PROPERTIES OF MINIMUM DIVERGENCE ESTIMATORS

BY GIUSEPPE RAGUSA

ABSTRACT. This paper shows that estimators defined in terms of Minimum Divergence are very general and indeed there is a one-to-one relationship between the Minimum Divergence (MD) and Generalized Empirical Likelihood (GEL) class of estimators. Newey and Smith (2004) show that in the GEL class, the Empirical Likelihood estimator can be singled out for having the smallest bias and being third order efficient. We show that all the estimators in the MD/GEL class that have the same asymptotic bias as the Empirical Likelihood estimator are third order equivalent, having the same Mean Square Error of order $O(n^{-2})$. A new estimator is suggested that is third order efficient and has bounded influence function. Boundedness of the influence function implies that the new estimator is well behaved under misspecification of the moment conditions.

1. INTRODUCTION

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

Date: This Version: October, 2005. First Version: April, 2003
Giuseppe Ragusa, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA, gragusa@rci.rutgers.edu.

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.

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. In a seminal paper, Newey and Smith (2004), NS henceforth, consider a generalization of EL, ET and CUE based on a saddle point problem. They refer to the estimators arising from this generalization as Generalized Empirical Likelihood (GEL) estimators.

This paper considers estimation 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. This paper makes the following main contributions. First, it proves that there is a relationship between the Generalized Empirical Likelihood (GEL) class of estimators –as defined by Newey and Smith (2004)– and the MD class. Every MD estimator has a GEL representation. Also, from given GEL estimator a MD problem can be given such that the two problem are equivalent. This result is very important for a series of reasons. It allows considering estimators that extend beyond the class defined by the Cressie-Read power discrepancy without abandoning the nice probabilistic features underpinning the MD class of estimators.

Comparisons between GMM and MD estimators 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 among 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.

The plan of the paper is as follows. Section 2 defines moment conditions models. Section 3 reviews the existing alternative to GMM estimation and testing framework. Section 4 briefly presents the existing alternatives to GMM, while Section 4 defines the Minimum Divergence class of estimators and establishes the duality results. Section 5 reviews the first order asymptotic properties. Section 6 introduces the higher order expansions. In Section 7 the bias of MD estimator is derived and a nice bias correction for Instrumental Variables models is discussed. Section 8 introduces third order efficient estimators whose influence function is bounded. Section 9 presents a numerical simulation. Section 10 concludes.

2. The model

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 i.i.d. observations on a data vector w with unknown probability distribution F_o. Also, let θ be a $k \times 1$ parameter vector and $q(w, \theta)$ be an $m \times 1$ vector of functions of the data observation w and the parameter θ, where $m \geq k$. The model consists of the following moment condition

$$E[q(w, \theta_o)] = \int q(w, \theta_o) \mathrm{d}F_o = 0 \tag{2.1}$$

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 (2.1) 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 (2.1) is extremely general and it is very common in many fields of economics. Simultaneous system of equations, dynamic panel data and many other models frequently encounter in economic have a econometric formulation equivalent to (2.1).

Empirical content is given to (2.1) by considering its sample counterpart

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

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

$$q_n(\theta) = 0 \tag{2.2}$$

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 (2.1) is overidentified and the existence of a solutions satisfying the the empirical restrictions is not guaranteed. A important estimator of (2.1) when $m > k$ is the Generalized Method of Moment (GMM) of Hansen (1982) which basic idea 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. The GMM estimator solves the following optimization step

$$\min_{\theta\in\Theta} Q_n(\theta, W_n) \tag{2.3}$$

where

$$Q_n(\theta, W_n) = n q_n(\theta)' W_n q_n(\theta)$$

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. Let $\tilde{\theta}$ denote a preliminary consistent estimator of θ_o. The efficient GMM estimator is defined as

$$\hat{\theta}^{\text{gmm}} = \arg\min_{\theta\in\Theta} Q_n(\theta, V(\tilde{\theta}))$$

where

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

3. Existing Alternatives to 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

$$\hat{\theta} = \left\{ \arg\max_{\pi,\theta} \frac{1}{n} \sum_{i=1}^{n} \log \pi_i \mid s.t. \sum_{i=1}^{n} \pi_i q_i(\theta) = 0, \ \sum_{i=1}^{n} \pi_i = 1, \ \pi_i > 0 \right\} \tag{3.1}$$

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

$$\hat{\theta} = \left\{ \arg\min_{\pi,\theta} \sum_{i=1}^{n} \pi_i \log \pi_i \mid s.t. \sum_{i=1}^{n} \pi_i q_i(\theta) = 0, \ \sum_{i=1}^{n} \pi_i = 1, \ \pi_i > 0 \right\} \tag{3.2}$$

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 (3.2) 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

$$\hat{\theta} = \left\{ \arg\min_{\pi,\theta} n \sum_{i=1}^{n} \pi_i^2 \mid s.t. \sum_{i=1}^{n} \pi_i q_i(\theta) = 0, \ \sum_{i=1}^{n} \pi_i = 1 \right\} \tag{3.3}$$

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. Give a matrix

A, let A^{-g} denote the Moore-Penrose inverse of A. The CUE estimator minimizes

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

NS show that $\arg\min_{\theta\in\Theta} \widetilde{Q}_n(\theta)$ is numerically equivalent to the estimator obtained by solving (3.3). 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 defined on an interval $\mathcal{V}$ that contains zero and $\Lambda_n(\theta) = \{\lambda | \lambda' q_i(\theta) \in \mathcal{V}, i = 1, \dots n\}$.[1] 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 $\psi(x) = \exp(x)$, and to CUE for $\psi(x) = x^2/2$.

4. 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

$$\hat{\theta} = \left\{ \arg\min_{\pi,\theta} \frac{1}{n} \sum_{i=1}^{n} \gamma(n\pi_i) \,|\, s.t. \sum_{i=1}^{n} \pi_i q(w_i, \theta) = 0, \;\; \sum_{i=1}^{n} \pi_i = 1, \; \pi_i \in (a_\gamma, b_\gamma) \right\} \tag{4.1}$$

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, that is $\gamma_r \equiv \gamma_r(1)$. Throughout the paper $\gamma(\cdot)$ will denote a function that satisfies the following requirements:

[1] In their specification they consider a problem defined 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.

Assumption 1. (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 $a_\gamma = 0$ 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 $\hat{\pi}_i$, provided that there exists a unique minimizer $\hat{\theta}$. Suppose $\hat{\pi}(\hat{\theta}) = (\hat{\pi}_1(\hat{\theta}), \hat{\pi}_2(\hat{\theta}), \ldots, \hat{\pi}_n(\hat{\theta}))$ and $\tilde{\pi}(\hat{\theta}) = (\tilde{\pi}_1(\hat{\theta}), \tilde{\pi}_2(\hat{\theta}), \ldots, \tilde{\pi}_n(\hat{\theta}))$ are both solution to 4.1. Then, for any $0 \leq \zeta \leq 1$, $\pi^\zeta(\hat{\theta}) = \zeta\hat{\pi}(\hat{\theta}) + (1-\zeta)\tilde{\pi}(\hat{\theta})$ is a feasible solution. But if $\gamma(\cdot)$ is strictly convex, $\sum_i^n \gamma(n\pi_i^\zeta(\hat{\theta})) < \zeta \sum_i^n \gamma(n\hat{\pi}_i(\hat{\theta})) + (1-\zeta)\sum_i^n \gamma(n\tilde{\pi}_i(\hat{\theta}))$, that is a contradiction, since by assumption $\hat{\pi}$ and $\tilde{\pi}$ are both solutions. The conditions on the the second derivative of $\gamma(\cdot)$ are normalizations and do not restrict the class of functions that may be considered as divergence. Such normalizations play a crucial roles in the derivations of the properties of the estimators.

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 (4.1) in pinning down optimal weights $\pi = (\pi_1, \ldots, \pi_n)$ 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 moment conditions $\pi_i \approx n^{-1}$ as $n \to \infty$ and $\gamma(1) = 0$, the shape of the divergence does not determine the (first order) 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 F_o in some meaningful

sense. Formally, the stochastic model for the random vector $w = (w_1, w_2, ..., w_n)$ is defined as

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

$$\mathcal{G}(\theta) = \left\{ G : \int q(w, \theta) \mathrm{d}G = 0 \right\} \tag{4.2}$$

For a given γ define the following functional

$$I_\gamma(R, G) = \begin{cases} \int \gamma\left(\frac{\mathrm{d}F_o}{\mathrm{d}G}\right) \mathrm{d}F_o & if\ G \ll F_o \\ +\infty & otherwise \end{cases} \tag{4.3}$$

The functional $I_\gamma(R, G)$, that is broadly speaking the population counterpart of (4.1), 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 $I_\gamma(G, F_o)$, formally $\inf_{G \in \mathcal{G}} I_\gamma(G, F_o)$. If the model is correctly specified (i.e. $F_o \in \mathcal{G}$) then clearly $F_o = \inf_{G \in \mathcal{G}} I_\gamma(G, F_o)$. Similarly, if $E_{F_o}[q(w, \theta)] \neq 0$ for $\theta \neq \theta_o$, F_o implicitly identifies θ_o. Note that under fairly weak conditions, when the model conditions are misspecified (i.e. $F_o \notin \mathcal{G}$), the solutions to $F_* = \inf_{G \in \mathcal{G}} I_\gamma(G, F_o)$ 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 (4.1): 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 F_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.

4.1. **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 (4.1) 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 $i=1,\ldots,n$, the following

$$\gamma_1(n\pi_i)-\lambda' q_i(\theta)-\eta=0 \tag{4.4}$$

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

$$\sum_{i=1}^{n}\pi_i\nabla_\theta q_i(\theta)'\lambda=0 \tag{4.5}$$

The Lagrange multiplier η in (4.4) can be eliminated as follow. Multiplying (4.4) 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 (4.4) 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

$$\pi_i = \frac{1}{n}\widetilde{\gamma}_1(\lambda' q_i(\theta)) \tag{4.6}$$

where here $\widetilde{\gamma}_1(\cdot)$ denotes the inverse function of $\gamma_1(\cdot)$. By substituting (4.6) into the constraint $\sum_{i=1}^n \pi_i q_i(\theta) = 0$ and into (4.5), the following first order conditions are obtained

$$\frac{1}{n}\sum_{i=1}^n \widetilde{\gamma}_1(\lambda' q_i(\theta)) q_i(\theta) = 0 \tag{4.7}$$

$$\frac{1}{n}\sum_{i=1}^n \widetilde{\gamma}_1(\lambda' q_i(\theta)) \nabla_\theta q_i(\theta)' \lambda = 0 \tag{4.8}$$

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 (4.7) and (4.8), the normalized weights $\{\omega_1, \omega_2, \ldots, \omega_n\}$ still solve (4.7) and (4.8) and, by construction, $\sum_{i=1}^n \omega_i = 1$.

Remark 4. The Lagrangian multiplier need 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 elimination of the Lagrange multiplier η is handy for it allows studying the asymptotic properties of the resulting estimators from a unified perspective.

The first order conditions (4.7) and (4.8) 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 $\omega_i = \exp(\lambda' q_i(\theta))/\sum_{i=1}^n \exp(\lambda' q_i(\theta))$ 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

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

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 divergences 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 1. *If the MD problem given in* (4.1) *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 1 is given in the Appendix. Theorem 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. Notice that this version of the duality relies on the first order conditions and simply establishes that MD and GEL solves the same set of first first order conditions. This requires that *a)* the moment function is differentiable; *b)* the solutions $\hat{\theta}$ and $\hat{\pi}$ are interior.

Example 1. [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 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 is instead is explicitly considered by slightly changing the assumption of Theorem 1. In particular, if $\sum_{i=1}^{n} \tilde{\gamma}_2(\hat{\eta} + \hat{\lambda}' q_i(\hat{\theta})) q_i(\hat{\theta}) q_i(\hat{\theta})'$ is non singular, the MD first order condition coincide with the first order conditions of the problem

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

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

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

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 1 is unavailable. Although Theorem 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. (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 2. *Consider the GEL problem given by*

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

where the function $\psi(\cdot)$ satisfies Assumption 2 and $\Lambda_n(\theta) = \{\lambda : \lambda' q_i(\theta) \in \mathcal{V}, i = 1, ..., n\}$. 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 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 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. However, Theorem 1 and 2 are based on equality of the first order conditions and they make assumption both on the invertibility of $\sum_i^n \tilde{\gamma}_2(\lambda' q_i(\theta)) q_i(\theta) q_i(\theta)'$ and $\sum_i^n \psi_1(\lambda' q_i(\theta)) q_i(\theta) q_i(\theta)'$ at the solution and on the differentiability of the moment function. The following result is a generalization of the duality and establishes that the solution of given a MD problem is the solution of a corresponding GEL problem, without invertibility and differentiability requirement.

Theorem 3. *Let $(\hat{\theta}, \hat{\lambda})$ denote the solution for the GEL problem for some $\psi(\cdot)$, satisfying Assumption 2 and $\hat{\lambda}' q_i(\hat{\theta}) \in \mathcal{V}$, $i = 1, \ldots, n$ Then $(\hat{\theta}, \{\hat{\omega}_i\}_i^n)$, where $\hat{\omega}_i = \psi_1(\hat{\lambda}' q_i(\hat{\theta}) / \sum_i^n \psi_1(\hat{\lambda}' q_i(\hat{\theta}))$, is solution for the corresponding MD problem with $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$.*

Theorem refth:equivgeneral extend the equality in situations where the moment conditions is non-differentiable as is the case for quantile restrictions on the moment. Notice that the above result does not assume nor require that the solution of the GEL is unique. Indeed, the function $f(\theta) = \min_\lambda \sum_i^n \psi(\lambda q_i(\theta)$ need not be convex in θ and uniqueness of the solution cannot be guaranteed.

Theorem 4. *Let $(\hat{\theta}, \{\hat{\pi}_i\}_{i=1}^n)$ denote the solution for the MD problem for some $\gamma(\cdot)$ satisfying Assumption 1, $\hat{\omega}_i = \tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta})) / \sum_{i=1}^n \tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta}))$ and $\hat{\lambda}' q_i(\hat{\theta}) \in \mathcal{A}$. Then $(\hat{\theta}, \hat{\lambda})$ solve the corresponding GEL problem with $\psi(x) = x\tilde{\gamma}_1(x) - \gamma(\tilde{\gamma}(x))$.*

The duality extends also to the value of the objectives functions $\gamma(\cdot)$ and $\psi(\cdot)$, as the following result shows.

Corollary 1. *If $\gamma(\cdot)$ satisfies Assumption 1, the value of the objective function evaluated at $\hat{\pi}_i = \tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta}))/n$ is equivalent to the negative value of the corresponding GEL objective function evaluated at $\hat{\theta}$ and $\hat{\lambda}$, that is*

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

When the objective function is evaluated at the normalized value of the weights $\hat{\omega}_i = \tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta})) / \sum_i \tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta}))$, the following disequality

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

5. 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 3. *(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 are derived under the assumption $\alpha = 2$.

Theorem 5. *Let Assumption 3 hold. Then (i) the solutions of the constrained optimization problem* (4.1)*,* $(\hat{\pi}, \hat{\theta})$*, exist with probability approaching 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 4. *(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 6. *Let Assumptions 3-4 hold. Then the sequence of solutions* $\hat{\theta}$ *and the vector of Lagrange multiplier* $\hat{\lambda}$ *are asymptotically normal and independent with*

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

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

Theorem 6 makes clear that MD and GMM 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 $F_o(w) = \int 1[w_i \leq w] \mathrm{d}F_o$ and $\sigma_w = F_o(w)(1 - F_o(w))$. Under the conditions of Theorem 6, the empirical distribution function given by $\mathbb{F}_n(w) = n^{-1} \sum_i^n 1[w_i \leq w]$ converges point-wise almost surely to the underlying distribution function of w and has limiting normal distribution described by

$$\sqrt{n}(\mathbb{F}_n(w) - F_o(w)) \xrightarrow{d} \mathscr{N}(0, \sigma_w)$$

Let $\mathbb{M}_n(w) = \sum_i^n 1[w_i \leq w]\hat{\omega}_i$, where $\hat{\omega} = (\hat{\omega}_1, \hat{\omega}_2, \ldots, \hat{\omega}_n)$ are the normalized MD weights.

Theorem 7. *Suppose Assumptions 3-4 hold. Then* $\mathbb{M}_n(w) - F_o(w) \xrightarrow{p} 0$ *point-wise and* $\mathbb{M}_n(w)$ *has limiting normal distribution given by*

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

where $q(w) = \int 1[w_i \leq w] q_i(w_i, \theta_o) \mathrm{d}F_o$.

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 $E[q(w_i, \theta_o)] = 0$. Theorem 7 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.

6. 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

$$(\hat{\theta} - \theta_o) = u_n + b_n + r_n + O_p(n^{-2}) \tag{6.1}$$

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 Hahn, Hausman, and Kuersteiner (2001a) and Hahn, Hausman, and Kuersteiner (2001b), among others.

Definition 1. [Higher Order Bias] If an estimator $\hat{\theta}$ of θ_o admits an expansion as in (6.1), 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 2. [Higher Order Mean Square Error] If an estimator $\hat{\theta}$ of θ_o admits an expansion as in (6.1), 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 the expectation of the corresponding terms in the expansion (6.1) 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 possesses 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. 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 (6.1) are not simple functions of w and smoothness assumptions must be imposed on $q(w, \theta)$.

The following notation is used. The partial derivatives with respect to θ_j, $j = 1, \ldots, k$ 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, $j, r = 1, \ldots, k$ 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 5. 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 $j, r, s = 1, \ldots, k$: *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 8. *Suppose Assumptions 1-5 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 8 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 Rilstone, 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 8 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 9. *Suppose Assumptions 1-5 hold. Then the asymptotic bias up to order $O_p(n^{-1})$ for a MD estimator of θ_o is given by*

$$B_{-1}(\hat{\theta}) = n^{-1}\left\{b_1 + (1 - \frac{\gamma_3}{2})b_2\right\} \tag{7.1}$$

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 9 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 1, it follows that, under Assumption 5, $\psi_3(0) = \gamma_3(1)$. 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 9. 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 estimators of $b_1(\theta_o)$, $b_2(\theta_o)$. It follows that the bias corrected MD estimator defined as

$$\hat{\theta}_{bc} = \hat{\theta} - n^{-1}\left\{\hat{b}_1 + (1 - \frac{\gamma_3}{2})\hat{b}_2\right\} \tag{7.2}$$

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

7.1. **Instrumental Variables Model.** 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

$$q_i(w_i, \theta) = z_i g(x_i, \theta) \tag{7.3}$$

where $g(x_i, \theta) : \mathscr{X} \times \mathbb{R}^k \to \mathbb{R}$ 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 6. (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 5 and Assumption 7 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 10. *Under Assumption 6 the $O(n^{-1})$ bias of the MD estimator is given by*

$$B_{-1}(\theta_o) = \tilde{S}_o \sigma_{gG}/\sigma_g^2/n - \tilde{B}_o \tilde{a}/n + (1 - \frac{\gamma_3}{2})\sigma_g^3 \sigma_z^3/n \tag{7.4}$$

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

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

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

The expression for the bias in (7.4) specializes immediately to the bias of the homoschedastic linear IV as given by

$$B_{-1}(\hat{\theta}) = -\tilde{S}_o E(x_i\varepsilon_i|z_i)/\sigma^2 + (1-\gamma_3/2)\sigma_g^3\tilde{B}_o\sigma_z^3$$

When $\gamma_3 = 2$, as mentioned in NS, is the bias of the Limited Information Maximum Likelihood estimator. Now consider the estimator that solves the following equations

$$0 = \frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_1(\lambda' z_i g_i(\theta))z_i g(x_i,\theta) \tag{7.5}$$

$$0 = \frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_1(\kappa\lambda' z_i g_i(\theta))G_i(\theta)' z_i'\lambda' \tag{7.6}$$

for some $\kappa > 0$. For $\kappa = 1$, these estimating equations are equivalent to those of MD estimators given in (4.7) and (4.8).

Theorem 11. *Let $\hat{\theta}$ and $\hat{\lambda}$ be the solution of the estimating equations in* (7.5) *and* (7.6)*. Then $\hat{\theta} = \theta_o + o_p(1)$ and has asymptotic bias given by*

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

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 (7.5) and (7.6) 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 (7.5) and (7.6) 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 (7.5) and (7.6).

8. 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 verifying 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 12. *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 12 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 7. 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 13. *Suppose Assumptions 1-7 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

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

where $q_{jr}^4(\theta_o) = E\left[q_{n,j}(\theta_o) q_{n,r}(\theta_o) q_i(\theta_o) q_i(\theta_o)'\right]$ 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 13 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 14. *Suppose Assumptions 1-2 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 (7.5)-(7.6) is third order efficient having the same $O(n^{-2})$ MSE than EL.

9. 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 improve the finite sample performance of MD estimators.

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

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

Heuristically, 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 14 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 solutions of the dual problem

$$\max_{\theta\in\Theta}\min_{\lambda\in\Lambda_n(\theta)}\frac{1}{n}\sum_{i=1}^{n}\psi(\lambda' q_i(\theta)) \tag{9.1}$$

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 2, the estimator associated with problem (9.1) corresponds to a MD estimator defined from a strictly convex divergence. Constructing a MD estimator with bounded IF in λ is equivalent to finding 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 $\psi(\upsilon)=\exp(h(\upsilon))$, 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 $\psi_1(\upsilon)=\exp(h(\upsilon))h_1(\upsilon)$. 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 (9.1) to have the same bias of EL, it must hold that

$$\left.\frac{\partial^2\exp(h(\upsilon))h_1(\upsilon)}{\partial^2\upsilon}\right|_{\upsilon=0}=2 \tag{9.2}$$

Assumption 2(ψ) also requires that $h_1(0)=1$ and $h_2(0)=1$. By expanding the derivative in (9.2), 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 (9.1) with $\psi(x)=[\exp(\sinh(x))-1]$ as Hyperbolic Tilting (HT) by analogy with the Exponential Tilting from which it originates.

Theorem 15. *[Hyperbolic Tilting] The Hyperbolic Tilting (HT) estimator is defined as the solution of the following problem*

$$\max_{\theta\in\Theta}\min_{\lambda}\frac{1}{n}\sum_{i=1}^{n}[\exp(\sinh(\lambda' q_i(\theta))) - 1] \tag{9.3}$$

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.

Unfortunately, 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 the simulations in Imbens, Spady, and Johnson (1998), ET seems to have nice finite sample properties in terms of size of resulting statistics for testing overidentified restrictions and restrictions on the parameters.

10. Numerical Examples

This section provides some simulation evidences on the performances of MD estimators. Three main designs are considered. In order to verify that the Hyperbolic Tilting estimator behaves well under misspecification, relative to Empirical Likelihood, Monte Carlo simulations are performed using a simple model for the mean and the variance of a normal distribution where one of the equation is potentially misspecified. In the second design, the performances of EL, ET, HT and GMM estimators are studied by using the experimental design of Hall and Horowitz(1996). The last design considers the performance of the estimators when applied to the linear instrumental variables model. For this last model the performance of the bias correction proposed in Section 2.5.1 are also assessed.

10.1. **Misspecified Models.**

The simple model is considered for the moment function

$$q(w_i,\theta) = \begin{bmatrix} w_i - \theta \\ (w_i-\theta)^2 - 1 \end{bmatrix} \tag{10.1}$$

In each Monte Carlo replication w is drawn from a normal distribution. Two models are considered. For the first model, $w \sim N(0,1)$. In this case the moment function is correctly specified, $E[q(w,\theta_o)] = 0$ for $\theta_o = 0$. For the second model $w \sim N(0, 0.64)$ in which case the moment function is misspecified, since $\|E[q(w,\theta)]\| > 0$ for any

θ. The estimator considered are the Empirical Likelihood, the Exponential Tilting, the Hyperbolic Tilting, the Two-Step GMM and the Iterated GMM estimator. The simulations are carried out for sample sizes $n = 1000, 5000$.

Table 1-2 report the result of the simulations for both models. Clearly, at the sample sizes considered here, the differences in the sampling distributions of EL, ET, HT, GMM and GMM10 are practically nonexistent for the correctly specified model. Notice, in particular, that the sampling variance of the estimator is equal to the asymptotic approximation value of 0.001 for $n = 1000$ and 0.0002 for $n = 5000$. The situation changes dramatically when we consider the misspecified model. The ratio of the sampling variances at $n = 1000$ and $n = 5000$ is for ET, HT, GMM and GMM10 very close to 1/5, the value predicted by a *sqrtn* asymptotic approximation to the distribution of the estimators. The behavior of EL is not in line with a *sqrtn* consistent estimator. The variance goes from 0.0030 to 0.0027, well above the 1/5 ratio. The difficulties of EL in dealing with misspecification are highlighted by Figure 10.1-10.2. Figure 10.1 plots the sampling density of EL,ET,HT for correctly specified model. As predictable, the difference are between these distributions are undetectable. Figure 10.2 plots the sampling distribution for EL, ET, HT under the misspecified model. The sampling distribution of EL shows clear signs of non-normality. Also, the degree of non-normality tends to increase with the sample size. A behavior clearly inconsistent with a $\sqrt{n}$ consistent estimator. Notice also that the sampling distribution of the HT estimator has a sampling distribution that is very close to the asymptotic approximation that would hold in absence of misspecification.

TABLE 1. Monte Carlo simulations: Normal model under correct specification

	Var	Mean	Median	Mse	Mad	Iqr
			$n = 1000$			
HT	0.0010	0.0003	0.0005	0.0010	0.0010	0.0421
ET	0.0010	0.0003	0.0005	0.0010	0.0010	0.0420
EL	0.0010	0.0003	0.0005	0.0010	0.0010	0.0420
GMM2	0.0010	0.0003	0.0005	0.0010	0.0010	0.0421
GMM10	0.0010	0.0003	0.0005	0.0010	0.0010	0.0421
			$n = 5000$			
HT	0.0002	−0.0002	−0.0004	0.0002	0.0002	0.0194
ET	0.0002	−0.0002	−0.0004	0.0002	0.0002	0.0194
EL	0.0002	−0.0002	−0.0004	0.0002	0.0002	0.0194
GMM2	0.0002	−0.0002	−0.0004	0.0002	0.0002	0.0194
GMM10	0.0002	−0.0002	−0.0004	0.0002	0.0002	0.0194

TABLE 2. Monte Carlo simulations: Normal model under misspecification

	Var	Mean	Median	Mse	Mad	Iqr
			$n = 1000$			
HT	0.0010	0.0005	0.0007	0.0010	0.0010	0.0426
ET	0.0014	0.0003	0.0003	0.0014	0.0015	0.0521
EL	0.0030	−0.0000	−0.0003	0.0030	0.0039	0.0844
GMM2	0.0015	0.0006	0.0013	0.0015	0.0015	0.0515
GMM10	0.0009	0.0005	0.0005	0.0009	0.0009	0.0410
			$n = 5000$			
HT	0.0002	−0.0002	−0.0005	0.0002	0.0002	0.0193
ET	0.0005	−0.0004	−0.0006	0.0005	0.0005	0.0288
EL	0.0027	−0.0007	−0.0015	0.0027	0.0052	0.0973
GMM2	0.0003	−0.0001	−0.0002	0.0003	0.0003	0.0234
GMM10	0.0002	−0.0002	−0.0004	0.0002	0.0002	0.0183

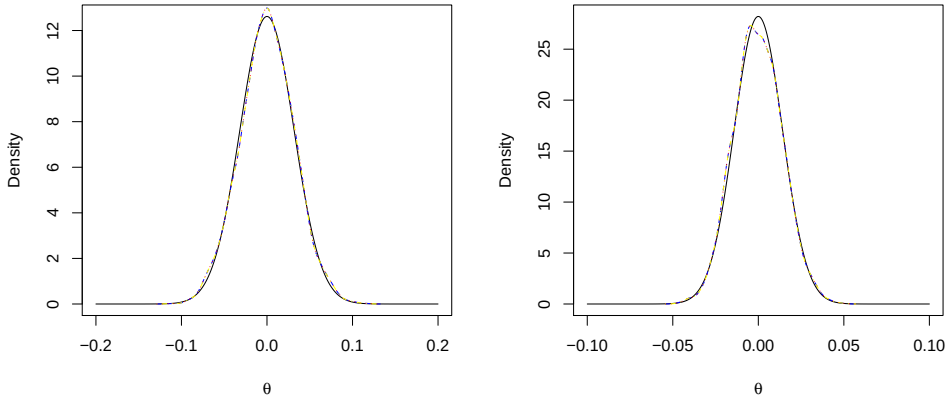

FIGURE 10.1. (Normal Model) Sampling distribution of HT (dashed), ET (dotted), EL (dotdash) and normal asymptotic approximation (solid line) under correct specification, for $n = 1000$ (left plot) and $n = 5000$ (right plot)

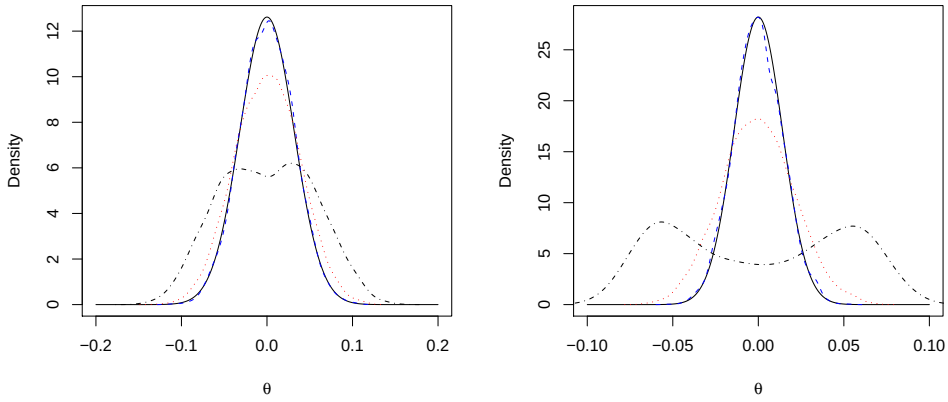

FIGURE 10.2. (Normal Model) Sampling distribution of HT (dashed), ET (dotted), EL (dotdash) and normal asymptotic approximation (solid line) under correct specification, for $n = 1000$ (left plot) and $n = 5000$ (right plot)

References

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

Hahn, J., J. A. Hausman, and G. Kuersteiner (2001a): "Bias Corrected Instrumental Variables Estimation for Dynamic Panel Models with Fixed Effects," Mimeo.

——— (2001b): "Higher order MSE of jacknife 2SLS," .

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

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.

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.

Judge, G., and R. Mittelhammer (2001): "Empirical evidence concerning the finite sample performance of EL-type structural equation estimators," .

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

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

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.

Owen, A. B. (1990): "Empirical likelihood ratio confidence regions," *The Annals of Statistics*, 18, 90–120.

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

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.

Rockafellar, T. R. (1970): *Convex Analysis*. Princeton University Press.

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

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

Proof to Theorem 1. Since, by assumption, the solutions of the MD problem are interior we have that $\hat{\lambda}'q(w_i,\hat{\theta}) \in \mathcal{A}$, 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

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

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

$$\frac{1}{n} \sum_{i=1}^{n} \tilde{\gamma}_1(\lambda' q_i(\theta)) q_i(\hat{\theta}) = 0 \tag{.2}$$

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 (.2) 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 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

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

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 $\tilde{\psi}_1(\cdot)$

$$\pi_i = \frac{1}{n} \psi_1(\lambda' q_i(\theta)) \tag{.3}$$

Substituting (.3) into $\sum_{i=1}^{n} \pi_i q_i(\theta) = 0$ gives

$$\sum_{i=1}^{n} \psi_1(\lambda' q_i(\theta)) q_i(\theta) = 0$$

and the result follows.

Proof to Theorem 3. Since by assumption $(\hat{\lambda}, \hat{\theta})$ are interior solutions, it follows that $\hat{\lambda}' q_i(\hat{\theta}) \in \mathcal{V}$. The normalized GEL weights are feasible for the MD problem, since $\{\hat{\pi}_i\}_{i=1}^{n}$ solve $\sum_{i=1}^{n} \hat{\pi}_i q_i(\hat{\theta}) = 0$ and $\sum_{i=1}^{n} \hat{\pi}_i = 1$. Let us show that $\hat{\pi}_i$ is optimal. Under the assumptions made on $\psi(\cdot)$, the Fenchel inequality applies (see Rockafellar (1970)),

$$\psi(s) = s\tilde{\gamma}_1(s) - \gamma(\tilde{\gamma}_1(s)) \geq st - \gamma(t)$$

for all $t \in (a_\gamma, b_\gamma)$ and for all $s \in \mathcal{V}$. Seting $s = \hat{\lambda}' q_i(\hat{\theta})$ and $t = p(w_i, \hat{\theta})$, where $p(w_i, \hat{\theta})$ denotes any feasible solutions, the above inequality can be written as

$$\gamma(\tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta})) - \hat{\lambda}' q_i(\hat{\theta}) \tilde{\gamma}_1(\hat{\lambda}' q_i(\hat{\theta}) \leq \gamma(p(w_i, \hat{\theta})) - \hat{\lambda}' q_i(\hat{\theta}) p(w_i, \hat{\theta}) \tag{.4}$$

Feasibility of $p(w_i,\hat{\theta})$ implies that it must satisfies the constraint given by $\sum_{i=1}^n p(w,\hat{\theta})q_i(\hat{\theta}) = 0$. Summing (.4) over $i = 1,\ldots,n$, we obtain

$$\sum_{i=1}^n \gamma(\tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))) \leq \sum_{i=1}^n \gamma(p(w_i,\theta))$$

By strictly convexity of $\gamma(\cdot)$ on (a_γ, b_γ) it follows that for all $x, y \in (a_\gamma, b_\gamma)$, $\gamma(y) > \gamma(x) + \gamma_1(x)(y-x)$, that in turns implies that

$$\begin{aligned}\sum_{i=1}^n \gamma\left(n\frac{\tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))}{\sum_{i=1}^n \tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))}\right) &\leq \sum_{i=1}^n \gamma(\tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))) \\ &\quad + \sum_{i=1}^n \gamma_1(\tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta})))\left(n\frac{\tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))}{\sum_{i=1}^n \tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))} - p(w_i,\theta)\right) \\ &= \sum_{i=1}^n \gamma(\tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta})))\end{aligned}$$

where the equality comes from the fact the$\gamma_1(\tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta}))) = \hat{\lambda}'q_i(\hat{\theta})$ and $\sum_i \hat{\lambda}'q_i(\hat{\theta})\tilde{\gamma}_1(\hat{\lambda}'q_i(\hat{\theta})) = \sum_i \hat{\lambda}'q_i(\hat{\theta})\psi_1(\hat{\lambda}'q_i(\hat{\theta})) = 0$, since $\tilde{\gamma}_1(x) = \psi_1(x)$ and by assumption $\hat{\lambda}$ solve the GEL first order condition. We also need to prove that $\hat{\theta}$ from the GEL is optimal for the MD problem. Let $\tilde{\theta} \in \Theta$ be any other MD feasible estimator. Feasibility here means that $\sum_{i=1}^n \tilde{\gamma}_1(\hat{\lambda}'q_i(\tilde{\theta}))q_i(\tilde{\theta}) = 0$. It holds that

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

Notice that the weak inequality is necessary because we are not assuming that the solutions in $\hat{\theta}$ is unique. Using $\psi(s) = s\tilde{\gamma}_1(s) - \gamma(\tilde{\gamma}_1(s))$ and the feasibility of $\tilde{\theta}$ in the MD problem

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

And the result follows.

Proof to Corollary 1. The results 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. 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 6. 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 7. 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 (2.1) 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 $\Delta_o(w)^{-1} = \tilde{Q}_o - q_o(w) V_o^{-1} q_o(w)$.

Appendix A. Asymptotic Expansions

This Appendix provides proofs for the theorems concerning higher order properties given in Section 6.

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 $j, r = 1, \ldots, g$, 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 vector 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, $[x]_{1,\ldots,k}$ denotes the first k elements of x.

Lemmas.

Lemma 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 2. *Suppose (i)* $\| \hat{\tau} - \tau_o \| = O_p(n^{-1/2})$*; (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* $j, r = 1, \ldots, g$*,*

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

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 3. *Suppose (i)* $\| \hat{\tau} - \tau_o \| = O_p(n^{-1/2})$*; (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* $j, r, p = 1, \ldots, g$*,*

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

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

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*

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

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

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

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 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 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

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

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

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

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 (A.1) 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

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

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

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

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

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

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

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

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 1 and collecting terms of similar order, we obtain

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

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

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

$$+ O_p(n^{-2})$$

as required.

Proof to Theorem 8. 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

$$Q_o = \begin{bmatrix} -S_o & B_o \\ B_o' & P_o \end{bmatrix}$$

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

$$J_n(\tau_o) = \begin{bmatrix} 0 & \frac{1}{n}\sum_i^n \nabla_\theta q_i(\theta_o)' \\ \frac{1}{n}\sum_i^n \nabla_\theta q(w,\theta_o) & \frac{1}{n}\sum_i^n q_i(\theta_o) q_i(\theta_o)' \end{bmatrix}$$

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

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

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

$$Q_o q_n = \begin{pmatrix} -B_o q_n \\ -P_o q_n \end{pmatrix} = \begin{pmatrix} u_n \\ l_n \end{pmatrix} \tag{A.3}$$

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 (A.3) 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-th elements of l_n and u_n respectively. Thus, the first k elements of the expansion for $(\hat{\tau} - \tau_o)$ are given by

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

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

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

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.

Proof to Theorem 9. 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) $E\left[S_o\Gamma_n l_n\right]$

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

(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

$$E[S_o \nabla_1] = 2 S_o E\left[\nabla_\theta q_i(\theta_o)' P_o q_i(\theta_o)\right] \tag{A.4}$$

(*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

$$\mu_j^{\theta\theta} = E\left[\frac{\partial q_i(\theta_o)}{\partial \theta_j \partial \theta}\right]; \;\; \mu_j^{\lambda\lambda} = \gamma_3 E[q_{i,j}(\theta_o) q_i(\theta_o) q_i(\theta_o)]$$

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

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

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

$$\begin{aligned}
\sum_{j=1}^{m} \mu_j^{\lambda\lambda} E(l_{n,j}l_n) &= \sum_{j=1}^{m} \gamma_3 E[q_{i,j}(\theta_o)q_i(\theta_o)q_i(\theta_o)']P_oe_j/n \\
&= \sum_{j=1}^{m} \gamma_3 E[q_i(\theta_o)q_i(\theta_o)'P_oe_jq_{i,j}(\theta_o)]/n \\
&= \gamma_3 E\left[q_i(\theta_o)q_i(\theta_o)\left(\sum_{j=1}^{m} P_oe_je_j'\right)q_i(\theta_o)\right]/n \\
&= \gamma_3 E\left[q_i(\theta_o)q_i(\theta_o)P_oq_i(\theta_o)\right]/n
\end{aligned}$$

and hence

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

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

$$\begin{aligned}
E(b_n) &= -E[B_o\Gamma_nu_n] + E[S_o\Gamma_n'u_n] - E[B_oV_nl_n] \\
&\quad + \frac{1}{2}S_oE(\nabla_1) - \frac{1}{2}B_oE(\nabla_2)
\end{aligned}$$

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

simplifying and rearranging yields

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

as required.

Proof to Theorem 10. Let $G_i(\theta_o) = G_i$ and $g_i(\theta_o) = g_i$. The first therm of the bias of Theorem 9 can be written as

$$B_oE[z_iG_iB_oz_ig_i] \tag{A.5}$$

Noting that $G_iB_oz_i$ is a scalar, we can write (A.5) as

$$B_oE[z_iz_i'B_o'G_i'g_i]$$

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

$$\begin{aligned} E[z_i z_i'] B_o' \sigma_{gG} &= E[z_i G_i] \tilde{S}_o \sigma_{gG} / \sigma_g^2 \\ B_o E[z_i z_i'] B_o' \sigma_{gG} &= B_o E[z_i G_i] \tilde{S}_o \sigma_{gG} / \sigma_g^2 \\ &= \tilde{S}_o \sigma_{gG} / \sigma_g^2 \end{aligned} \tag{A.6}$$

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 expectations,

$$E[z_i z_i' P_o z_i g_i^3] = E[z_i z_i' P_o z_i] \sigma_g^3 \tag{A.7}$$

Substituting (A.6) and (A.7) into the formula for the bias gives the desired result.

Proof to Theorem 11. When the first order conditions are modified, the expectations of the terms in the expansions hold with the exception of the expectation in (A.4), because it involves $\beta_{jr}^{\theta\theta}$. Multiplying the Lagrange multiplier re-scale 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 (A.4) 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 for 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

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

as required.

Proof to Theorem 12. Let $B_{-1}(\hat{\theta}_{el})-B_{-1}$. Since by assumption $(\hat{\theta}_{el}-\theta_o)=u_n+b_n+r_n$, noting that $E(u_nB_{-1}')=0$ and dropping terms $o(n^{-2})$, we have

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

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

Proof to Theorem 13. The validity of the expansion can be proved along the lines of Theorem 9 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

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

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

$$\Delta_n\equiv-\frac{1}{6}Q_oD_o\left\{Q_om_n\otimes Q_om_n\otimes Q_om_n\right\} \tag{A.8}$$

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

$$\begin{aligned} MSE(\hat{\theta}) - MSE(\tilde{\theta}) &= \frac{1}{6} S_o \sum_{r=1}^{m}\sum_{j=1}^{m} \left[\tilde{\mu}_{jr}^{\lambda\lambda\lambda} - \hat{\mu}_{jr}^{\lambda\lambda\lambda}\right] E(l_{n,j}l_{n,r}l_n u_n') \\ &\quad + \frac{1}{6} \sum_{r=1}^{m}\sum_{j=1}^{m} \left[\tilde{\mu}_{jr}^{\lambda\lambda\lambda} - \hat{\mu}_{jr}^{\lambda\lambda\lambda}\right] E(l_{n,j}l_{n,r}u_n l_n')S_o' \end{aligned} \tag{A.9}$$

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.

Proof to Theorem 14. Expanding the summations in expressions (A.9) 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 all 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.

Rutgers University