ and $B_oV_oB_o' = S_o$, we have

$$\sum_{j=1}^{k}\mu_j^{\theta\theta}E[u_{n,j}u_n] = \sum_{j=1}^{k}E\left[\frac{\partial q_i(\theta_o)}{\partial\theta_j\partial\theta}\right]S_oe_j/n$$

and noting that $E(l_{n,j}l_n) = P_oV_oP_oe_j = P_oe_j$ yields

$$\begin{aligned} \sum_{j=1}^{m}\mu_j^{\lambda\lambda}E(l_{n,j}l_n) &= \sum_{j=1}^{m}\gamma_3E[q_{i,j}(\theta_o)q_i(\theta_o)q_i(\theta_o)']P_oe_j/n \\ &= \sum_{j=1}^{m}\gamma_3E[q_i(\theta_o)q_i(\theta_o)'P_oe_jq_{i,j}(\theta_o)]/n \\ &= \gamma_3E\left[q_i(\theta_o)q_i(\theta_o)\left(\sum_{j=1}^{m}P_oe_je_j'\right)q_i(\theta_o)\right]/n \\ &= \gamma_3E\left[q_i(\theta_o)q_i(\theta_o)P_oq_i(\theta_o)\right]/n \end{aligned}$$

and hence

$$E(\nabla_2) = \gamma_3 E\left[q_i(\theta_o)q_i(\theta_o)P_o q_i(\theta_o)\right]/n + \sum_{j=1}^{k} E\left[\frac{\partial q_i(\theta_o)}{\partial\theta_j\partial\theta}\right] S_o e_j/n$$

Substituting expression of part *(i)-(v)* into the expectation for the term of order n^{-1} in the asymptotic expansion we obtain

$$\begin{aligned} E(b_n) &= -E[B_o\Gamma_n u_n] + E[S_o{\Gamma'}_n u_n] - E[B_o V_n l_n] \\ &\quad +\frac{1}{2}S_o E(\nabla_1) - \frac{1}{2}B_o E(\nabla_2) \end{aligned}$$

$$\begin{aligned} B_{-1}(\theta_o) &= B_o E\left[\nabla_\theta q_t(\theta)B_o q_t(\theta_o)\right]/n \\ &\quad -S_o E\left[\nabla_\theta q_i(\theta)P_o q_i(\theta_o)\right]/n \\ &\quad +B_o E\left[q_i(\theta_o)q_i(\theta_o)P_o q_i(\theta_o)\right]/n \\ &\quad +S_o E\left[\nabla_\theta q_i(\theta_o)' P_o q_i(\theta_o)\right]/n - \frac{1}{2}\gamma_3 B_o E\left[q_i(\theta_o)q_i(\theta_o)' P_o q_i(\theta_o)\right]/n \\ &\quad -\frac{1}{2}B_o\sum_{j=1}^{k} E\left[\frac{\partial q_i(\theta_o)}{\partial\theta_j\partial\theta}\right] S_o e_j/n \end{aligned}$$

simplifying and rearranging yields

$$\begin{aligned} B_{-1}(\theta) &= B_o E\left[\nabla_\theta q_t(\theta)B_o q_t(\theta_o)\right]/n + \left(1-\frac{\gamma_3}{2}\right)B_o E\left[q_i(\theta_o)q_i(\theta_o)' P_o q_i(\theta_o)\right]/n \\ &\quad -\frac{1}{2}B_o\sum_{j=1}^{k} E\left[\frac{\partial q_i(\theta_o)}{\partial\theta_j\partial\theta}\right] S_o e_j/n \end{aligned}$$

as required.

B.3 Proof to Theorem 7.3

Let $G_i(\theta_o) = G_i$ and $g_i(\theta_o) = g_i$. The first therm of the bias of Theorem 7.2 can be written as

$$B_o E[z_i G_i B_o z_i g_i] \tag{49}$$

Noting that $G_i B_o z_i$ is a scalar, we can write (49) as

$$B_o E[z_i z_i' B_o' G_i' g_i]$$

By the law of iterated expectation

$$B_o E[z_i z_i' B_o' G_i' g_i] = B_o E[z_i z_i'] B_o' \sigma_{gG}$$

and hence

$$\begin{aligned} E[z_i z_i'] B_o' \sigma_{gG} &= E[z_i G_i] \tilde{S}_o \sigma_{gG} / \sigma_g^2 \\ B_o E[z_i z_i'] B_o' \sigma_{gG} &= B_o E[z_i G_i] \tilde{S}_o \sigma_{gG} / \sigma_g^2 \\ &= \tilde{S}_o \sigma_{gG} / \sigma_g^2 \qquad (50) \end{aligned}$$

where $\tilde{S}_o = E[G_i(\theta_o)' z_i'] E[z_i z_i']^{-1} E[z_i G_i(\theta_o)]$. For the second term, using the law of iterated expactations,

$$E[z_i z_i' P_o z_i g_i^3] = E[z_i z_i' P_o z_i] \sigma_g^3 \qquad (51)$$

Substituting (50) and (51) into the formala for the bias gives the desired result.

B.4 Proof to Theorem 7.4

When the first order conditions are modified, the expactations of the terms in the expansions hold with the exception of the expectation in (48), because it involves $\beta_{jr}^{\theta\theta}$. Multiplying the Lagrange multiplier rescale the derivatives by κ

$$\beta_j^{\lambda\lambda} = \kappa E\left[\nabla_\theta q_t(\theta_o)' e_j q_i(\theta_o)' + q_{t,j} \nabla_\theta q_t(\theta_o)'\right]$$

Hence, the expectation in (48) becomes

$$E[S_o \nabla_1] = 2\kappa S_o E\left[\nabla_\theta q_i(\theta_o)' P_o q_i(\theta_o)\right]$$

For the instrumental variable case the expectation above is given by

$$\begin{aligned} 2\kappa E\left[G_i(\theta_o)' z_i' P_o z_i g_i(\theta_o)\right] &= 2\kappa \sigma_{gG}^2 E[z_i' P_o z_i] \\ &= 2\kappa \sigma_{gG}^2 E[Trace\{z_i' P_o z_i\}] \\ &= 2\kappa \sigma_{gG}^2 E[Trace\{P_o z_i z_i'\}] \end{aligned}$$

Notice that

$$\begin{aligned} P_o E[z_i z_i'] &= V_o^{-1} E[z_i z_i'] - V_o^{-1} \Gamma_o S_o \Gamma_o V_o^{-1} E[z_i z_i'] \\ &= \sigma_g^{-2}(I_m - V_o^{-1}\Gamma_o S_o \Gamma_o) \end{aligned}$$

Using the properties fo the $Trace$, it follows that $Trace\{P_o z_i z_i'\} = Trace\{\sigma_g^{-2}(I_m - V_o^{-1}\Gamma_o S_o \Gamma_o)\} = \sigma_g^{-2}(m-k)$, and $2\kappa S_o E\left[G_i(\theta_o)' z_i' P_o z_i g_i(\theta_o)\right] = 2\kappa \tilde{S}_o \sigma_{gG}^2 (m-k)\sigma_g^{-2}$. Substituting the terms in the expansion gives

$$\begin{aligned} B_{-1}(\theta_o) &= \tilde{S}_o \sigma_{gG} \sigma_g^{-2}/n \\ &\quad - \tilde{S}_o \sigma_{gG}^2 \sigma_g^{-2}(m-k)/n \\ &\quad + B_o E[z_i z_i' P_o z_i]\sigma_g^3/n \\ &\quad + \kappa \tilde{S}_o \sigma_{gG}^2 \sigma_g^{-2}(m-k)/n \\ &\quad - \frac{1}{2}\gamma_3 B_o E[z_i z_i' P_o z_i]\sigma_g^3/n \\ &\quad - \frac{1}{2} B_o \sum_{j=1}^{k} E\left[z_i \frac{\partial g_i(\theta_o)}{\partial\theta\partial\theta_j}\right] S_o e_j/n \end{aligned}$$

Collecting terms gives

$$\begin{aligned} B_{-1}(\theta_o) &= \tilde{S}_o \sigma_{gG} \sigma_g^{-2}(\kappa(m-k) - (m-k-1)) \\ &\quad + (1 - \frac{\gamma_3}{2}) B_o E[z_i z_i' P_o z_i]\sigma_g^3/n \\ &\quad - \frac{1}{2} B_o \sum_{j=1}^{k} E\left[z_i \frac{\partial g_i(\theta_o)}{\partial\theta\partial\theta_j}\right] S_o e_j/n \end{aligned}$$

as required.

B.5 Proof to Theorem 7.5

Let $B_{-1}(\hat{\theta}_{el}) - B_{-1}$. Since by assumption $(\hat{\theta}_{el} - \theta_o) = u_n + b_n + r_n$, noting that $E(u_n B_{-1}') = 0$ and dropping terms $o(n^{-2})$, we have

$$\begin{aligned} \mathcal{M}_{-2}(\hat{\theta}_{el} - B_{-1}(\hat{\theta}_{el})) &= E(\hat{\theta}_{el} - B_{-1} - \theta_o)(\hat{\theta}_{el} - B_{-1} - \theta_o)' \\ &= E(u_n + b_n + r_n - B_{-1} - \theta_o)(u_n + b_n + r_n - B_{-1} - \theta_o)' \\ &= E(u_n u_n' + r_n u_n' + u_n r_n' + b_n b_n' + B_{-1} B_{-1}{}' - b_n B_{-1}' - B_{-1} b_n') \end{aligned}$$

By the definition of $O(n^{-1})$ bias, $E(b_n B'_{-1}) = B_{-1}B'_{-1}$ and the result follows.

B.6 Proof to Theorem 7.6

The validity of the expansion can be proved along the lines of Theorem 7.2 by verifying the conditions of Lemma B.3. By Assumption D the CLT can be applied to the elements of the Jacobian giving $\|J_n - J_o\| = O_p(n^{-1/2})$. Similarly, the elements of the matrix collecting the second derivatives is also bounded in probability of order $n^{-1/2}$, by the same argument given in the proof of Theorem 4.3. By Assumption D, the third derivatives are bounded by $\frac{1}{n}\sum_i^n \mathcal{B}(w_i)^5\|\tau - \tau_o\|$ and Lemma B.3 holds with $\bar{\mathcal{C}}(w_i) = \mathcal{B}(w_i)^5$.

Application of Lemma B3 gives the first conclusion of the theorem, where the expansion for $(\hat{\theta} - \theta_o)$ and $\hat{\lambda}$ are given, respectively, by the first k and the last m elements of

$$\begin{aligned}(\hat{\tau} - \tau_o) &= -Q_o m_n + Q_o Z_n Q_o m_n \\ &\quad -\frac{1}{2}Q_o H_o \left\{[Q_o m_n \otimes Q_o m_n] \otimes \left[+Q_o \tilde{V}_m(\tau_0) Q_o \bar{m}_T(\tau_0)\right]\right\} \\ &\quad -\frac{1}{6}Q_o D_o \left\{Q_o m_n \otimes Q_o m_n \otimes Q_o m_n\right\}\end{aligned}$$

By inspection of the matrix collecting the second derivatives of m_n, it follows that differences in the expansions of two MD estimators with $\gamma_3 = 2$ appear only in the term

$$\Delta_n \equiv -\frac{1}{6}Q_o D_o \left\{Q_o m_n \otimes Q_o m_n \otimes Q_o m_n\right\} \tag{52}$$

Let$\beta_{jr}^{\theta\theta\theta} = E\left[\frac{\partial[m_i]_{1,k}(\tau_o)}{\partial\theta_j\partial\theta_r\partial\theta}\right]$; $\mu_{jr}^{\theta\theta\theta} = E\left[\frac{\partial[m_i]_{h,m}(\tau_o)}{\partial\theta_j\partial\theta_r\partial\theta}\right] and \beta_{jr}^{\theta\theta\lambda} = E\left[\frac{\partial[m_i]_{1,k}(\tau_o)}{\partial\theta_j\partial\theta_r\partial\lambda}\right]$; $\mu_{jr}^{\theta\theta\lambda} = E\left[\frac{\partial[m_i]_{h,m}(\tau_o)}{\partial\theta_j\partial\theta_r\partial\lambda}\right]$

so on for the other cross partial derivatives. The first k elements of Δ_n are given by

$$\begin{aligned}\Delta_n^{\theta} = - \quad & \frac{1}{6}\Bigg\{ -B_o\sum_{j=1}^{k}\sum_{r=1}^{k}\beta_{jr}^{\theta\theta\theta}u_{n,j}u_{n,r}u_n + S_o\sum_{j=1}^{k}\sum_{r=1}^{k}\mu_{jr}^{\theta\theta\theta}u_{n,j}u_{n,j}u_n \\ & -B_o\sum_{j=1}^{k}\sum_{r=1}^{k}\beta_{jr}^{\theta\theta\lambda}u_{n,j}u_{n,r}l_n + S_o\sum_{j=1}^{k}\sum_{r=1}^{k}\mu_{jr}^{\theta\theta\lambda}u_{n,j}u_{n,r}l \\ & -B_o\sum_{j=1}^{k}\sum_{r=1}^{m}\beta_{jr}^{\theta\lambda\theta}u_{n,j}l_{n,r}u_n + S_o\sum_{j=1}^{k}\sum_{r=1}^{m}\mu_{jr}^{\theta\lambda\theta}u_{n,j}l_{n,r}u_n \\ & -B_o\sum_{j=1}^{k}\sum_{r=1}^{m}\beta_{jr}^{\theta\lambda\lambda}u_{n,j}l_{n,r}l_n + S_o\sum_{j=1}^{k}\sum_{r=1}^{m}\mu_{jr}^{\theta\lambda\lambda}u_{n,j}l_{n,r}l_n \\ & -B_o\sum_{j=1}^{k}\sum_{r=1}^{k}\beta_{jr}^{\lambda\theta\theta}l_{n,j}u_{n,r}u_n + S_o\sum_{j=1}^{k}\sum_{r=1}^{k}\mu_{jr}^{\lambda\theta\theta}l_{n,j}u_{n,r}u_n \\ & -B_o\sum_{j=1}^{m}\sum_{r=1}^{k}\beta_{jt}^{\lambda\theta\lambda}l_{n,j}u_{n,r}l_n + S_o\sum_{j=1}^{m}\sum_{r=1}^{k}\mu_{jt}^{\lambda\theta\lambda}l_{n,j}u_{n,r}l_n \\ & -B_o\sum_{j=1}^{m}\sum_{r=1}^{m}\beta_{jr}^{\lambda\lambda\lambda}l_{n,j}l_{n,r}l_n + S_o\sum_{r=1}^{m}\sum_{j=1}^{m}\mu_{jr}^{\lambda\lambda\lambda}l_{n,j}l_{n,r}l_n \Bigg\}\end{aligned}$$

Inspecting the expected value of the third derivatives of of $m_i(\tau_o)$ shows that EL and any MD with $\gamma_3 = 2$ have the same partial derivatives with exception of $\mu_{jr}^{\lambda\lambda\lambda}$, $j, r = 1, \ldots, m$. This implies that for any MD estimators with $\gamma_3 = 2$

$$\Delta_n^{\theta} = -\frac{1}{6}S_o\sum_{r=1}^{m}\sum_{j=1}^{m}\mu_{jr}^{\lambda\lambda\lambda}l_{n,j}l_{n,r}l_n$$

Let $\hat{\theta}$ denote the EL estimator and $\tilde{\theta}$ denote any MD estimator with $\gamma_3 = 2$. The difference of the MSE for EL and any other MD estimators with $\gamma_3 = 2$ is then given by

$$\begin{aligned} MSE(\hat{\theta}) - MSE(\tilde{\theta}) &= E\left[(\Delta_{n,el}^{\theta} - \Delta_{n,umd}^{\theta})u_n'\right] \\ &\quad +E\left[u_n(\Delta_{n,el}^{\theta'} - \Delta_{n,umd}^{\theta'})\right]\end{aligned}$$

where $\Delta_{n,el}^{\theta}$ and $\Delta_{n,umd}^{\theta}$ denote the differences in the expansion for EL and an any MD estimator with $\gamma_3 = 2$, respectively. Thus, the expression of the difference between the

MSE errors above reduces to

$$
\begin{aligned}
MSE(\hat{\theta}) - MSE(\tilde{\theta}) &= \frac{1}{6} S_o \sum_{r=1}^{m} \sum_{j=1}^{m} \left[\tilde{\mu}_{jr}^{\lambda\lambda\lambda} - \hat{\mu}_{jr}^{\lambda\lambda\lambda} \right] E(l_{n,j} l_{n,r} l_n u_n') \qquad (53) \\
&\quad + \frac{1}{6} \sum_{r=1}^{m} \sum_{j=1}^{m} \left[\tilde{\mu}_{jr}^{\lambda\lambda\lambda} - \hat{\mu}_{jr}^{\lambda\lambda\lambda} \right] E(l_{n,j} l_{n,r} u_n l_n') S_o'
\end{aligned}
$$

Noting that $\hat{\mu}_{jr}^{\lambda\lambda\lambda} = -\gamma_4 E(q_{i,j} q_{i,r} q_i q_i')$ and the for EL $\gamma_4 = 6$ gives

$$
\begin{aligned}
MSE(\hat{\theta}) - MSE(\tilde{\theta}) &= \left(1 - \frac{\gamma_4}{6}\right) S_o \sum_{r=1}^{m} \sum_{j=1}^{m} \left[E(q_{i,j} q_{i,r} q_i q_i') \right] E(l_{n,j} l_{n,r} l_n u_n') \\
&\quad + \left(1 - \frac{\gamma_4}{6}\right) \sum_{r=1}^{m} \sum_{j=1}^{m} \left[E(q_{i,j} q_{i,r} q_i q_i') \right] E(l_{n,j} l_{n,r} u_n l_n') S_o'
\end{aligned}
$$

as required.

B.7 Proof to Theorem 7.7

Expanding the summations in expressions (**??**) in the proof of Theorem 9.3 yields

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}) - \mathrm{MSE}(\tilde{\theta}) &= \left(1 - \frac{\gamma_4}{6}\right) \frac{1}{n^4} S_o \sum_{j=1}^{m} \sum_{r=1}^{m} q_{jr}^4 E(\sum_{\nu=1}^{n} \sum_{\varsigma=1}^{n} \sum_{\alpha=1}^{n} \sum_{\eta=1}^{n} l_{\nu,j} l_{\varsigma,r} l_\alpha u_\eta') \\
&\quad + \left(1 - \frac{\gamma_4}{6}\right) \frac{1}{n^4} \sum_{j=1}^{m} \sum_{r=1}^{m} q_{jr}^4 E(\sum_{\nu=1}^{n} \sum_{\varsigma=1}^{n} \sum_{\alpha=1}^{n} \sum_{\eta=1}^{n} l_{\nu,j} l_{\varsigma,r} u_\alpha l_\eta') S_o
\end{aligned}
$$

Consider the first term, since the second is the transpose of the first one. For $\nu = \varsigma = \alpha = \eta = i$, ($n$ times), it follows from Assumption D that $n^{-3} E(l_{i,j} l_{i,r} l_i u_i') = O(n^{-3})$. When three indexes are equal by independence and since $E(l_i) = E(u_i) = 0$, $E(t_\eta l_\alpha' l_{j,\nu} l_{r,\varsigma}) = 0$. Thus the summation can reduced to

$$
\begin{aligned}
\frac{1}{n^4} \sum_{\nu=1}^{n} \sum_{\varsigma=1}^{n} \sum_{\alpha=1}^{n} \sum_{\eta=1}^{n} l_{\nu,j} l_{\varsigma,r} u_\alpha l_\eta' &= n^{-2} \left[E(l_{1,j} l_{1,r}) E(u_2 l_2') + E(l_{2,j} l_{2,j}) E(u_1 l_1') \right. \\
&\quad \left. + E(l_{1,j} u_1) E(l_2' l_{2,r}) + E(l_{1,r} u_1) E(l_2' l_{2,j}) \right] \\
&\quad + o(n^{-2})
\end{aligned}
$$

The results then follows by noting that the oll the expectation involving l_i and u_i are zero, since by orthogonality of $P_o V_o B_o' = 0$,

$$\begin{aligned} E(l_{i,r} u_i') &= e_r' E(P_o q_i(\theta_o) q_i(\theta_o)' B_o') \\ &= e_r' E(P_o V_o B_o') = 0 \end{aligned}$$

Noting that the same holds for the expectations involved in the transpose, the result follows.