Minimum Divergence, Generalized Empirical Likelihoods, and Higher Order Expansions*

Giuseppe Ragusa †

June 15, 2010

Abstract

This paper studies the Minimum Divergence (MD) class of estimators for econometric models specified through moment restrictions. We show that MD estimators can be obtained as solutions to a tractable lower dimensional optimization problem. This problem is similar to the one solved by the Generalized Empirical Likelihood estimators of Newey and Smith (2004), but it is equivalent to it only for a subclass of divergences. The MD framework provides a coherent testing theory: tests for overidentification and parametric restrictions in this framework can be interpreted as semiparametric versions of Pearson-type goodness of fit tests. The higher order properties of MD estimators are also studied and it is shown that MD estimators that have the same higher order bias as the Empirical Likelihood (EL) estimator also share the same higher order Mean Square Error and are all higher order efficient. We identify members of the MD class that are not only higher order efficient, but also, unlike the EL estimator, well behaved when the moment restrictions are misspecified.

JEL Classification: C12, C13, C23

Keywords: Minimum Divergence, GMM, Generalized Empirical Likelihood, Higher Order Efficiency, Misspecified Models.

*Parts of this paper circulated under a different title in the past. Very helpful comments on previous drafts were received from Graham Elliott, Hal White, Roger Gordon, Eli Berman, Francesca Mazzolari, Seth Pruitt and seminar participants at UCSD, UCI, Rutgers, University of Washington, UBC, Princeton, and Boston College. I also thanks Esfandiar Maasoumi (the editor), the associated editor, and two anonymous referees for very helpful suggestions.

†Dipartimento di Scienze Economiche e Aziendali, LUISS Guido Carli, Viale Romania 32, 00197 Rome. email: gragusa@luiss.it

1 Introduction and motivations

Econometric models are often postulated in terms of moment restrictions:

$$\int q(w, \theta_0) F(\mathrm{d}w) = 0, \tag{1}$$

where $w \in \mathcal{W} \subseteq \mathbb{R}^L$ is a random vector with unknown probability distribution F, and $q(w, \theta)$ is an $M \times 1$ vector of functions of w and the parameter $\theta \in \Theta \subset \mathbb{R}^K$, $q : \mathcal{W} \times \Theta \mapsto \mathbb{R}^M$. Given a random sample from w, $(w_1, \ldots, w_N)$, the objective is to estimate θ_0. Simultaneous systems of equations, dynamic panel data, and many other models frequently employed in econometrics have a formulation equivalent to (1).

The traditional way of estimating θ_0 is by the Generalized Method of Moments (GMM) of Hansen (1982). GMM estimators are consistent and asymptotically normal in a broad array of setups (see, among others, Gallant and White (1988) and Newey and McFadden (1994)). Despite GMM's desirable asymptotic properties and limited computational requirements, there has been increasing concern over its performance in applications. A vast literature documents that inference based on GMM has unsatisfying finite sample performance (see the articles in the 1996 special issue of the Journal of Business and Economic Statistics).

New estimators have been proposed that tend to perform better than GMM in some settings. The Continuous Updating Estimator (CUE) of Hansen et al. (1996), the Empirical Likelihood (EL) estimator of Qin and Lawless (1994) and Imbens (1997), and the Exponential Tilting (ET) of Kitamura and Stutzer (1997) are three of the most known examples.

Hansen et al. (1996) show through Monte Carlo simulations that CUE is nearly median unbiased. Simulations in Imbens (2002) suggest that EL and ET estimators have lower bias than GMM in nonlinear models. Mittelhammer et al. (2005) find that EL has lower bias than two-stage least squares in linear structural models. Imbens et al. (1998) present Monte Carlo evidence on the performance of the overidentification test statistics based on EL, ET and CUE, and find them to have lower size distortion than corresponding GMM statistics. Kitamura (2001) shows that EL is optimal in terms of large deviations for testing overidentifying restrictions. EL has been adapted to a wide array of settings. Notably, Guggenberger and Smith (2005) explore the behavior of EL in the weak instrumental variables scenario. Kitamura et al. (2004) apply EL to models defined through smooth conditional moment restrictions. Both Otsu (2008) and Whang (2006) apply EL to the estimation of parameters identified by conditional quantile restrictions.

Newey and Smith (2004) (NS henceforth) study the theoretical properties of EL, ET, CUE by embedding them into the Generalized Empirical Likelihood (GEL) class of

estimators. They show that all GEL estimators have lower asymptotic bias than GMM. In particular, EL has the smallest $O(N^{-1})$ bias. The bias corrected EL is also second order efficient in the sense of Pfanzagl and Wefelmeyer (1979), suggesting that EL is a preferable member of the GEL class under the higher order bias/efficiency criterion.

This paper studies the properties of the Minimum Divergence (MD) class of estimators for parameters satisfying moment restrictions like (1). First, we show that MD estimators can be obtained as the solution to a saddle point problem whose criterion function is very similar to the one that GEL estimators optimize. However, the MD framework encompasses the GEL: using convex analysis arguments, we derive the condition under which the GEL and MD estimators coincide. Second, we show that the equivalence between MD estimators and solution to an optimization problem is complete: not only any MD estimator can be interpreted as solving a saddle point problem for a given criterion function; for any criterion function and corresponding saddle point problem, there exists an underlying MD problem whose solution is the same as the one to the saddle point problem.

The MD class of estimators is not new and it has a long history that precedes the generalization suggested by the Empirical Likelihood estimator. Its roots can be traced back to the Maximum Entropy (ME) principle introduced by Jaynes (see, Jaynes, 1984, 2003). In the econometric literature, the Maximum Entropy principle when moment restrictions are present was studied by Golan et al. (1996). For an excellent review of applications of the ME principle see Golan (2008). Also related to MD is the generalized minimum contrast class of estimators proposed by Pfanzagl (1979) and studied in detail by Bickel (1998, Ch. 7). Corcoran (1998) analyzes the performances of goodness of fit tests based on minimizing divergences when the moment function does not depend on a parameter. The idea of using duality for the ME dates back to Agmon et al. (1979) (see also Golan et al. (1996); Golan (2002, 2008) and references therein). Kitamura (2007) discusses duality for MD. Broniatowski and Keziou (2004b) and Broniatowski and Keziou (2004a) discuss the relationship between Empirical Likelihood type estimators and measures of divergences with emphasis on duality results also relevant for moment condition models.

Expressing the estimation problem in terms of divergence minimization is particularly appealing from a statistical point of view, since it provides a framework to understand the analogy between testing theory developed for parametric models and testing theory appropriate to the semiparametric setting considered here. We show that overidentification test statistics based on the saddle point criterion functions are semiparametric versions of Pearson-type goodness of fit tests.

We also study the higher order properties of MD estimators. We show that MD

estimators with the same $O(N^{-1})$ bias as EL also share the same higher order Mean Square Error (MSE). This result implies that these MD estimators are as efficient as the EL up to order $O(N^{-2})$, and that there are many higher order efficient MD estimators with competitive $O(N^{-1})$ bias.

Since higher order considerations alone are not sufficient for selecting a member of the MD class of estimators to be used in applications, we turn to misspecification robustness as an additional criterion. Results in Schennach (2007) suggest that if the moment restrictions are misspecified, the EL estimator may be ill-behaved and may not be $\sqrt{N}$-consistent. Schennach (2007) proposes a new estimator that is a combination of EL and ET, which while having the same higher order properties of EL, is well behaved under misspecification. The existence of higher order efficient estimators in the MD class distinct from the EL estimator allows us to identify estimation procedures that are higher order efficient, behave well under misspecification, and are based on minimization of divergences.

A word on notation. If A is a matrix, $\|A\| = \sqrt{\text{Tr}\, AA'}$ denotes its Frobenious norm. This reduces to the usual Euclidean norm when A is a vector. Throughout the paper, vectors are columns unless transposed. Random vectors and their realizations are denoted by lower case letters. All limits are taken as $N \to \infty$. The qualifiers "with probability one" and "with probability approaching one" are abbreviated as "w.p.1" and "w.p.a.1", respectively. The symbols O_p and o_p are the stochastic order symbols. Finally, the following notation for functions and their derivatives is used. If f is a function $f : \mathbb{R} \mapsto \mathbb{R}$, $f_r(x) := d^r f(x)/d^r x$, for all $r = 1, 2, \ldots$ for which f is differentiable. If the inverse function of f is defined, we set $\tilde{f}(x) := f^{-1}(x)$; similarly, for the inverse of the derivatives of f, we set $\tilde{f}_r(x) := f_r^{-1}(x)$.

2 Minimum Divergence estimators

Given a random sample of size N $(w_1, \ldots w_N)$, a Minimum Divergence estimator for the parameter vector θ_0 that satisfies (1) is the solution to

$$\min_{\theta\in\Theta,\pi_1,\ldots,\pi_N} \sum_{i=1}^{N} \gamma(N\pi_i)/N, \quad \gamma \in \mathcal{G},$$
$$s.t. \quad \sum_{i=1}^{N} \pi_i q(w_i, \theta) = 0, \ \sum_{i=1}^{N} \pi_i = 1, \ N\pi_i \in D_\gamma, \tag{2}$$

where $\mathcal{G}$ denotes the class of convex and twice continuously differentiable divergence functions, $\gamma : D_\gamma \subseteq \mathbb{R} \mapsto \mathbb{R}_+$ with D_γ convex and $\text{int}(D_\gamma) = (a_\gamma, b_\gamma)$, $a_\gamma < 1 < b_\gamma$;

$\gamma(1) = 0$, $\gamma_1(1) = 0$, $\gamma_2(x) > 0$ for $x \in (a_\gamma, b_\gamma)$, $\gamma_2(1) = 1$. A strictly positive second derivative in the interior of the domain implies that any $\gamma \in \mathcal{G}$ is strictly convex. Note that $\gamma(1) = 0$ and $\gamma_2(1) = 1$ are normalizations that are not restrictive. Let $\rho : D_\rho \subseteq \mathbb{R} \mapsto \mathbb{R}_+$ be convex and twice continuously differentiable on the convex set D_ρ, $\rho_2(x) > 0$, for $x \in \text{int}(D_\rho)$. If ρ does not satisfy the normalizations, the function $\bar{\rho}(x) := \rho(x)/\rho_2(1) - x\rho(1)/\rho_2(1) - \rho(0)/\rho_2(1)$ will and $\bar{\rho} \in \mathcal{G}$.

The MD problem in (2) defines a collection of estimators indexed by γ ranging in $\mathcal{G}$. Notably, it encompasses the EL estimator, for $\gamma^{el}(x) = -\ln x + x - 1$, the ET, for $\gamma^{et}(x) = x \ln x - x + 1$, the CUE, for $\gamma^{cue}(x) = x^2/2 - x + .5$, and estimators based on the Cressie-Read family of divergences (Cressie and Read, 1984), for $\gamma^{cr}(x; \alpha) = \frac{x^{\alpha+1}-1}{\alpha(\alpha+1)} - \frac{1}{\alpha}x + \frac{1}{\alpha}$, $\alpha \neq \{0, -1\}$.[1]

The Fisher consistency of the MD procedure can be shown heuristically as follows. The function $\sum_{i=1}^N \gamma(N\pi_i)/N$ is minimized when $\pi_i = N^{-1}$, $(i = 1, \ldots, N)$. From all the feasible vectors $(\pi_1, \ldots, \pi_N)$ and parameters $\theta \in \Theta$, the MD problem will select a θ that gives a weighting that is the closest to assigning N^{-1} to each sample point. As $N \to \infty$, the moment restrictions in (1) imply that $\theta \approx \theta_0$ and $\pi_i \approx N^{-1}$ will solve (2). Intuitively, since $\gamma(1) = 0$ for all $\gamma \in \mathcal{G}$, the specific member of $\mathcal{G}$ used in the procedure does not determine the first order asymptotic behavior of the estimator; features of γ in a neighborhood of 1 do, however, determine the finite sample properties of the estimator.

Remark 1. In the exactly identified case, that is, $M = K$, if there exists a $\dot{\theta} \in \Theta$ such that $\sum_{i=1}^N q(w_i, \dot{\theta})/N = 0$, then the MD estimator of θ_0 is $\dot{\theta}$ and the optimal weights are given by $\pi_i = N^{-1}$ $(i = 1, \ldots N)$. Thus, in this case, the MD estimator coincides with the Method of Moment estimator.

Remark 2. Problem (2) is feasible if the set $C(\theta) = \{y_i \in D_\gamma,\, i \leqslant N : \sum_{i=1}^N y_i q_i(w_i, \theta) = 0\}$ is non empty for at least some $\theta \in \Theta$. If $a_\gamma < 0$, then the problem is always feasible, that is, $C(\theta)$ is non empty for all $\theta \in \Theta$. If $a_\gamma = 0$ or $a_\gamma > 0$, then, for a given sample of N observations on w, the set $C(\theta)$ may be empty for all $\theta \in \Theta$.

Remark 3. The solution to the MD problem is not in general unique in θ. Strict convexity of γ does, however, imply that the optimal π_i's are unique. Suppose that $\dot{\theta}, \ddot{\theta} \in \Theta$ both minimize (2), that is $\sum_{i=1}^N \gamma(N\pi_i(\dot{\theta}))/N = \sum_{i=1}^N \gamma(N\pi_i(\ddot{\theta}))/N$, where $\pi_i(\dot{\theta})$ and $\pi_i(\ddot{\theta})$, $(i = 1, \ldots, N)$, denote the optimal weights that correspond to $\dot{\theta}$ and $\ddot{\theta}$. We have that $\bar{\pi}_i := \zeta\pi_i(\dot{\theta}) + (1 - \zeta)\pi_i(\ddot{\theta})$ is feasible for any $0 \leq \zeta \leq 1$. However, strict convexity of γ implies that $\sum_{i=1}^N \gamma(N\bar{\pi}_i) < \zeta \sum_{i=1}^N \gamma(N\pi_i(\dot{\theta})) + (1 - \zeta) \sum_{i=1}^N \gamma(N\pi_i(\ddot{\theta}))$, which is a contradiction. Thus, $\pi_i(\dot{\theta}) = \pi_i(\ddot{\theta})$ $(i = 1, \ldots, N)$.

[1] It should be noted that for some values of α, γ^{cr} is not (strictly) convex everywhere on its domain. In these cases, we restrict γ^{cr} to be defined on the largest convex interval containing 1 on which γ^{cr} is strictly convex. For instance, for $\alpha = 2$, γ^{cr} is strictly convex on $(0, +\infty)$, so we consider $\gamma^{cr}(\cdot, 2) : D_\gamma \mapsto \mathbb{R}$, $D_\gamma = [0, +\infty)$.

2.1 First order conditions

In the overidentified case, $M > K$, the solution to (2) can, under some conditions, be obtained through the method of Lagrange multipliers. The Lagrangian of the constrained optimization problem is

$$\mathcal{L}(\theta, \pi, \eta, \lambda) = \sum_{i=1}^{N} \gamma(N\pi_i)/N - \lambda' \sum_{i=1}^{N} \pi_i q(w_i, \theta) - \eta\big(\sum_{i=1}^{N} \pi_i - 1\big),$$

where $\lambda \in \mathbb{R}^M$ and $\eta \in \mathbb{R}$ are the Lagrange multipliers associated with the two constraints. If the moment function $q_i(\theta) := q(w_i, \theta)$ is differentiable on Θ, an interior solution to (2) must set to zero the partial derivatives of $\mathcal{L}(\theta, \pi, \eta, \lambda)$. Let $G_i(\theta) = \partial q_i(\theta)/\partial\theta$. The partial derivatives of $\mathcal{L}(\theta, \pi, \eta, \lambda)$ with respect to θ and π are, respectively,

$$\sum_{i=1}^{N} \pi_i G_i(\theta)'\lambda = 0; \quad \gamma_1(N\pi_i) - \lambda' q_i(\theta) - \eta = 0 \ \ (i = 1, \ldots, N).$$

By twice continuous differentiability of γ on D_γ, and strict positivity of γ_2 on D_γ, γ_1 is monotone on D_γ. Let $\mathcal{A}_\gamma = \{y : \gamma_1(x) = y,\, x \in D_\gamma\}$ be the image of the first derivative of γ and

$$\Lambda_N(\theta) = \big\{(\eta, \lambda') \in \mathbb{R}^{M+1} : \eta + \lambda' q_i(\theta) \in \mathcal{A}_\gamma, \text{ for all } i \leqslant N\big\}.$$

For any $(\eta, \lambda') \in \Lambda_N(\theta)$, we can invert the first order condition $\gamma_1(N\pi_i) - \lambda' q_i(\theta) - \eta = 0$ to obtain that $\pi_i = \tilde{\gamma}_1(\eta + \lambda' q_i(\theta))/N$ $(i = 1, \ldots, N)$. Replacing this expression for the weights into the constraints, we have that, for a given $\theta \in \Theta$, if there exists $(\eta, \lambda') \in \Lambda_N(\theta)$ solving the equations

$$\sum_{i=1}^{N} \tilde{\gamma}_1(\eta + \lambda' q_i(\theta)) q_i(\theta)/N = 0, \quad \sum_{i=1}^{N} \tilde{\gamma}_1(\eta + \lambda' q_i(\theta))/N = 1,$$

then the optimal π_i's must take the form $\pi_i(\theta) = \tilde{\gamma}_1(\eta + \lambda' q_i(\theta))/N$. When optimizing over $\theta \in \Theta$, if $q_i(\theta)$ is differentiable on Θ, the first order condition for θ must be taken into account. So, if there exists $\hat{\theta} \in \text{int}(\Theta)$ and $(\hat{\eta}, \hat{\lambda}') \in \Lambda_N(\hat{\theta})$ such that

$$\begin{aligned}
&\sum_{i=1}^{N} \tilde{\gamma}_1(\hat{\eta} + \hat{\lambda}' q_i(\hat{\theta}))/N = 1, \quad \sum_{i=1}^{N} \tilde{\gamma}_1(\hat{\eta} + \hat{\lambda}' q_i(\hat{\theta})) q_i(\hat{\theta})/N = 0, \\
&\sum_{i=1}^{N} \tilde{\gamma}_1(\hat{\eta} + \hat{\lambda}' q_i(\hat{\theta})) G_i(\hat{\theta})'\hat{\lambda}/N = 0,
\end{aligned} \tag{3}$$

then $\hat{\theta}$ and $\pi_i(\hat{\theta}) = \tilde{\gamma}_1(\hat{\eta}+\hat{\lambda}'q_i(\hat{\theta}))/N$ $(i = 1, \ldots, N)$ solve the MD problem. Note that the set $\mathcal{A}_\gamma$ determines whether the optimal solution can be attained by Lagrangian techniques. If the image of the derivative of the divergence is the real line, i.e. $\mathcal{A}_\gamma = \{y : -\infty < y < +\infty\}$, all $(\eta, \lambda') \in \mathbb{R}^{M+1}$ are in $\Lambda_N(\hat{\theta})$ and the only requirement is that $(\theta', \eta, \lambda')$ solves (3).

From a statistical point of view, the first order conditions in (3) could be used to estimate θ_0 (Imbens, 1997). If θ_0 is the unique solution to (1), the system of equations has a unique solution w.p.a.1, $(\theta', \eta, \lambda') = (\theta_0', 0, 0)$, and it can be shown that the root of (3) is a consistent and asymptotically normal distributed estimator of θ_0. There are however problems in using (3) directly for estimation. For instance, the inverse of the first derivative of γ may not have a expression for some $\gamma \in \mathcal{G}$. Even if $\tilde{\gamma}_1$ has a closed-form expression, $q(\cdot, \theta)$ may not be differentiable on Θ. Also, computing MD estimators as solutions to (2) leaves open the possibility that the equations in (3) have multiple roots even if (2) has a unique minimum.[2]

2.2 Duality

An alternative to working with the first order conditions (3) is working directly with the extremum problem in (2). However, the constrained optimization problem involves solving for $N + K$ variables, and it becomes computationally challenging even for small N. We show here that the MD problem can be re-casted in terms of an attractive saddle point problem in $M + K + 1$ variables.

Let $\mathcal{P}$ denote the class of convex and twice continuously differentiable functions, $\psi : D_\psi \subseteq \mathbb{R} \mapsto \mathbb{R}_+$ with D_ψ convex and $\text{int}(D_\psi) = (a_\psi, b_\psi)$, $a_\psi < 0 < b_\psi$; $\psi(0) = 0$, $\psi_1(0) = \psi_2(0) = 1$, $\psi_2(x) > 0$ for $x \in D_\psi$. Consider the following saddle point problem

$$\sup_{\theta\in\Theta} \min_{(\eta,\lambda)\in\Lambda_N^\dagger(\theta)} P_N(\eta, \lambda, \theta), \quad P_N(\eta, \lambda, \theta) = \sum_{i=1}^{N} \psi(\eta + \lambda' q_i(\theta))/N - \eta, \ \psi \in \mathcal{P}, \tag{4}$$

where

$$\Lambda_N^\dagger(\theta) = \big\{(\eta, \lambda') \in \mathbb{R}^{M+1} : \eta + \lambda' q_i(\theta) \in D_\psi, \text{ for all } i \leqslant N\big\}.$$

If $q(\cdot, \theta)$ is differentiable on Θ, a solution $\hat{\theta} \in int(\Theta)$ and $(\hat{\eta}, \hat{\lambda}') \in \Lambda_N^\dagger(\hat{\theta})$ must satisfy the

[2]The multiple roots problem could be addressed by selecting, among all the roots of the first order conditions, the one that minimizes the MD objective function. That is, if $(\theta_j', \eta_j, \lambda_j')$ $(j = 1, \ldots, J)$, solve the first order conditions, one can form $\pi_i^j = \tilde{\gamma}_1(\eta_j + \lambda_j' q_i(\theta_j))/N$ $(i = 1, \ldots, N)$ and choose the solution that satisfies $\min_{j\in\{1,\ldots,J\}} \sum_i^N \gamma(N\pi_i^j)$. It is however difficult to recover all J roots to the estimating equations, especially when M and/or K are large.

following first order conditions

$$\begin{aligned}&\sum_{i=1}^{N}\psi_1(\hat{\eta}+\hat{\lambda}'q_i(\hat{\theta}))/N=1,\quad \sum_{i=1}^{N}\psi_1(\hat{\eta}+\hat{\lambda}'q_i(\hat{\theta}))q_i(\hat{\theta})/N=0,\\&\sum_{i=1}^{N}\psi_1(\hat{\eta}+\hat{\lambda}'q_i(\hat{\theta}))G_i(\hat{\theta})'\hat{\lambda}/N=0.\end{aligned}\tag{5}$$

The first order conditions in (5) differ from (3) in that $\tilde{\gamma}_1$ is substituted with ψ_1.

The following theorems make the relationship between the solutions to (2) and the solutions to (4) explicit. The result is not established in terms of first order conditions. Instead it applies more generally even when the moment function $q(\cdot,\theta)$ is not differentiable. Let $\hat{q}_i := q_i(\hat{\theta})$, $\hat{\pi}_i := \tilde{\gamma}_1(\hat{\eta}+\hat{\lambda}'\hat{q}_i)/N$, $\widehat{\Gamma}_N := \sum_{i=1}^{N}\gamma(N\hat{\pi}_i)/N$, and $\widehat{P}_N := \sum_{i=1}^{N}\psi(\hat{\eta}+\hat{\lambda}'\hat{q}_i)/N-\hat{\eta}$.

Theorem 1. *Suppose $\hat{\theta}\in\Theta$ and $(\hat{\eta},\hat{\lambda}')\in\Lambda_N(\hat{\theta})$ solve (4) for some $\psi\in\mathcal{P}$. Then $\hat{\theta}$ and $\hat{\pi}_i$ $(i=1,\ldots,N)$ solve (2) when $\gamma(x)=x\tilde{\psi}_1(x)-\psi(\tilde{\psi}_1(x))$. For this choice of the divergence, it holds that: $\gamma\in\mathcal{G}$, $\psi_1(x)=\tilde{\gamma}_1(x)$ for $x\in D_\psi$, $D_\psi=\mathcal{A}_\gamma$, and $\widehat{P}_N=-\widehat{\Gamma}_N$.*

Proof. See Appendix A.

The next result establishes the converse of Theorem 1: for any divergence $\gamma\in\mathcal{G}$, there exists a function $\psi\in\mathcal{P}$ such that if $\hat{\theta}\in\Theta$ and $\hat{\pi}_i=\tilde{\gamma}_1(\hat{\eta}+\hat{\lambda}'\hat{q}_i)/N$ $(i=1,\ldots,N)$ solve (2), then $(\hat{\theta}',\hat{\eta},\hat{\lambda}')$ solves (4).

Theorem 2. *Suppose $\hat{\theta}\in\Theta$ and $\hat{\pi}_i=\tilde{\gamma}_1(\hat{\eta}+\hat{\lambda}'\hat{q}_i)/N$ $(i=1,\ldots N)$ solve (2) for some $\gamma\in\mathcal{G}$. Then $(\hat{\theta},\hat{\eta},\hat{\lambda}')$ solves (4) when $\psi(x)=x\tilde{\gamma}_1(x)-\gamma(\tilde{\gamma}_1(x))$. For this choice of ψ, it holds that: $\psi\in\mathcal{P}$, $\gamma_1(x)=\tilde{\psi}_1(x)$ for $x\in D_\gamma$, $\mathcal{A}_\gamma=D_\psi$, and $\widehat{P}_N=-\widehat{\Gamma}_N$.*

Proof. See Appendix A.

Theorem 1 and Theorem 2 establish the complete equivalence between the MD problem in (2) and the saddle point problem in (4): not only any MD estimator can be interpreted as solving a saddle point problem for a given $\psi\in\mathcal{P}$; for any criterion function $\psi\in\mathcal{P}$, there exists an underlying MD problem whose solution is the same as the one to the saddle point problem.

Remark 4. If $q(\cdot,\theta)$ is differentiable on Θ, Theorems 1-2 imply that solutions to (2) and (4) solve the same first order conditions, since (3) and (5) are equivalent if $\psi_1(x)=\tilde{\gamma}_1(x)$ for $x\in D_\psi$. Even for $q(\cdot,\theta)$ differentiable, however, Theorems 1-2 give a more general result than simple first order conditions equivalence: the objective functions in (2) and (4) are shown to be equal at $(\hat{\theta},\hat{\pi}_1,\ldots,\hat{\pi}_N)$ and $(\hat{\eta},\hat{\lambda}',\hat{\theta}')$.

In some cases, for a given divergence there exists a closed form ψ function. For example, as shown in Table 1, the divergences of EL, ET, CUE and CR imply: $\psi^{el}(x) = -\ln(1-x)$, $\psi^{et}(x) = \exp x - 1$, $\psi^{cue}(x) = x^2/2 + x$, and $\psi^{cr}(x;\alpha) = \left[(1+\alpha x)^{\frac{1+\alpha}{\alpha}} - 1\right]/(1+\alpha)$. In other cases, though, for a given divergence in $\mathcal{G}$, the implied ψ does not have a closed form expression. This situation is problematic inasmuch as MD estimators are in practice defined as solutions to (4). The importance of Theorem 1 is that it shows that any MD estimator can be defined from the "bottom-up" as the solution to (4) for a given $\psi \in \mathcal{P}$. The implied divergence may not have a closed form expression, but this does not present a practical difficulty: what is needed to give a sound theoretical foundation to the estimation procedure in (4) is only the existence of an implied divergence, not its closed form expression.

Remark 5. When the divergence implied by a given $\psi \in \mathcal{P}$ is not available in closed form, its features can still be studied since the inverse function of $\psi(x)$ can be obtained by numerically solving $\psi_1(x) = y$ for $y \in D_\psi$. In a later section, we follow this approach to obtain graphical representation of divergences implied by certain functions $\psi \in \mathcal{P}$ with attractive statistical properties. We then compare them to the divergences of EL and ET.

Remark 6. Theorem 1 does not make any uniqueness claim about the solution. Uniqueness of $\hat{\theta}$ as a solution to (4) is not guaranteed because the function $P_N(\eta, \lambda, \theta)$ is not necessarily (strictly) concave in θ. Theorem 1 only says that every θ that solves (4) will also solve the corresponding MD problem. However, by the same arguments in Remark 4, the optimal weights will be unique.

2.3 The GEL problem

The GEL estimator of Newey and Smith (2004) solves the following optimization problem:

$$\sup_{\theta\in\Theta} \min_{\tau\in T_N(\theta)} P_N(\tau,\theta), \quad P_N(\tau,\theta) = \sum_{i=1}^{N} \psi(\tau' q_i(\theta))/N, \ \psi \in \mathcal{P}, \tag{6}$$

where $T_N = \{\tau \in \mathbb{R}^M : \tau' q_i(\theta) \in D_\psi, \text{ for all } i \leqslant N\}$. NS show that the first order conditions of the optimization problem in (6) and the first order conditions of the MD problem in (2) agree for $\gamma^{cr}(x;\alpha)$ and $\psi^{cr}(x;\alpha)$.[3] We give here sufficient conditions under which the GEL solutions coincide with the solutions to an MD problem for a generic $\gamma \in \mathcal{G}$. First, we introduce the concept of generalized homogeneous functions.

[3]The Cressie-Read family of divergences considered by NS is slightly different from the one considered here. The difference is due to the normalizations that insure that $\gamma^{cr} \in \mathcal{G}$.

Name	$\gamma(x)$	$\gamma_1(x)$	$\tilde{\gamma}_1(x)$	$\psi(x)$	$\psi_1(x)$	$\tilde{\psi}_1(x)$	$\mathcal{A}$	$\psi_3(0)$
Empirical Likelihood	$-\ln x + x - 1$	$1 - \frac{1}{x}$	$1/(1-x)$	$-\ln(1-x)$	$1/(1-x)$	$1 - 1/x$	$(-\infty, 1)$	2
Exponential Tilting	$x \ln x - x + 1$	$\ln x$	$\exp x$	$\exp x - 1$	$\exp x$	$\ln x$	$(-\infty, +\infty)$	1
CUE	$x^2/2 - x + .5$	$x - 1$	$1 + x$	$x^2/2 + x$	$1 + x$	$x - 1$	$(-\infty, +\infty)$	0
Hellinger Divergence	$-4(\sqrt{x} - 1) + 2(x-1)$	$2 - \frac{2}{\sqrt{x}}$	$(1 - .5x)^{-2}$	$2(1 - x/2)^{-1} - 2$	$(1 - .5x)^{-2}$	$2 - 2/\sqrt{x}$	$(-\infty, 2)$	
Cressie Read Family, $\alpha \neq \{-1, 0\}$	$\frac{x^{\alpha+1}-1}{\alpha(\alpha+1)} - \frac{(x-1)}{\alpha}$	$\frac{x^{\alpha}-1}{\alpha}$	$(1 + \alpha\ x)^{1/\alpha}$	$\frac{(1+\alpha x)^{\frac{1+\alpha}{\alpha}}-1}{1+\alpha}$	$(1 + \alpha\ x)^{1/\alpha}$	$x^{\alpha}/\alpha - 1/\alpha$	‡	$1 - \alpha$
Hyperbolic Tilting	NA	NA	NA	$e^{\sinh x} - 1$	$\cosh x\, e^{\sinh x}$	NA		2
Quartic Tilting	NA	NA	NA	$\begin{cases} h(x) & x > \nu \\ \frac{e^{c_1 x}}{c_2} - c_3 & x \leqslant \nu \end{cases}$ †	$\begin{cases} h_1(x) & x > x_0 \\ \frac{c_1}{c_2} e^{c_1 x} & x \leqslant x_0 \end{cases}$ †	NA	$(-\infty, +\infty)$	2

Table 1: Divergence and dual functions

‡ For the Cressie Read family of divergences the shape of the set $\mathcal{A}$ depends on α. If $\alpha > 0$ and $(1+\alpha)/\alpha \in \mathbb{N}$ an even number, then $\mathcal{A} = (-\infty, +\infty)$.

† $h(x) = e^{((1+x)^4 - 4x - 1)/12} + x - 1$, $\nu < 0$, $c_1 = h_1(\nu)/(c_3 + h(\nu))$, $c_2 = e^{c_1 \nu}/(h(\nu) + c_3)$, and $c_3 = h_1(\nu)^2/h_2(\nu) - h(\nu)$.

Definition 1. Let $a, h : A \subseteq \mathbb{R} \to B \subseteq \mathbb{R}$. A function $f : C \subseteq \mathbb{R} \to E \subseteq \mathbb{R}$ is generalized homogeneous if $f(\kappa x) = a(\kappa) + h(\kappa) f(x)$ for all $x \in C$ and any constant $\kappa \in A$ such that $\kappa x \in C$.

Let $\tilde{q}_i := q_i(\tilde{\theta})$, $\tilde{\pi}_i := \psi_1(\tilde{\tau}'\tilde{q}_i)/N$, $\tilde{\omega}_i := \tilde{\pi}_i / \sum_{i=1}^N \tilde{\pi}_i$, $\widetilde{\Gamma}_N := \sum_{i=1}^N \gamma(\tilde{\gamma}_1(N\tilde{\pi}_i))/N$, $\widetilde{\Gamma}_N^\dagger := \sum_{i=1}^N \gamma(\tilde{\gamma}_1(N\tilde{\omega}_i))/N$, and $\widetilde{P}_N := \sum_{i=1}^N \psi(\tilde{\tau}'\tilde{q}_i)/N$.

Theorem 3. *Suppose $(\tilde{\theta}', \tilde{\tau}')$ solves (6) for some $\psi \in \mathcal{P}$. If $\tilde{\psi}_1$ is generalized homogeneous, then $\tilde{\theta}$ and $\tilde{\omega}_i$ $(i = 1, \ldots, N)$ solve (2) when $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$. For this choice of γ it holds: $\gamma \in \mathcal{G}$, $D_\psi = \mathcal{A}_\gamma$, $\psi_1(x) = \tilde{\gamma}_1(x)$, $x \in D_\psi$, and $\widetilde{P}_N = -\widetilde{\Gamma}_N^\dagger = \widehat{P}_N$.*

Proof. See Appendix A □

If the inverse function of the first derivative of $\psi \in \mathcal{P}$ is generalized homogeneous, GEL estimators and MD estimators coincide for γ given in Theorem 3. Therefore, from Theorem 1, if $\tilde{\theta}$ solves the GEL problem then it must also solve (4), $\widetilde{P}_N = -\widetilde{\Gamma}_N^\dagger = \widehat{P}_N = -\widehat{\Gamma}_N$, and $\hat{\pi}_i = \tilde{\omega}_i$ $(i = 1, \ldots, N)$.

When $\tilde{\psi}_1$ is not homogeneous, GEL and MD problems are in general solved by different values of θ. In fact, the GEL problem can be shown to be equivalent to an MD problem that does not constrain the weights to sum to one:

$$\min_{\theta \in \Theta, p_1, \ldots, p_N} \sum_{i=1}^N \gamma(Np_i), \quad \gamma \in \mathcal{G}$$
$$s.t. \sum_{i=1}^N p_i q_i(\theta) = 0, \;\; Np_i \in (a_\gamma, b_\gamma), \; i = 1, \ldots, N. \tag{7}$$

Let $\tilde{p}_i = \psi_1(\tilde{\tau}'\tilde{q}_i)/N$.

Theorem 4. *Suppose $(\tilde{\theta}', \tilde{\tau}')$ solves (6) for some $\psi \in \mathcal{P}$. Then $\tilde{\theta}$ and $\tilde{\pi}_i$ $(i = 1, \ldots, N)$ solve (7) when $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$. For this choice of γ it holds: $\gamma \in \mathcal{G}$, $\mathcal{A}_\gamma = D_\psi$, $\psi_1(x) = \tilde{\gamma}_1(x)$, $x \in D_\psi$, and $\widetilde{P}_N = -\sum_{i=1}^N \gamma(N\tilde{p}_i)/N \geqslant \widehat{P}_N$.*

Proof. See Appendix A □

EL, ET, CUE, and in general members of the Cressie Read family posses the generalized homogeneous property. Generalized homogeneity of $\tilde{\psi}_1$ may be difficult to assess in general and especially when $\tilde{\psi}_1$ does not have a closed form expression.[4] For this reason in the remainder of the paper we consider estimators solving (4); the small computational cost (the inner optimization is with respect to $M + 1$ instead of M parameters) is outweighed by the fact that regardless of the homogeneity of the inverse of ψ_1, solving (4) is equivalent to solving (2).

[4]Since by Theorems 1 and 2 $\tilde{\psi}_1(x) = \gamma_1(x)$, $\tilde{\psi}_1$ does not have a closed form expression any time the corresponding divergence does not have a closed form expression.

3 Asymptotic

In this section we derive the asymptotic distribution of estimators defined as solutions to (4). We make the following assumptions.

(A1) (a) $\theta_0 \in \Theta$ is the unique solution to $E[q(w,\theta)] = 0$; (b) Θ is compact; (c) $q(\cdot,\theta)$ is continuous at each $\theta \in \text{int}(\Theta)$, w.p.1; (d) $E\left[\sup_{\theta\in\Theta} \|q(w,\theta)\|^2\right] < \infty$; (e) $\Omega = E[q_i(\theta_0)q_i(\theta_0)']$ is non-singular.

(A2) (a) $\theta_0 \in \text{int}(\Theta)$; (b) $q(w,\theta)$ is continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0; (c) $E\left[\sup_{\theta\in\mathcal{N}} \|G_i(\theta)\|\right] < \infty$; (d) $\text{Rank}(G) = K$, $G = E\left[G_i(\theta_0)\right]$.

Theorem 5. *If A1 holds,* $\hat{\theta} \xrightarrow{p} \theta_0$, $\hat{\eta} = o_p(N^{-1/2})$, *and* $\hat{\lambda} = O_p(N^{-1/2})$.

Proof. See Appendix A □

The consistency proof uses ideas from Kitamura et al. (2004). Not surprisingly, the Lagrange multiplier $\hat{\eta}$ converges to zero faster than $\sqrt{N}$, implying that the first order asymptotic properties of GEL and MD estimators coincide: the asymptotic distribution of $\hat{\lambda}$ and $\hat{\theta}$ is identical to the asymptotic distribution of the GEL parameters $\tilde{\tau}$ and $\tilde{\theta}$ (see, NS, Theorem 2.2), even when the generalized homogeneity property does not hold, as the next result makes clear.

Theorem 6. *If A1 and A2 hold,*

$$\sqrt{N}\begin{pmatrix} \hat{\lambda} \\ \hat{\theta} - \theta_0 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} P & 0 \\ 0 & \Sigma \end{pmatrix}\right),$$

where $\Sigma = (G'\Omega^{-1}G)^{-1}$, $P = \Omega^{-1}(I_M - G\Sigma G'\Omega^{-1})$.

Proof. See Appendix A □

The weights $\hat{\pi}_i$ $(i = 1,\ldots,N)$ can be used to construct an efficient estimate of the distribution function of w. For any Borel set A, the probability $p_A := P(w \in A)$ can be estimated by

$$\hat{p}_A = \sum_{i=1} 1(x \in A)\hat{\pi}_i = \sum_{i=1} 1(x \in A)\tilde{\gamma}_1(\hat{\eta} + \hat{\lambda}'\hat{q}_i)/N.$$

The following theorem summarizes the properties of this estimator.

Theorem 7. *If A1 and A2 hold,*

$$\hat{p}_A \xrightarrow{p} p_A, \quad \sqrt{N}(\hat{p}_A - p_A) \xrightarrow{d} \mathcal{N}(0, V_A),$$

where $V_A = p_A(1-p_A) - E[q(w,\theta)1(w \in A)]' PE[q(w,\theta)1(w \in A)]$. *Further,* $\hat{p}_A$ *is efficient in the sense that* V_A *reaches the semiparametric efficiency bound.*

Proof. See Appendix A □

Semiparametric efficient estimators for p_A that incorporate the information about the moment restrictions have been proposed and analyzed in the GMM context by Back and Brown (1993) and Brown and Newey (1998). Newey and Smith (2004), Ramalho and Smith (2005), and Brown and Newey (2002) discuss estimation of efficient probability under (1) in the GEL context using the normalized weights $\tilde{\omega}_i = \tilde{\pi}_i / \sum_{i=1}^N \tilde{\pi}_i$ $(i = 1, \ldots, N)$.

4 Testing overidentifying restrictions

In the GEL framework, test statistics are based either on (i) the GEL objective function (Smith, 1997; Newey and Smith, 2004); (ii) a quadratic form in the Lagrange multipliers (Imbens et al. (1998)); (iii) implied probabilities (Ramalho and Smith, 2005). The results in Section 2 can be used to cast the statistics proposed in the literature in a unified framework. Specifically, all the statistics can be expressed in terms of the divergence of the underlying MD problem.

The overidentification test statistic based on the GEL criterion function proposed by Newey and Smith (2004) is given by

$$GEL(\tilde{\theta}) = -2 \sum_{i=1}^{N} \psi(\tilde{\tau}' \tilde{q}_i).$$

The corresponding statistic based on the MD saddle point problem is

$$D(\hat{\theta}) = -2 \Big[\sum_{i=1}^{N} \psi(\hat{\eta} + \hat{\lambda}' \hat{q}_i) - N\hat{\eta} \Big].$$

From Theorem 3, if $\tilde{\psi}_1$ is a generalized homogeneous function, then

$$GEL(\tilde{\theta}) = D(\hat{\theta}) = 2 \sum_{i=1}^{N} \gamma(N\hat{\pi}_i)/N.$$

If $\tilde{\psi}_1$ is not generalized homogeneous, the equality above does not hold and we have instead

$$D(\hat{\theta}) = 2 \sum_{i=1}^{N} \gamma(N\hat{\pi}_i)/N, \quad GEL(\tilde{\theta}) = 2 \sum_{i=1}^{N} \gamma(N\tilde{p}_i)/N, \quad GEL(\tilde{\theta}) \leqslant D(\hat{\theta}).$$

The inequality $GEL(\tilde{\theta}) \leqslant D(\hat{\theta})$ follows from the fact that the GEL optimization is equivalent to an MD problem in which the weights are not restricted to sum to one: once the restriction is removed, the minimum attained in (7) must be lower or equal to the minimum attained in (2).

Imbens et al. (1998) propose statistics for testing (1) based on the Lagrange multipliers of EL and ET. In our setup, the corresponding statistics are given by

$$LM_{\omega}(\tilde{\theta}) = N\tilde{\tau}'\tilde{\Omega}_{\omega}\tilde{\tau}, \quad LM_{\pi}(\hat{\theta}) = N\big(\hat{\eta}^2 + \hat{\lambda}'\hat{\Omega}_{\pi}\hat{\lambda}\big),$$

where $\tilde{\Omega}_{\omega} = \sum_{i=1}^{N} \tilde{\omega}_i q_i(\tilde{\theta}) q_i(\tilde{\theta})'$ and $\hat{\Omega}_{\pi} = \sum_{i=1}^{N} \hat{\pi}_i q_i(\hat{\theta}) q_i(\hat{\theta})'$ are consistent for Ω. The Lagrange multipliers can be scaled by any consistent estimator of Ω without affecting the asymptotic distribution of the statistics. For instance, Imbens et al. (1998) consider scaling the Lagrange multipliers by a robust weighting matrix given by

$$\hat{\Omega}_r = \hat{\Omega}_{\pi}\left[\sum_{i=1}^{N} \hat{\pi}_i^2 q_i(\hat{\theta}) q_i(\hat{\theta})'\right]^{-1} \hat{\Omega}_{\pi}.$$

The intuition behind these statistics is simple: if the moment conditions are satisfied, $(\hat{\eta}, \hat{\lambda}') \xrightarrow{p} 0$ and $\tilde{\tau} \xrightarrow{p} 0$ and so will the LM statistics. Using our equivalence results, we can cast these statistics into a more coherent framework. In fact, if $\tilde{\psi}_1$ is a generalized homogeneous function, then

$$LM_{\omega}(\tilde{\theta}) = LM_{\pi}(\hat{\theta}) = 2\sum_{i=1}^{N} \hat{\pi}_i \gamma_1 (N\hat{\pi}_i)^2,$$

otherwise,

$$LM_{\pi}(\hat{\theta}) = 2\sum_{i=1}^{N} \hat{\pi}_i \gamma_1 (N\hat{\pi}_i)^2, \quad LM_{\omega}(\tilde{\theta}) = 2\sum_{i=1}^{N} \tilde{\omega}_i \gamma_1 (N\tilde{\pi}_i)^2.$$

This characterization shows that when $\tilde{\Omega}_{\omega}$ and $\hat{\Omega}_{\pi}$ are used to scale the Lagrange multipliers, the LM statistics can be thought of as a semiparametric version of score statistics, where the score is based on the first derivative of the divergence.

When $\psi_3 \neq 2$, the Lagrange multiplier $\hat{\eta}$ can be used to test the overidentifying restrictions using the following statistic

$$LM_{\eta}(\hat{\theta}) = \frac{N\hat{\eta}}{(1 - \psi_3(0)/2)}.$$

For the ET case, we have that $\hat{\eta} = -\sum_{i=1}^{N} \psi(\hat{\lambda}' q_i(\hat{\theta}))/N$. Also, since $\psi_3(0) = 1$, we have

that $LM_\eta(\hat{\theta}) = -2\sum_{i=1}^{N}\psi(\hat{\lambda}'q_i(\hat{\theta}))$.

Proposition 1. *If A1-A2 hold,*

$$D(\hat{\theta}), GEL(\tilde{\theta}), LM_\eta(\hat{\theta}), LM_\pi(\hat{\theta}), LM_\omega(\tilde{\theta}) \xrightarrow{d} \chi^2_{(M-K)}.$$

Proof. See Appendix A □

The $\chi^2_{(M-K)}$ calibration can be easily shown to hold even if $(\hat{\theta}', \hat{\eta}, \hat{\lambda}')$ are replaced by $\sqrt{N}$ equivalent estimators. It also holds when the divergence defining $\sum_{i=1}^{N}\gamma(N\hat{\pi}_i)/N$ is different from the divergence under which $\hat{\pi}_i$ $(i = 1, \ldots, N)$ are optimal. Thus, one can obtain $(\dot{\theta}', \dot{\eta}, \dot{\lambda}')$ by solving (4) with $\psi^{el}(x) = -\ln(1-x)$, but test for overidentified restrictions using $2\sum_{i=1}^{N}\gamma(N\dot{\pi}_i)$ with the CUE divergence $\gamma^{cue}(x) = x^2/2 - x + .5$ and EL weights, $\dot{\pi}_i^{el} = (1 - \dot{\eta} - \dot{\lambda}'q_i(\dot{\theta}))^{-1}/N$, that is:

$$2\sum_{i=1}^{N}\gamma^{cue}(N\dot{\pi}_i^{el}) = \sum_{i=1}^{N}\left(N\dot{\pi}_i^{el}\right)^2 - N.$$

Through Monte Carlo simulations, Ramalho and Smith (2005) show that this particular test statistic has competitive size properties.

5 Higher order expansions

In this section we investigate the higher order properties of MD estimators. The analysis is similar to the one in NS, but it emphasizes different points. NS focus their exploration on the asymptotic differences between GEL and GMM estimators. We are instead concerned with the ranking—in terms of higher order efficiency—of estimators in the MD class. In the following, some basic higher order asymptotic concepts are introduced. For further details on higher order asymptotic concepts, see Rothenberg (1984), Ullah (2004), and references therein.

A higher order asymptotic analysis begins with a $O_p(N^{-2})$ expansion of $(\hat{\theta} - \theta_0)$:

$$(\hat{\theta} - \theta_0) = \frac{i_N}{\sqrt{N}} + \frac{b_N}{N} + \frac{c_N}{N\sqrt{N}} + \frac{r_N}{N^2}, \tag{8}$$

where i_N, b_N, c_N and r_N are $O_p(1)$. These terms are also tractable being sums and products of certain sample moments of functions of the underlying random vectors. The first order asymptotic behavior of $\hat{\theta}$ is entirely determined by the influence function i_N. In the discussion that follows, we restrict ourselves to estimators that admit a $O_p(N^{-2})$ expansion and are consistent for θ_0, in which case $E[i_N] = 0$.

The higher order bias of $\hat{\theta}$ is derived by taking the expectations of the $O_p(N^{-1})$ term in the expansion.

Definition 2. The $O(N^{-1})$ bias of $\hat{\theta}$ is given by $E\big[b_N\big]/N$.

The higher order MSE of $\hat{\theta}$ is obtained by calculating $E[(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)']$ using the asymptotic expansion and dropping terms that are of order $o(N^{-2})$.

Definition 3. The $O(N^{-2})$ MSE of $\hat{\theta}$ is given by

$$MSE_2(\hat{\theta}) = E\big[i_N i_N'\big]/N + \Xi/N,$$

where

$$\Xi := E[b_N b_N'/N] + E\big[(b_N/\sqrt{N} + c_N/N)i_N'\big] + E\big[i_N(b_N/\sqrt{N} + c_N/N)'\big].$$

The first term of the higher order MSE, $E[i_N i_N']$ coincides with the first-order asymptotic variance of the estimator; the second term Ξ contains terms of order $O(N^{-1/2})$ and $O(N^{-1})$.

First order efficient estimators of θ_0 could be compared on the basis of their MSE_2, by saying that $\hat{\theta}$ is not worse than $\tilde{\theta}$ if $MSE_2(\hat{\theta}) - MSE_2(\tilde{\theta})$ is a positive semi-definite matrix. A more sensible approach, that excludes many unreasonable estimators, is to consider only estimators that are adjusted to be unbiased of order $O(N^{-1})$.

An estimator $\hat{\theta}$ can be bias adjusted to order $O(N^{-1})$ by defining $\hat{\theta}^c = \hat{\theta} - \widehat{E[b_N]}/N$, where $\widehat{E[b_N]}$ denotes an estimator of $E[b_N]$. Since $E[b_N]$ depends on unknown quantities, $\widehat{E[b_N]}$ will be constructed using $\hat{\theta}$ and sample versions of the corresponding population moments that enter $E[b_N]$. Under the same assumptions under which the $O_p(N^{-2})$ expansion holds, we usually have that

$$\widehat{E[b_N]} = E[b_N] + \xi_N/\sqrt{N} + O_p(N^{-1}),$$

where $\xi_N = O_p(1)$. Using this approximation and the expansion of $\hat{\theta}$ we obtain a valid $O_p(N^{-2})$ expansion of $\hat{\theta}^c$:

$$(\hat{\theta}^c - \theta_0) = (\hat{\theta} - \theta_0) - \frac{\widehat{E[b_N]}}{N} = \frac{i_N}{\sqrt{N}} + \frac{b_N - E[b_N]}{N} + \frac{c_N - \xi_N}{N\sqrt{N}} + O_p(N^{-2}). \tag{9}$$

This expansion, which shows immediately that $\hat{\theta}^c$ is unbiased of order $O(N^{-1})$, can be used to derive an expression for the higher order MSE of $\hat{\theta}^c$.

Definition 4. The $O(N^{-2})$ MSE of the bias corrected estimator is

$$MSE_2(\hat{\theta}^c) = MSE_2(\hat{\theta}) + \Xi^c/N,$$

where

$$\Xi^c := E[(\xi_N - i_n\xi_N)/N] + E[b_N b_N'/N] - E[b_N]E[b_N]/N.$$

The notion of second order efficiency is based on the MSE_2 of the bias corrected estimator.

Definition 5. The bias corrected estimator $\hat{\theta}^c$ is second order efficient if for any other bias corrected estimator $\tilde{\theta}^c$ there exists a positive definite matrix Π such that $MSE_2(\hat{\theta}^c) - MSE_2(\tilde{\theta}^c) = \Pi + o(N^{-2})$.

Pfanzagl and Wefelmeyer (1979) and Ghosh et al. (1980) show that the bias corrected maximum likelihood estimator is second order efficient and thus optimal among other bias corrected estimators.[5] NS show that the bias corrected EL estimator is higher order efficient, having the smallest $O(N^{-2})$ MSE among all the bias corrected estimators based on the same moment conditions.

5.1 Higher order comparison of MD estimators

We use the following notation. Components of vectors are indexed using superscripts. Thus $\hat{\theta}^2$ denotes the second component of the vector $\hat{\theta}$. Matrix are denoted componentwise adopting the index notation. So, a_{ij} is the element (i, j) of the matrix A. Raised indexes denote inverse matrix: a^{ij} denotes the (i, j) element of A^{-1}. We use the summation convention for matrix product (see McCullagh, 1987). In any expression, a twice repeated index (occurring twice as a subscript, twice as a superscript, or once as a subscript and once as a superscript) shall automatically stand for its sum over the values of the repeated index. We work with three sets of indexes: (i) $a, b, c, d, e, f, g, h \in \{1, \ldots, M+K+1\}$, (ii) $j, k, \ell, m, n, o \in \{1, \ldots, M\}$, (iii) $r, s, t, u, v, w \in \{M+1, \ldots, M+K\}$. Let $\beta = (\lambda', \theta', \eta)'$ and define

$$Q_{i,1}(\beta) := \psi_1(\eta + \lambda' q_i(\theta))q_i(\theta)$$
$$Q_{i,2}(\beta) := \psi_1(\eta + \lambda' q_i(\theta))G_i(\theta)'\lambda$$

[5]Second order efficiency is also referred to as third order efficiency (see, for instance, Pfanzagl and Wefelmeyer (1979)). If one considers the $O(N^{-2})$ asymptotic expansion of MSE, it is the second term, Ξ^c/N, that defines the efficiency of the estimator, hence the second order. If one approaches the problem through Edgeworth expansions, one has three terms in powers of $1/\sqrt{N}$, hence third order. We call it second order because we are analyzing higher order efficiency by means of MSE expansions.

$$Q_{i,3}(\beta) := \psi_1(\eta + \lambda' q_i(\theta)) - 1$$

The first order conditions of the MD estimator can be conveniently rewritten as

$$Q(\hat{\beta}) := \sum_{i=1}^{N} Q_i(\hat{\beta})/N = 0$$

where $Q_i(\beta) = (Q_{i,1}(\beta)', Q_{i,2}(\beta)', Q_{i,3}(\beta))'$. We define the following moments of the derivatives of the first order conditions:

$$\mu_{ab} \equiv E\left[\frac{\partial Q^a(\beta_0)}{\partial \beta^b}\right], \quad \mu_{abc} \equiv E\left[\frac{\partial^2 Q^a(\beta_0)}{\partial \beta^b \partial \beta^c}\right], \quad \mu_{abcd} \equiv E\left[\frac{\partial^3 Q^a(\beta_0)}{\partial \beta^b \partial \beta^c \partial \beta^d}\right], \ldots,$$

where $\beta_0 = (0, \theta_0', 0)'$. We also let:

$$Z_a = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Q_i^a(\beta_0), \qquad Z_{ab} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial Q_i^a(\beta_0)}{\partial \beta^b} - \sqrt{N}\mu_{ab},$$

$$Z_{abc} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial Q_i^a(\beta_0)}{\partial \beta^b \partial \beta^c} - \sqrt{N}\mu_{abc}, \qquad Z_{abcd} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial Q_i^a(\beta_0)}{\partial \beta^b \partial \beta^c \partial \beta^d} - \sqrt{N}\mu_{abcd},$$

and so forth.

Following McCullagh (1987), we expand $\sum_{i=1}^{N} Q_i(\hat{\beta})/N = 0$ around β_0 by means of Taylor expansions. Let $\hat{\delta}^a = \sqrt{N}(\hat{\beta}^a - \beta_0^a)$. Then,

$$\begin{aligned} 0 =& N^{1/2} Z_a + (N^{1/2} Z_{ab} + N\mu_{ab})\hat{\delta}^b/N^{1/2} + (N^{1/2} Z_{abc} + N\mu_{abc})\delta^b\delta^c/2N \\ &+ (N^{1/2} Z_{abcd} + N\mu_{abcd})\hat{\delta}^b\hat{\delta}^c\hat{\delta}^d/6N^{3/2} + o_p(N^{-3/2}). \end{aligned}$$

The validity of the previous expansion can be verified under the following assumptions:

(A3) There is $b(w)$ with $E[b(w_i)^6] < \infty$ such that for $0 \leq j \leq 4$ and all w, $\partial^j q(w,\theta)/\partial\theta^j$ exists on a neighborhood $\mathcal{N}$ of θ_0, $\sup_{\theta\in\mathcal{N}} \|\partial^j q(w,\theta)/\partial\theta^j\| \leq b(w)$, and for each $\theta \in \mathcal{N}$, $\|\partial^4 q(w,\theta)/\partial\theta^4 - \partial^4 q(w,\theta_0)/\partial\theta^4\| \leq b(w)\|\theta - \theta_0\|$, and $\psi(x)$ is four times continuously differentiable with Lipschitz fourth derivative in a neighborhood of zero.

To obtain a $O_p(N^{-2})$ expansion for $\hat{\beta}^a$ of the type in (8), one proceeds by telescopic substitution of lower order expansions to obtain

$$\hat{\beta}^a - \beta_0 = i_N^a/\sqrt{N} + b_N^a/N + c_N^a/N\sqrt{N} + O_p(N^{-2}), \tag{10}$$

where, for $\mu^{a,b,c} = \mu^{ad}\mu^{be}\mu^{cf}\mu_{def}$ and $\mu^{a,b,c,d} = \mu^{ae}\mu^{bf}\mu^{cg}\mu^{dh}\mu_{efgh}$,

$$
\begin{aligned}
i_N^a &= -\mu^{aj} Z_j \\
b_N^a &= \mu^{ab}\mu^{cj} Z_{bc} Z_j - \mu^{a,j,k} Z_j Z_k/2 \\
c_N^a &= -\mu^{ab}\mu^{cd}\mu^{ej} Z_{bc} Z_{de} Z_j + \mu^{a,j,c}\mu^{dk} Z_j Z_{cd} Z_k \\
&\quad -\mu^{ab}\mu^{cjk} Z_{bc} Z_j Z_k + \mu^{a,j,c}\mu^{k,\ell,f}\mu_{cf} Z_j Z_k Z_\ell \\
&\quad -\mu^{ab}\mu^{jc}\mu^{kd} Z_{bcd} Z_j Z_k/2 + \mu^{a,j,k,\ell} Z_j Z_k Z_\ell/6.
\end{aligned}
$$

Following Definition 2, the bias of the MD estimator of θ_0 in (1) and of the Lagrange parameter λ is obtained by taking expectations of b_N^r, $r \in (M+1, \ldots, M+K)$, and of b_N^j, $j \in (1, \ldots, M)$, respectively. Here we give expressions for the bias in which the expectations of higher order derivatives of Q_i are substituted for with expectations of higher order derivatives of q_i:

$$
\begin{aligned}
E[b_N^r] &= (1-\psi_3/2)\mu^{rj}\mu^{k\ell} E\left[q_i^j q_i^k q_i^\ell\right] \\
&\quad +\mu^{rj}\Big\{\mu^{sk} E\big[(\partial q_i^j/\partial\beta^s) q_i^k\big] - \mu^{st} E\big[\partial q_i^j/\partial\beta^s\partial\beta^t\big]/2\Big\} \qquad (11)
\end{aligned}
$$

$$
\begin{aligned}
E[b_N^j] &= (1-\psi_3/2)\mu^{jk}\mu^{\ell m} E[q_i^k q_i^\ell q_i^m] \\
&\quad +\mu^{jk}\Big\{\mu^{r\ell} E\big[(\partial q_i^j/\partial\beta^r) q_i^\ell\big] - \mu^{rs} E\big[\partial q_i^j/\partial\beta^r\partial\beta^s\big]/2\Big\}.
\end{aligned}
$$

The bias of the estimator of θ_0 and the bias of the Lagrange multiplier have the same structure. In particular, only the third derivative of ψ evaluated at 0 affects their magnitude. MD estimators have the same bias of order $O(N^{-1})$ if *(i)* the generalized third moments of q_i are zero, that is, $E[q_i^j q_i^k q_i^\ell] = 0$, all $j, k, \ell \in (1, \ldots, M)$, or *(ii)* $\psi_3 = 2$. Notably, EL has $\psi_3 = 2$.

The expression for the higher order MSE of MD estimators could be obtained by substituting i_N^r, b_N^r and c_N^r into the expression in Definition 3. The resulting expression is however too complex to be of any help for carrying out higher order comparisons. Calculations can be greatly simplified if one focuses on the difference between the higher order MSE of two MD estimators with the same higher order bias. Let $\hat\beta_\psi = (\lambda_\psi', \theta_\psi', \eta_\psi)'$ and $\hat\beta_{\psi'} = (\lambda_{\psi'}', \theta_{\psi'}', \eta_{\psi'})'$ denote the solutions to the MD problem obtained from $\psi \in \mathcal{P}$ and $\psi' \in \mathcal{P}$, respectively. These two estimators expand as

$$
(\hat\beta_\psi^r - \beta_0^r) = i_N^r/\sqrt{N} + b_{\psi,N}^r/N + c_{\psi',N}^r/N\sqrt{N} + O_p(N^{-2}),
$$

and

$$(\hat{\beta}^r_{\psi'} - \beta^r_0) = i^r_N/\sqrt{N} + b^r_{\psi',N}/N + c^r_{\psi',N}/N\sqrt{N} + O_p(N^{-2}),$$

where i^r_N is not indexed by ψ or ψ' because all MD estimators have the same influence function. Let $MSE_2^{r,s}(\hat{\beta}_\psi)$ denote the element (r,s) of the $O(N^{-2})$ MSE of $\hat{\beta}_\psi$. Then the expression for the difference in higher order MSE is given by

$$\begin{aligned} MSE_2^{r,s}(\hat{\beta}_\psi) - MSE_2^{r,s}(\hat{\beta}_{\psi'}) &= \frac{1}{N\sqrt{N}}\left[\text{cov}(b^r_{\psi,N} - b^r_{\psi',N}, i^s_N) + \text{cov}(b^s_{\psi,N} - b^s_{\psi,N} i^s_N)\right] \\ &\quad + \frac{1}{N^2}\left[\text{cov}\left(i^r_N, c^s_{\psi,N} - c^s_{\psi',N}\right) + \text{cov}\left(i^s_N, c^r_{\psi,N} - c^r_{\psi',N}\right)\right]. \end{aligned}$$

We are now ready to establish the main result of this section, that if two MD estimators are obtained from two divergences such that $\psi_3 = \psi'_3$, then they have the same higher order MSE.

Theorem 8. *If A1-A3 hold and $\psi_3 = \psi'_3$, then (i) $b^r_{\psi,N} = b^r_{\psi',N}$, and (ii) $MSE_2^{r,s}(\hat{\beta}_\psi) - MSE_2^{r,s}(\hat{\beta}_{\psi'}) = o(N^{-2})$.*

Proof. See Appendix B □

The first conclusion of Theorem 8 implies that $\text{cov}\left(i^s_N, b^r_{\psi,N} - b^r_{\psi',N}\right) = 0$ and thus the difference between the $O(N^{-2})$ MSE of the two estimators reduces to

$$MSE_2^{r,s}(\hat{\beta}_\psi) - MSE_2(\hat{\beta}_{\psi'}) = \frac{1}{N^2}\left[\text{cov}\left(i^r_N, c^s_{\psi,N} - c^s_{\psi',N}\right) + \text{cov}\left(i^s_N, c^r_{\psi,N} - c^r_{\psi',N}\right)\right].$$

In general, $\text{cov}\left(i_N, c_{\psi,N} - c_{\psi',N}\right)$ is of order $O(1)$, but, as established in the second part of the proof, if $\psi_3 = \psi'_3$, then $\text{cov}\left(i_N, c_{\psi,N} - c_{\psi',N}\right) = O(N^{-1})$, and its contribution to $MSE_2^{r,s}(\hat{\beta}_\psi) - MSE_2^{r,s}(\hat{\beta}_{\psi'})$ is negligible. In particular, we show in the proof that $\text{cov}\left(i^r_N, c^s_{\psi,N} - c^s_{\psi',N}\right) = (\psi_4 - \psi'_4)\Delta^{r,s} + O(N^{-1})$. Showing that $\Delta^{r,s} = O(N^{-1})$ involves checking the asymptotic order of the expectations that enter its expression.

5.2 Second order efficiency

Theorem 8 has an important implication on the second order efficiency of bias corrected MD estimators. The estimator of $E[b^r_{\psi,N}]$ for an MD estimator $\hat{\beta}^r_\psi$ is obtained by substituting population moments in (11) with correspondent sample moments. To describe the specific form of the bias correction for MD, we need to introduce some notation. Let $\mu_{ab,N}(\beta) = \sum_{i=1}^N \frac{\partial Q^a_i(\beta)}{\partial \beta^b}/N$, let $\mu^{ab}_N(\beta)$ denote the element (a,b) of the matrix with

elements $\{\mu_{ab,N}(\beta)\}$, and

$$A_N^{j,k,\ell}(\beta) = \frac{1}{N}\sum_{i=1}^N q_i^j(\theta)q_i^k(\theta)q_i^\ell(\theta), B_N^{j,s,k}(\beta) = \frac{1}{N}\sum_{i=1}^N \frac{\partial q_i^j(\theta)}{\partial \beta^s} q_i^k(\theta), C_N^{j,s,t}(\beta) = \frac{1}{N}\sum_{i=1}^N \frac{\partial q_i^j(\theta)}{\partial \beta^s \partial \beta^t}.$$

Then, the bias adjusted MD estimator is $\bar{\beta}_\psi^r = \hat{\beta}_\psi^r - \widehat{E[b_N^r]}/N$, where the estimator of the bias term is

$$\begin{aligned}\widehat{E[b_{\psi,N}^r]} &= (1-\psi_3/2)\mu_N^{rj}(\hat{\beta}_\psi)\mu_N^{k\ell}(\hat{\beta}_\psi)A_N^{j,k,\ell}(\hat{\beta}_\psi) \\ &\quad +\mu_N^{rj}(\hat{\beta}_\psi)\Big\{\mu_N^{sk}(\hat{\beta}_\psi)B_N^{j,s,k}(\hat{\beta}_\psi) - \mu_N^{st}C_N^{j,s,k}(\hat{\beta}_\psi)/2\Big\}.\end{aligned}$$

Under the assumptions sufficient for the asymptotic expansions for $\hat{\beta}_\psi$, the estimator of the bias admits a $O_p(N^{-1})$ asymptotic expansion.[6]

Theorem 9. *If Assumption A1-A3 are satisfied, then (i)* $\widehat{E[b_{\psi,N}^r]} = E[b_{\psi,N}^r] + \xi_{\psi,N}^r/\sqrt{N} + O_p(N^{-1})$*, where* $\xi_{\psi,N}^r = O_p(1)$ *depends on* ψ *only through* ψ_3*; (ii)* $E[\bar{\beta}_\psi^r] = o(N^{-1})$.

Proof. See Appendix B □

Since $\xi_{\psi,N}$ only depends on ψ trough ψ_3, the bias corrected estimators, $\hat{\beta}_{el}^r - \widehat{E[b_{el,N}^r]}/N$, and $\hat{\beta}_\psi^r - \widehat{E[b_{\psi,N}^r]}/N$, $\psi_3 = 2$, admit the same expansion for their bias correction. By Theorem 8, $b_{el,N}^r = b_{\psi,N}^r$ and $MSE_2(\hat{\beta}_{el}^r) = MSE_2(\hat{\beta}_\psi^r)$. Thus, the two bias corrected estimators will have the same higher order MSE.

Corollary 1. *If Assumption A1-A3 are satisfied, all bias corrected GEL-MD estimators with* $\psi_3 = 2$ *are second order efficient in the sense of Definition 5.*

That bias corrected versions of estimators with $\psi_3 = 2$ are as efficient as the EL. NS mention that any GEL estimator for which $\psi_3 = 2$ and $\psi_4 = 6$ are second order efficient: the analysis of this section shows that $\psi_4 = 6$ is not necessary. This has the effect of enlarging the class of higher order efficient (bias corrected) estimators; for instance, none of the estimators proposed in Section 6 satisfy $\psi_4 = 6$, yet they have the same higher order MSE as EL. Corollary 1 should be interpreted with care. The second order efficiency concept does not say that EL and the other MD-GEL estimators with $\psi_3 = 2$ are optimal, but only that their bias corrected versions are optimal. If one uses the uncorrected estimators, it is possible to have MD estimators with $\psi_3 \neq 2$ with better MSE than estimators with $\psi_3 = 2$. Yet, if the bias contribution to the MSE is high, EL and other estimators with $\psi_3 = 2$ will have lower MSE than other estimators of

[6]The bias term can also be estimated by replacing the empirical distribution by one based on the MD probabilities defined in Section 3. Defining the correction using the MD probabilities does not change the results of this section since it has no effect on the asymptotic variance of $\hat{\beta}_\psi$.

the MD class. Notice that by Theorem 8, the EL and other MD estimators—not their bias corrected versions—with ψ_3 will have the same $o(N^{-2})$. So, if one's objective is to minimize bias, EL and MD with $\psi_3 = 2$ should be considered on equal footing, as far as second order MSE is concerned.

There are few caveats that should be kept in mind when ranking estimators on their higher order MSE. It is well known from Srinivasan (1970) that it is possible for an estimator whose finite sample distribution does not have finite moments to admit a valid asymptotic expansion. This is particularly relevant for the class of estimators considered here since Kunitomo and Matsushita (2003) and Guggenberger (2008) suggest that EL does not have finite moments in a linear simultaneous equations setting. Note however that the higher order moments derived in this section correspond to the moments based on Edgeworth approximations to the sampling distribution (see, Sargan, 1974). If the finite sample moments do not exist, it is sensible to interpret the moments based on (8) as moments of an approximating distribution.

Even if from a theoretical standpoint it is legitimate to carry out a comparison of MD estimators based on higher order moments, one should ask what is the practical significance of this enterprise. First, higher order comparisons of estimators are still based on large N asymptotic arguments and their rates of success in describing the relative performance of MD estimators depend on the particular stochastic environment under consideration. Second, if MD estimators do not have finite moments for a particular design, the approximation provided by the Edgeworth approximation will be poor and so will be the ranking. Monte Carlo evidence, presented in Section 7, indicates that the "no moments" problem of these estimators limited to the simultaneous equations setting.

6 Behavior under misspecification

A moment condition model is said to be misspecified if

(MS) $\| \int q(w, \theta) F(dw) \| > 0$ for all $\theta \in \Theta$.

There are at least two important reasons why it is relevant to consider the behavior of estimators when the model is misspecified. First, it is sometimes reasonable to interpret conditions in (1) as mere approximations of reality. Second, even when the conditions in (1) are interpreted as the true model, misspecification is a relevant case for hypothesis testing, since it naturally arises under the alternative hypothesis that the overidentifying restrictions do not hold.

The MD problem provides a convenient setting for estimating parameters defined by moment conditions that are misspecified. The population version of the MD problem

can be interpreted as selecting—among all the distributions that satisfy the moment conditions—the probability measure that is the closest to the true but unknown distribution F. Formally, the population counterpart of the MD estimator is

$$\theta^*_\gamma := \inf_{\theta\in\Theta} \inf_{Q\in\mathcal{Q}(\theta)} \int \gamma(dQ/dF)dF, \quad \gamma \in \mathcal{G},$$

where, letting $Q \ll F$ denote absolutely continuity of Q w.r.t. F,

$$\mathcal{Q}(\theta) = \left\{ Q : \int q(w,\theta)\mathrm{d}Q = 0, \ \int dQ = 1, \quad Q \ll F \right\}.$$

(See Komunjer and Ragusa (2009) for details on this type of problems). If the model is correctly specified, $F \in \bigcup_{\theta\in\Theta} \mathcal{Q}(\theta)$ and $\theta^*_\gamma = \theta_0$, all $\gamma \in \mathcal{G}$, and $F = \inf_{Q\in\mathcal{Q}(\theta_0)} \int \gamma(dQ/dF)dF$, where θ_0 solves $\int q(w,\theta_0)dF = 0$. When the model is misspecified, $F \notin \bigcup_{\theta\in\Theta} \mathcal{Q}(\theta)$ and, in general, θ^*_γ will vary with $\gamma \in \mathcal{G}$. In correspondence of θ^*_γ, the pseudo-true value, there exists a pseudo-true probability measure, $F^*_\gamma := \inf_{Q\in\mathcal{Q}(\theta^*_\gamma)} \int \gamma(dQ/dF)dF$. This definition of pseudo-true parameters is the extension of White's(1982) definition of pseudo-true parameters and pseudo-true probability measures. A recent discussion of the appropriateness of explicitly assuming misspecification can be found in Maasoumi (2007).

A statistical model may defines both a "true" and "pseudo-true" parameter. Consider the simultaneous equations model, where $y = X\theta_0 + u$ and $X = Z\pi_0 + v$. If $E(Z'u) = 0$, θ_0 can be estimated by exploiting the moment conditions $E[Z'(y - X\theta_0)] = 0$. If the instruments Z are correlated with u the moment conditions are misspecified, and MD estimators estimate a different (pseudo-true) parameter for each choice of γ. Nevertheless, θ_0 is defined whether or not $E[Z'u] = 0$: the definition of θ_0 has to do with the data generating process; the definition of the pseudo-true parameter is related to the estimation procedures.

Under misspecification, some members of the MD class of estimators may fail to converge at usual rate $\sqrt{N}$ to its pseudo-true value. This happens because the equivalence of the MD problem (2) and the saddle point problem (4) may fail to hold for some $\gamma \in \mathcal{G}$, rendering estimation of θ^*_γ unfeasible. The equivalence between (2) and (4) is based entirely on Lagrangian type arguments: the equivalence holds, as seen in Section 2, if there exist $(\hat\eta, \hat\lambda') \in \mathbb{R}^{M+1}$ and $\hat\theta \in \Theta$ such that

$$\hat\eta + \hat\lambda' q_i(\hat\theta) \in \mathcal{A}_\gamma, \text{ for all } i \leqslant N, \tag{12}$$

the constraints are satisfied for $\hat\pi_i = \tilde\gamma_1(\hat\eta + \hat\lambda' q_i(\hat\theta))/N$, and, for all feasible $\bar\pi_i$, $i \leqslant N$, $\sum_{i=1}^N \gamma(N\hat\pi_i)/N \leqslant \sum_{i=1}^N \gamma(N\bar\pi_i)/N$. Under correct specification, there exists a solution, given by $\hat\lambda = o_p(N^{-1/2})$, $\hat\theta = \theta_0 + o_p(N^{-1/2})$, and $\hat\eta = o_p(N^{-1/2})$, that satisfies (12)

w.p.a.1. However, if the model is misspecified and $\mathcal{A}_\gamma$ does not span all $\mathbb{R}$, then (12) may fail to hold even as $N \to \infty$.

For instance, consider the EL estimator. Its divergence ($\gamma^{el}(x) = -\ln x + x - 1$) implies that $\mathcal{A}_\gamma = (-\infty, +1)$. Since the Lagrange multiplier η can be eliminated in this case (see Theorem 3), condition (12) becomes

$$\max_{i \leq N} \tau' q_i(\theta) < 1. \tag{13}$$

We show now that under misspecification there does not exist a $\sqrt{N}-$convergent Lagrange multiplier that solves the EL problem. Let $\dot{\theta}$ and $\dot{\tau}$ denote the solution to the EL problem and the associated Lagrange multiplier, respectively. Suppose that $q(w, \theta)$ is unbounded in every direction, i.e. $\sup_{w \in \mathcal{W}} \upsilon' q(w, \theta) = +\infty$ for all $\|\upsilon\| = 1$ and all $\theta \in \Theta$. As shown in the proof of Lemma 3, Assumption A1 implies that $b_N := \max_{i \leqslant N} \sup_{\theta \in \Theta} \|q_i(\theta)\| = o_p(N^{1/2})$. If the Lagrange multiplier is $N^{-1/2}$ bounded in probability, we can write, $\dot{\tau} = \rho\xi + O_p(N^{-1/2})$, where $\rho = \|\tau_0\|$ and $\xi \in \mathbb{R}^M$, $\|\xi\| = 1$. The parameter τ_0 here is to be seen as a pseudo true value defined as the solution to the dual problem under (MS). Uniformly on $(i = 1, \ldots, N)$,

$$\dot{\tau}' q_i(\dot{\theta}) \leqslant (\rho + O_p(N^{-1/2}))\|q_i(\dot{\theta})\| \leqslant (\rho + O_p(N^{-1/2}))b_N = \rho o_p(N^{1/2}) + o_p(1).$$

To satisfy (13), ρ must be 0 which gives that $\dot{\tau}' q_i(\dot{\theta}) = o_p(1)$ uniformly on $(i = 1, \ldots, N)$. This implies that $\tilde{\gamma}_1(\dot{\tau}' q_i(\dot{\theta}))/N = N^{-1} + o_p(1)$, uniformly as well; but under misspecification, $\pi_i = N^{-1} + o_p(1)$ $(i = 1, \ldots, N)$ and $\dot{\theta}$ are not asymptotic solutions to the MD problem.

In an interesting paper, Schennach (2007) shows that the EL does not converge to the pseudo-true value at rate $\sqrt{N}$ under misspecification. Schennach (2007) proposes a procedure that mixes EL and ET to obtain an estimator that is $\sqrt{N}$ consistent under (MS) and has good higher order properties.

There is a simple way to avoid the pitfalls of MD procedures under MS, that is, choosing divergences with $\mathcal{A}_\gamma = (-\infty, +\infty)$. ET, CUE and all members of the CR family with parameter α equal to an odd integer have $\mathcal{A}_\gamma = (-\infty, +\infty)$. Unfortunately, when the moment conditions are correctly specified, these estimators are not higher order efficient. We identify MD estimators with $\mathcal{A}_\gamma = \mathbb{R}$ that have the same higher order MSE when the model is correctly specified.

We proceed by first deriving functions $\psi \in \mathcal{P}$ with full domain, $D_\psi = \mathbb{R}$ and such that $\psi_3(0) = 2$. We then use Theorem 1 to derive the underlying divergences. We start

by considering a modification of ψ^{et}, that is,

$$\psi(x) = \exp h(x) - xC_1 - C_2, \qquad C_1 = \frac{h_1(0)}{h_1(0) + h_2(0)}, \quad C_2 = \frac{1}{h_1(0) + h_2(0)},$$

where $h : \mathbb{R} \mapsto \mathbb{R}$ is four times continuously differentiable. Since $dom(h) = \mathbb{R}$, then, by construction, $D_\psi = (-\infty, +\infty)$. With the normalization $\exp h(c) - xC_1 - C_2$ belongs to $\mathcal{P}$ if $h_2(x) > h_1(x)^2$, $x \in \mathbb{R}$. It is easy to verify that if $h_3(0) = 1$, the estimator based on $\exp h(x)$ will be higher order efficient. We define the following estimators.

Definition 6 (Quartic Tilting)**.** The Quartic Tilting (QT) estimator solves (4) with

$$\psi^{qt}(x;\nu) = \begin{cases} e^{[(1+x)^4 - 4x - 1]/12} + x - 1 & x > \nu \\ c_1 e^{c_2 x} & x \leqslant \nu \end{cases}$$

where $\nu < 0$, and for $h(x) = e^{[(1+x)^4-4x-1]/12} + x - 1$, $c_1 = h_1(\nu)/(h_1(\nu)^2/h_2(\nu) - h(\nu) + h(\nu))$, $c_2 = e^{c_1\nu}/(h(\nu) + h_1(\nu)^2/h_2(\nu) - h(\nu))$.

Definition 7 (Hyperbolic Tilting)**.** The Hyperbolic Tilting (HT) estimator solves (4) with $\psi^{ht}(x) = \exp\sinh x - 1$.

It is easy to verify that $\psi_3^{qt} = \psi_3^{ht} = 2$. The underlying divergences however cannot be given explicitly because neither the inverse function of ψ^{qt} nor the one of ψ^{ht} have a closed form expression. Nevertheless, as pointed out in Remark 6, the characterization of the divergence in Theorem 1 allows us to obtain at least a graphical representation by numerically inverting ψ_1^{qt} and ψ_1^{ht} and calculating $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$, for all x in the image of ψ_1^{qt} and ψ_1^{ht}. The resulting divergences are plotted in Figure 1, which, for reference, also plots γ^{el} and γ^{et}.

An alternative approach consists in modifying the ψ^{el}. As in Owen (2001), we define, for $\varepsilon \in (0, 1)$,

$$\psi^{el}(x;\varepsilon) = \begin{cases} -\log(1-x) & \text{if } x \in (-\infty, \varepsilon) \\ -\log(1-\varepsilon) + \frac{x-\varepsilon}{1-\varepsilon} + \frac{(x-\varepsilon)^2}{2(1-\varepsilon)^2} & \text{if } x \in [\varepsilon, +\infty) \end{cases}.$$

Owen (2001) points out that as $\varepsilon \to 1$ the function $\psi^{el}(x;\varepsilon)$ converges to $\psi^{el}(x)$; he suggests using $\varepsilon_N = 1 - o(N^{-1})$. Under MS, setting $\varepsilon = 1 - o(N^{-1})$ as $N \to \infty$ will limit the span of $\mathcal{A}_\gamma$ and make the estimator based on $\psi^{el}(x, \varepsilon_N)$ susceptible to the same misspecification issues as EL. However, setting ε to a constant, say $\bar{\varepsilon} \in (0, 1)$, does not affect the higher order asymptotic efficiency of $\psi^{el}(x;\varepsilon)$ and it does not restrict the span of $\mathcal{A}_\gamma$. The divergence underlying $\psi^{el}(x, \bar{\varepsilon})$ can be easily recovered using Theorem 1. The

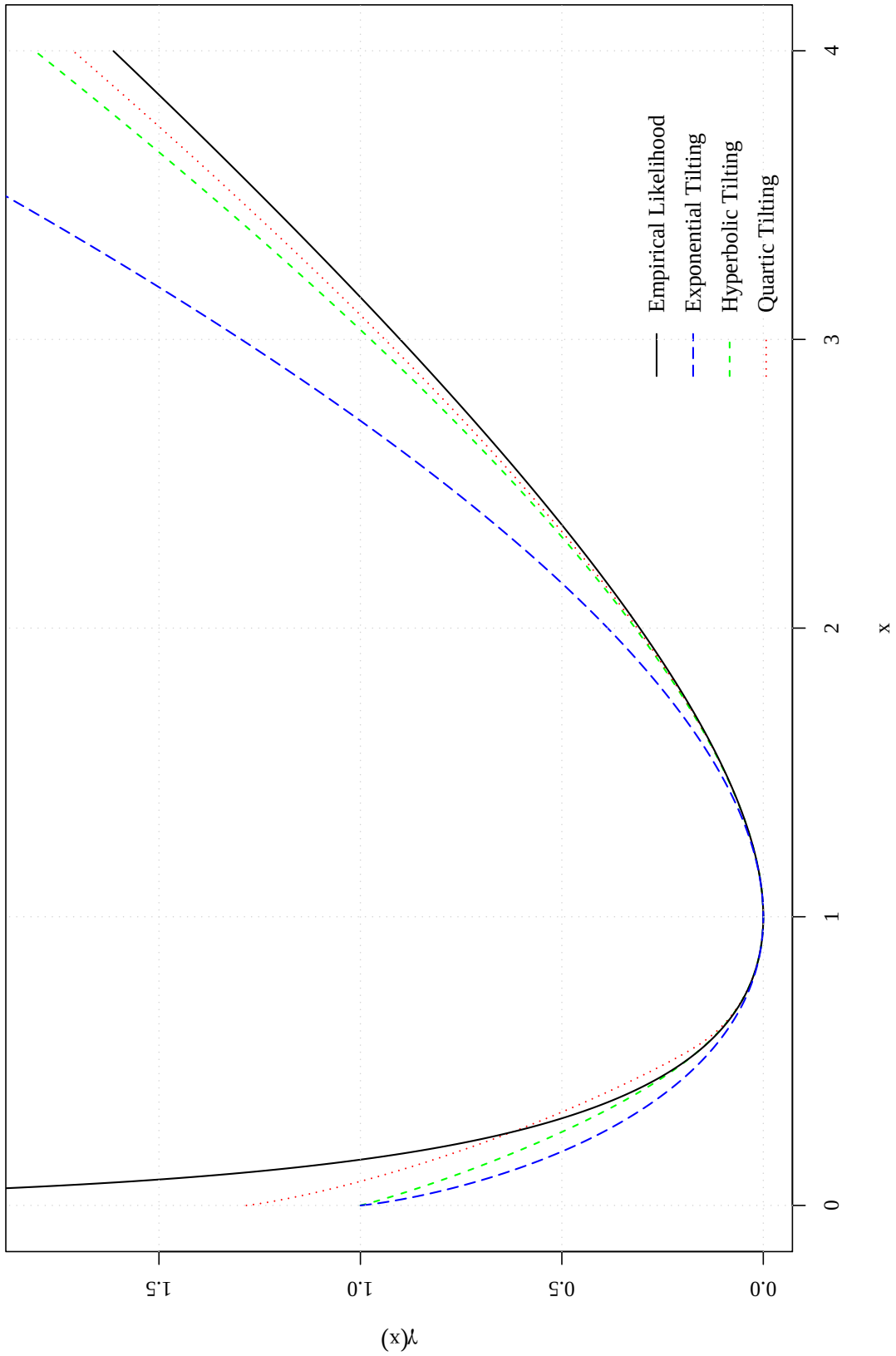


Figure 1: Implied divergence functions. For Quartic Tilting and Hyperbolic Tilting the divergences are obtained by numerically inverting the first derivative of ψ^{qt} and ψ^{ht}on a grid of points covering $(0, 4)$ to obtain $\tilde{\psi}_1^{qt}$ and $\tilde{\psi}_1^{ht}$ and then by calculating $\gamma^{qt}(x) = x\tilde{\psi}_1^{qt}(x) - \psi^{qt}(\tilde{\psi}_1^{qt}(x))$ and $\gamma^{ht}(x) = x\tilde{\psi}_1^{ht}(x) - \psi^{ht}(\tilde{\psi}_1^{ht}(x))$.

first derivative of $\psi^{el}(x,\bar{\varepsilon})$ is given by

$$\psi_1^{el}(x;\bar{\varepsilon}) = \begin{cases} \frac{1}{1-x} & \text{if } x \in (-\infty, \bar{\varepsilon}) \\ \frac{x-\bar{\varepsilon}}{(1-\bar{\varepsilon})^2} - \frac{1}{1-\bar{\varepsilon}} & \text{if } x \in [\bar{\varepsilon}, +\infty) \end{cases},$$

with inverse

$$\tilde{\psi}_1^{el}(x;\bar{\varepsilon}) = \begin{cases} 1 - 1/x & \text{if } x \in \left(0, \frac{1}{1-\bar{\varepsilon}}\right) \\ (x-1)(1-2\bar{\varepsilon}) + \bar{\varepsilon}^2 x & \text{if } x \in \left[\frac{1}{1-\bar{\varepsilon}}, +\infty\right) \end{cases}.$$

Applying the transformation $\gamma^{el}(x;\bar{\varepsilon}) = x\tilde{\psi}_1^{el}(x;\bar{\varepsilon}) - \psi^{el}(\tilde{\psi}_1^{el}(x;\bar{\varepsilon});\bar{\varepsilon})$, we obtain

$$\gamma^{el}(x;\bar{\varepsilon}) = \begin{cases} -\log(x) + x - 1 & \text{if } x \in \left(0, \frac{1}{1-\bar{\varepsilon}}\right] \\ \log(1-\bar{\varepsilon}) + 0.5 + (2\bar{\varepsilon}-1)x + 0.5(\bar{\varepsilon}-1)^2 x^2 & \text{if } x \in \left[\frac{1}{1-\bar{\varepsilon}}, +\infty\right) \end{cases}.$$

The divergence $\gamma^{el}(x,\bar{\varepsilon})$ is plotted in Figure 2 together with $\gamma^{el}(x)$ and $\gamma^{et}(x)$. Although the differences between $\gamma^{el}(x)$ and $\gamma^{el}(x;\bar{\varepsilon})$ are small, the behavior of the underlying estimators is—under misspecification—very different, as shown in the simple numerical example below.

Remark 7. The function $\psi^{el}(x;\varepsilon)$, $\varepsilon = 1 - o(N^{-1})$, is proposed by Owen as a computational device to avoid explicitly constraining the Lagrange multipliers of EL to belong to $\Lambda(\theta)$. Under correct specification, one could let $\varepsilon = 1 - o(N^{-\delta})$, $\delta > 0$, without affecting the asymptotic behavior of the resulting estimator.

Under MS, estimators based on $\gamma^{el}(x,\bar{\varepsilon})$ and $\gamma^{qt}(x;\nu)$ will converge to a pseudo-true value that depends on the specific value of $\bar{\varepsilon}$ and ν used. Under correct specification, only the behavior of the divergence in a neighborhood of 1 is important and, hence, the resulting estimator is asymptotically unaffected by the particular choice $\bar{\varepsilon}$ and ν.

Note that $\psi^{ht}(x)$ and $\psi^{qt}(x;\nu)$ do not satisfy the generalized homogeneity conditions of Theorem 3 and, thus, the estimators obtained from solving the GEL problem in (6) with these functions do not correspond to minimum divergence estimators. To obtain estimators that correspond to solutions to an MD problem one need to solve (4).

7 Monte Carlo experiments

This section describes the design and the results of two Monte Carlo experiments. The objective of the first Monte Carlo is to verify the practical significance of the higher order theoretical results of Section 5. The second experiment is designed to shed light

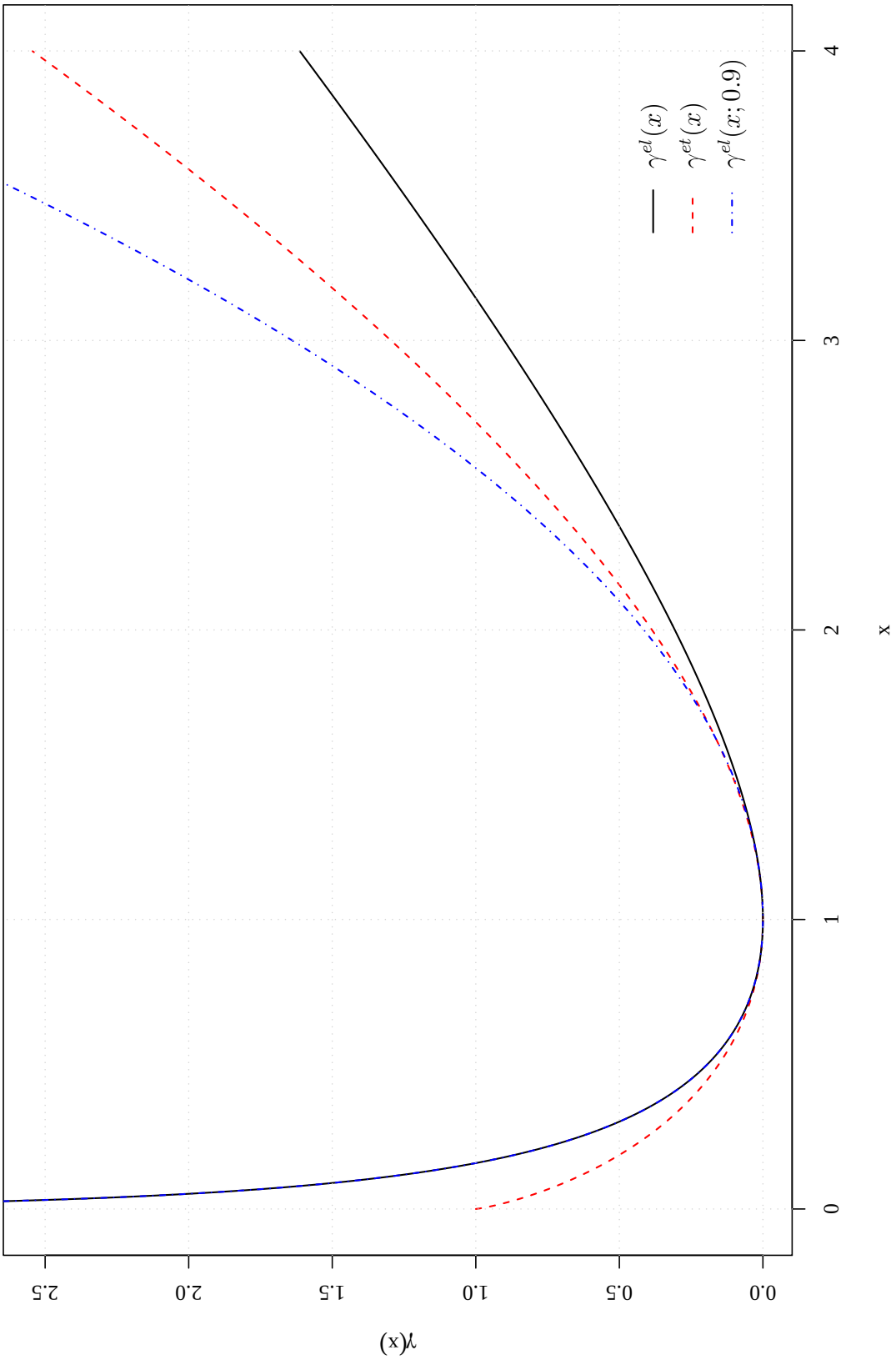


Figure 2: Empirical Likelihood, Exponential Tilting and Modified Empirical Likelihood divergences.

on the behavior of MD under misspecification. In both cases, we consider estimators obtained by solving (4) with $\psi^{el}(x; 1 - N^{-1})$, $\psi^{el}(x, .99)$, $\psi^{et}(x)$, and $\psi^{qt}(x; -1.5)$. These estimators are referred to as EL, EL2, ET and QT, respectively. The efficient GMM estimator is also calculated. The outer maximization step in (4) is carried out by a global optimization routine based on an evolutionary search algorithm (see, Michalewicz et al., 1993 and Mebane, Jr. et al., 2009).[7] We choose a global optimization routine because Guggenberger (2008) shows that Monte Carlo results for EL and ET estimators obtained by derivative based solvers tend to be very misleading.[8]

The first experiment follows Hall and Horowitz (1996). The parameter is defined by the moment condition

$$Eq(Z, X_1, X_2, \theta_0) = 0, \quad q(Z, X_1, X_2, \theta) := Z\{\exp[-.72 - \theta(X_1 + X_2) + 3X_2] - 1\},$$

where $\theta_0 = 3$, X_1 and X_2 are random variables with distribution

$$(X_1, X_2) \sim \mathcal{N}(0, \text{diag}(.16, .16)).$$

The random vector Z varies across two different designs. In the first design $Z = (1, X_2)$; in the second $Z = (1, X_2, X_3, X_4)$, where X_3 and X_4 are independently distributed with standardized χ^2_1 and t-distribution with 5 degrees of freedom, respectively. In the first design, due to the symmetry of the normal distribution, the third moments of $q(Z, X_1, X_2, \theta)$ are equal to zero. The second design adds positive skewness through X_3 and a larger kurtosis through X_4. The sample sizes are $n = 100$, $n = 200$, and $n = 400$. For each design, 10000 replications are performed.

Table 2 summarizes the results. We first comment on the case $Z = (1, X_2)$. At sample size $n = 100$ we see that ET and GMM have higher bias than the other estimators, while EL, EL2 and QT biases are similar and statistically indistinguishable. At this sample size, the MSE of EL2 and QT are not statistically different from the MSE of EL. At $n = 200$ the ET is the only estimator whose bias is larger and significantly different from that of EL. The MSE of EL, EL2, QT and GMM is comparable, while the MSE of GMM and ET is larger and statistically different from the MSE of EL. At $n = 400$ the bias of ET and GMM is still generally larger than the bias of the other estimators, although the

[7] Using a global optimizer with evolutionary algorithm increases computation time considerably. For instance, using a Newton algorithm for the outer step, the computation time of a typical MD estimator for the second design is of about .77 second; with the global optimizer the same estimator is calculated in approximately 46 second.

[8] Guggenberger (2008) finds that, at least for the linear instrumental variables model, solving (4) for EL and ET by evaluating $P_N(\theta) = \min_{(\eta,\lambda')\in\Lambda_N(\theta)} P_N(\eta, \lambda, \theta)$ on a grid $\Theta_B = \{\theta_1, \theta_2, \ldots, \theta_B\}$ and taking $\sup_{\theta\in\Theta_B} P_N(\theta)$ gives results than are considerably different than those obtained with a derivative based optimizer such as Nelder-Mead, BFGS, and quasi-Newton.

	$Z = (1, X_2)$					$Z = (1, X_2, X_3, X_4)$				
	$Bias$	Var	Med	MSE	MAE	$Bias$	Var	Med	MSE	MAE
			$n = 100$					$n = 100$		
EL	0.0567	0.0850	3.0340	0.0882	0.2801	0.1071	0.0920	3.0853	0.1035	0.2952
$EL2$	0.0624‡	0.0837	3.0406	0.0876†	0.2771	0.1061†	0.0908	3.0817	0.1021†	0.2885
ET	0.0754	0.0877	3.0487	0.0934	0.2808	0.1806	0.1158	3.1475	0.1484	0.3130
QT	0.0643†	0.0864	3.0413	0.0905†	0.2834	0.1319	0.1064	3.0955	0.1238	0.3060
GMM	0.0751	0.0878	3.0481	0.0934	0.2818	0.1772	0.1199	3.1376	0.1513	0.3155
			$n = 200$					$n = 200$		
EL	0.0334	0.0401	3.0253	0.0412	0.1972	0.0574	0.0411	3.0466	0.0444	0.1966
$EL2$	0.0303†	0.0404	3.0199	0.0413†	0.1984	0.0629†	0.0422	3.0507	0.0462†	0.2012
ET	0.0407	0.0406	3.0294	0.0422†	0.1964	0.0959	0.0458	3.0833	0.0550	0.2103
QT	0.0333†	0.0393	3.0226	0.0404†	0.1950	0.0604†	0.0425	3.0492	0.0461†	0.2012
GMM	0.0388	0.0407	3.0294	0.0422†	0.1989	0.0948	0.0468	3.0824	0.0558	0.2117
			$n = 400$					$n = 400$		
EL	0.0180	0.0199	3.0130	0.0202	0.1414	0.0313	0.0201	3.0267	0.0211	0.1397
$EL2$	0.0190†	0.0201	3.0140	0.0205†	0.1359	0.0273†	0.0204	3.0224	0.0211†	0.1422
ET	0.0215†	0.0200	3.0178	0.0204†	0.1396	0.0538	0.0212	3.0474	0.0241	0.1470
QT	0.0199†	0.0198	3.0144	0.0202†	0.1414	0.0316†	0.0201	3.0260	0.0211†	0.1400
GMM	0.0226†	0.0198	3.0170	0.0203†	0.1406	0.0550	0.0213	3.0478	0.0243	0.1452

† The null hypothesis that the bias (MSE) of EL is equal to the bias (MSE) of the estimator of the corresponding row cannot be rejected at the 1% significance level.

Table 2: Bias, Variance, Median, Mean Square Error, and Mean Absolute Error of MD-GEL estimators and GMM. Each entry is based on 10000 simulations. EL, EL2, ET, QT denote, in order, MD-GEL estimators obtained from $\psi^{el}(x; \varepsilon_N)$, $\psi^{el}(x, \bar{\varepsilon})$, $\psi^{et}(x)$, and $\psi^{qt}(x; \nu)$, with $\bar{\varepsilon} = 0.99$, $\varepsilon_N = 1 - N^{-1}$, and $\nu = -1.5$. The GMM estimator is the two step efficient estimator that uses the identity matrix as first step weighting matrix.

differences are not statistically significant with respect to the EL. At these sample size, the MSE's of all the estimators are very similar.

In the second design, $Z = (1, X_2, X_3, X_4)$, the picture is quite different. As expected, ET and GMM have larger biases at all sample sizes, while EL2 and QT have biases that are comparable with EL. The only exception being the case $n = 100$ in which QT has significant larger bias than EL, but much smaller than the bias of ET and GMM. The same results hold true for the analysis of the MSE. EL, EL2 and QT have very similar MSE (except QT for $n = 100$), while ET and GMM tend to have larger MSE.

It is interesting to reconcile the Monte Carlo results with the higher order asymptotic findings. In the first design, due to the symmetry of $q(Z, X_1, X_2, \theta)$, all the MD-GEL estimators considered have the same theoretical higher order bias regardless of the value of ψ_3. The Monte Carlo experiment suggests that estimators with $\psi_3 = 2$ tend to perform better for smaller sample sizes, as it takes $n = 400$ for ET to approach the bias of EL. In line with Theorem 8, the MSE of estimators with $\psi_3 = 2$ is found to be very close in both designs, while ET has a different (and, in this case larger) MSE in the second design.[9]

To verify that QT, HT and the estimator based on the modified EL divergence behave well under misspecification we run a small scale Monte Carlo experiment, considering the same design of Schennach (2007). The moment condition is

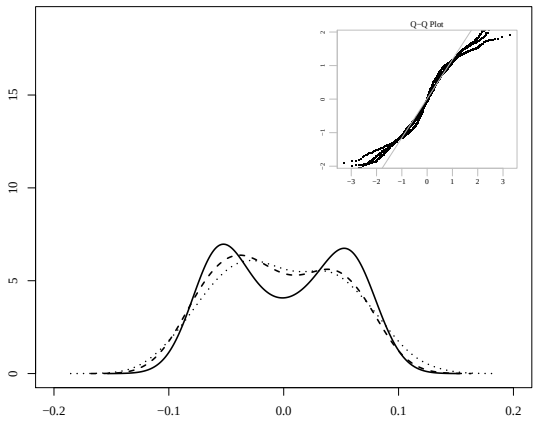
$$Eq(X, \theta_0) = 0, \quad q(X, \theta) = \begin{bmatrix} X - \theta \\ X^2 - \theta^2 - 1 \end{bmatrix}. \tag{14}$$

We consider again two designs. In the first, $X \sim \mathcal{N}(0, 1)$, when the moment condition is correctly specified. In the second, X is drawn from $X \sim \mathcal{N}(0, 0.64)$ and the moment condition is misspecified.[10] We run 1000 replications for each of the three sample sizes considered $N = \{1000, 2500, 5000\}$.

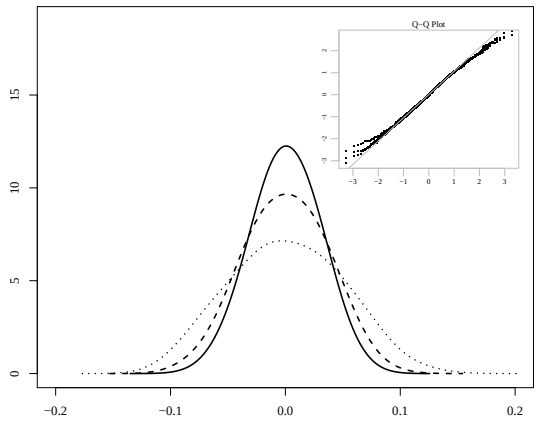
The results of this Monte Carlo experiment are summarized in Figure 3. For the three sample sizes considered, we plot sampling densities and normal quantile-to-quantile plots of the standardized estimators. The sampling density of EL depicted in Figure 3(a) shows clear signs of non-normality which are exacerbated as the sample size increases. The variance of this estimator does not appear to be shrinking at rate N: the variance of this estimator is equal to 0.00293 when $n = 1000$, 0.00262 when $n = 5000$, and 0.00270 when $n = 1000$. The densities of ET, QT, GMM, and of EL2 are in line with the sampling distribution of a $\sqrt{N}$ consistent estimator: inspection of the quantile-to-

[9]Sensitivity analysis with respect to the value of $\bar{\varepsilon}$ and ν shows that the results remain virtually unchanged when different values for those two parameters are used.

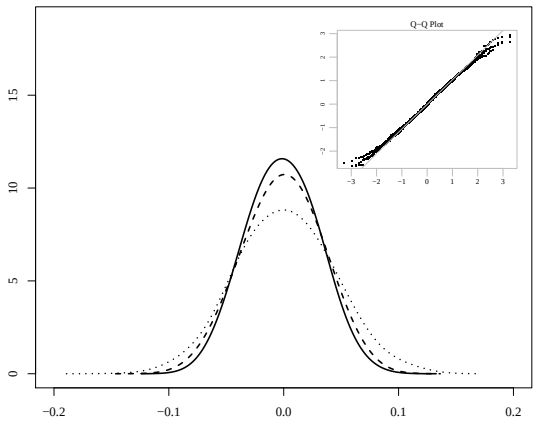
[10]Note however that for this particular design, it can be shown that the "true" and the "pseudo-true" parameters coincide, $\theta_0 = \theta^*_\gamma = 0$, $\gamma \in \mathcal{G}$.



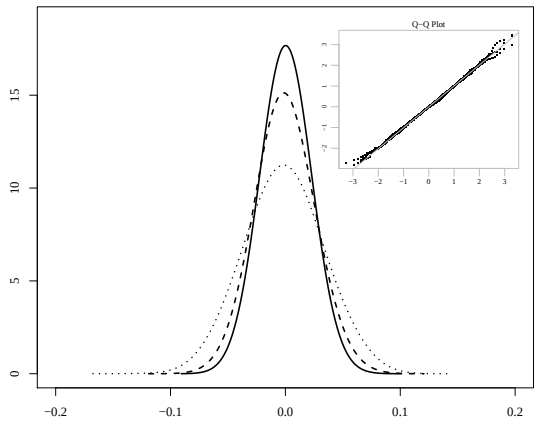
(a) Empirical Likelihood: $\psi^{el}(x; 1 - N^{-1})$



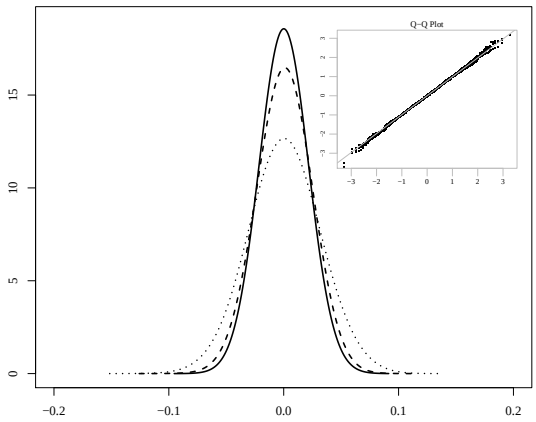
(b) Empirical Likelihood: $\psi^{el}(x; .99)$



(c) Exponential Tilting



(d) Quartic Tilting



(e) GMM

Figure 3: Behavior of MD and GMM estimators under misspecification. Each plot reports the sampling distribution of a particular estimator for the three sample sizes considered $\{1000, 2500, 5000\}$, shown in dotted, dashed, and solid lines, respectively. Each panel also contains plots of the quantiles of the normalized sampling distribution of the estimator against the quantiles of the standard normal distribution.

quantile plots clearly indicate no signs of departures from normality. Also, the variances of these estimators shrink at rate N. For instance, the variance of QT goes from 0.00039 when $n = 5000$ to 0.00018 when $n = 10000$.

We recognize that it can be extremely misleading and somewhat deceiving to provide simulation results under misspecification as the results will be much more sensitive to the simulation design. We choose the example above because is used by Schennach (2007) and we do not intend to draw general conclusions out of it. Nevertheless, we cannot avoid to notice that—at least for this example—there exist MD-GEL estimators that have the same higher order MSE of EL but whose behavior is not pathological when the moment conditions are misspecified.

8 Conclusion

This paper studies the Minimum Divergence class of estimators for econometric models specified through moment conditions. We extend the analysis of Newey and Smith (2004) and show that MD estimators defined in terms of strictly convex divergences can always be calculated as solutions to a computationally tractable optimization problem. The problem is similar to the optimization setting that defines the GEL estimators of Newey and Smith and it is identical when a condition on the inverse function of the first derivative is satisfied. The MD framework allows a coherent presentation and unification of a series of tests that have been presented as alternative to the overidentification test of Hansen (1982). MD estimators that have the same higher order bias of EL share the same higher order MSE. Since EL is higher order efficient, this result implies that there are many higher order efficient MD estimators. Schennach (2007) shows that the asymptotic distribution of the EL may not be normal if the moment condition is misspecified. We give examples of estimators that are second order efficient under correct specification and do not misbehave when the moment condition does not hold exactly.

There are many important aspects of MD estimators that still remain to be explored. Estimators who have small bias and are higher order efficient are often preferable. However, concerns for real applications include the small sample properties of test procedures (in terms of size and power) and of confidence intervals (in terms of coverage). The only work that deals with optimality of overidentification test statistics is Kitamura (2001), where it is demonstrated that tests based on the EL objective function are uniformly most powerful in the Hoeffding sense. Unfortunately, the empirical size of overidentification tests based on EL is, in simulations, often found to be far from the nominal level. Further, different divergences give rise to test statistics that perform very differently in terms of size. What is the combination estimator/test that performs better (and in which

statistical environment) is still an open question. Chen and Cui (2007) have shown that EL is Bartlett correctable under the setting consider in this paper. It would be interesting to extend their analysis and derive conditions on the class of divergences under which Bartlett correctability can be proved. In Monte Carlo simulations not reported here tests of overidentifying restrictions based on the divergences proposed in Section 6 tend to perform extremely well in terms of size. We leave exploration of this aspect for future work.

A Mathematical Appendix

Lemma 1. *Suppose $\psi \in \mathcal{G}$. Then the function $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$ belongs to $\mathcal{G}$, its domain is $D_\gamma = (l, u)$, $l = \lim_{u \searrow a_\psi} \psi_1(u)$ and $u = \lim_{u \nearrow b_\psi} \psi_1(u)$, and $\Lambda_N^\dagger(\theta) = \Lambda_N(\theta)$, $T_N^\dagger(\theta) = \Lambda_N(\theta)$ and $T_N^\dagger(\theta) = T_N(\theta)$ for $\theta \in \Theta$.*

Proof. Strict convexity of ψ on D_ψ implies that the inverse function of $\psi_1(x)$ is well defined for every $x \in D_\psi$, $\tilde{\psi}_1 : S \to (a_\psi, b_\psi)$, $S = (a_{\psi'}, b_{\psi'})$, $a_{\psi'} = \lim_{u \searrow a_\psi} \psi_1(u)$ and $b_{\psi'} = \lim_{u \nearrow b_\psi} \psi_1(u)$. The function $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$ is defined on S, and, by twice continuous differentiability of ψ on D_ψ, it is twice continuously differentiable on S with

$$\gamma_1(x) = \tilde{\psi}_1(x) + x\frac{d\tilde{\psi}_1(x)}{dx} - \psi_1(\tilde{\psi}_1(x))\frac{d\tilde{\psi}_1(x)}{dx} = \tilde{\psi}_1(x).$$

The inverse function $\tilde{\psi}_1(x)$ is strictly increasing on S. Therefore, $\gamma_1(x)$ is strictly increasing on S and $\gamma(x)$ is strictly convex on S. The normalizations $\psi_1(0) = 1$ and $\psi(0) = 0$ imply that $\tilde{\psi}_1(1) = 0$ and $\gamma_1(1) = \tilde{\psi}_1(1) - \psi(\tilde{\psi}_1(1)) = 0$. This and strictly convexity imply that γ attains its minimum 0 at $x = 1$, thus $\gamma(x) \geq 0$ for $x \in S$. Since $\psi_2(x) > 0$ on $x \in D_\psi$ the inverse function theorem gives that $\gamma_2(x) = 1/\psi_2(\tilde{\psi}_1(x))$; since $\psi_2(0) = 1$, and $\tilde{\psi}_1(1) = 0$ it follows that $\gamma_2(1) = 1$. The last assertion follows from noting that $\{y : \gamma_1(x) = y, x \in S\} = \text{dom}\psi_1$. *Q.E.D.*

Lemma 2. *Suppose $\gamma \in \mathcal{G}$. Then the function $\psi(x) = x\tilde{\gamma}_1(x) - \gamma(\tilde{\gamma}_1(x))$ belongs to $\mathcal{P}$, its domain is $D_\psi = (l, u)$, $l = \lim_{u \searrow a_\gamma} \gamma_1(u)$ and $u = \lim_{u \nearrow b_\gamma} \gamma_1(u)$, and $\Lambda_N^\dagger(\theta) = \Lambda_N(\theta)$, $T_N^\dagger(\theta) = \Lambda_N(\theta)$ and $T_N^\dagger(\theta) = T_N(\theta)$ for $\theta \in \Theta$.*

Proof. Strict convexity of γ on D_γ implies that the inverse function of $\gamma_1(x)$ is defined for $x \in D_\gamma$, $\tilde{\gamma}_1 : S \to (a_\gamma, b_\gamma)$, $S = (a_{\gamma'}, b_{\gamma'})$, $a_{\gamma'} = \lim_{u \searrow a_\gamma} \gamma_1(u)$ and $b_{\gamma'} = \lim_{u \nearrow b_\gamma} \gamma_1(u)$. The function $\psi(x) = x\tilde{\gamma}_1(x) - \gamma(\tilde{\gamma}_1(x))$ is defined on S, and, by twice continuous differentiability of γ on (a_γ, b_γ), it is twice continuously differentiable on S with

$$\psi_1(x) = \tilde{\gamma}_1(x) + x\frac{d\tilde{\gamma}_1(x)}{dx} - \gamma_1(\tilde{\gamma}_1(x))\frac{d\tilde{\gamma}_1(x)}{dx} = \tilde{\gamma}_1(x).$$

The inverse function $\tilde{\gamma}_1(x)$ is strictly increasing on S. Therefore, $\psi_1(x)$ is strictly increasing on S and $\psi(x)$ is strictly convex on S. The normalizations $\gamma_1(1) = 0$ and $\gamma(1) = 0$ imply that $\tilde{\gamma}_1(0) = 1$ and $\psi_1(0) = \tilde{\gamma}_1(0) - \gamma(\tilde{\gamma}_1(0)) = 1$. This and strictly convexity imply that ψ it attains its minimum 0 at $x = 0$, thus $\gamma(x) \geq 0$ for $x \in S$. Since $\gamma_2(x) > 0$ on $x \in D_\gamma$ the inverse function theorem gives that $\psi_2(x) = 1/\gamma_2(\tilde{\gamma}_1(x))$; since $\gamma_2(1) = 1$, and $\tilde{\gamma}_1(0) = 1$ it follows that $\psi_2(0) = 1$. The last assertion follows from noting that $\{y : \gamma_1(x) = y, x \in S\} = \text{dom}\psi_1$. *Q.E.D.*

Proof of Theorem 1

Apply Lemma 1 to conclude that, for $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$, $\gamma \in \mathcal{G}$, $\psi_1(x) = \tilde{\gamma}_1(x)$, $x \in D_\psi$, $\Lambda_N(\theta) = \Lambda_N^\dagger(\theta)$ for $\theta \in \Theta$. We need to show that $\hat{\Gamma}_N \leqslant \sum_{i=1}^N \gamma(Np_i)/N$ for all feasible p_i, $i = 1, \ldots, N$. First notice that $\gamma(N\hat{\pi}_i) = N\hat{\pi}_i\tilde{\psi}_1(N\hat{\pi}_i) - \psi(\tilde{\psi}_1(N\hat{\pi}_i))$; summing over $(i = 1, \ldots, N)$, using $\psi_1(x) = \tilde{\gamma}_1(x)$, $\sum_{i=1}^N \psi_1(\hat{\eta} + \hat{\lambda}'\hat{q}_i)/N = 1$, and $\sum_{i=1}^N \psi_1(\hat{\eta} + \hat{\lambda}'\hat{q}_i)\hat{q}_i/N = 0$ give $\widehat{\Gamma}_N = -\widehat{P}_N$. Let $(\bar{\eta}, \bar{\lambda}')' = \arg\min_{(\eta,\lambda')' \in \Lambda_N^\dagger(\bar{\theta})} P_N(\eta, \lambda, \bar{\theta})$ and $\bar{\pi}_i = \psi_1(\bar{\eta} + \bar{\lambda}'q_i(\bar{\theta}))/N$, $(i = 1, \ldots, N)$. Optimality of $\hat{\eta}$, $\hat{\lambda}$ and $\hat{\theta}$ implies that $\widehat{P}_N \geqslant P_N(\bar{\eta}, \bar{\lambda}, \bar{\theta})$ for all $\bar{\theta} \in \Theta$. We have that $\gamma(N\bar{\pi}_i) = N\bar{\pi}_i\tilde{\psi}_1(N\bar{\pi}_i) - \psi(\tilde{\psi}_1(N\bar{\pi}_i))$. Summing over $(i = 1, \ldots, N)$ and noting that $\sum_{i=1}^N \psi_1(\bar{\eta} + \bar{\lambda}'q_i(\bar{\theta}))/N = 1$, and $\sum_{i=1}^N \psi_1(\bar{\eta} + \bar{\lambda}'q_i(\bar{\theta}))\hat{q}_i/N = 0$ imply that $\sum_{i=1}^N \gamma(N\bar{\pi}_i)/N = P_N(\bar{\eta} + \bar{\lambda}'q_i(\bar{\theta}))$ which, in turns, implies

$$-\widehat{P}_N = \widehat{\Gamma}_N \leqslant \sum_{i=1}^N \gamma(N\bar{\pi}_i)/N = -P_N(\bar{\eta} + \bar{\lambda}'q_i(\bar{\theta})). \tag{A.1}$$

This last result establishes that $\hat{\pi}_i$, $(i = 1, \ldots, N)$, solve the MD problem for all the feasible weights of type $\tilde{\gamma}_1(\eta + \lambda'q_i(\theta))/N$, which are optimal for $\theta \in \Theta$. *Q.E.D.*

Proof of Theorem 2

Apply Lemma 2 to obtain that, for $\psi(x) = x\tilde{\gamma}_1(x) - \gamma(\tilde{\gamma}_1(x))$, $\gamma \in \mathcal{G}$, $\psi_1(x) = \tilde{\gamma}_1(x)$, $x \in D_\psi$, $\Lambda_N(\theta) = \Lambda_N^\dagger(\theta)$ for $\theta \in \Theta$. For every $s \in D_\psi$ and every $t \in D_\gamma$, the Fenchel inequality (see Rockafellar, 1970, pag. 218) yields

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

Let $\hat{p}_i$, $(i = 1, \ldots, N)$, be feasible at $\theta = \hat{\theta}$, that is $N\hat{p}_i \in (a_\gamma, b_\gamma)$, $\sum_{i=1}^N \hat{p}_i = 1$, $\sum_{i=1}^N \hat{p}_i\hat{q}_i = 0$. Evaluating the Fenchel inequality at $t = N\hat{p}_i$ and $s = \hat{\eta} + \hat{\lambda}'\hat{q}_i$, summing over $(i = 1, \ldots, N)$, and using $\tilde{\gamma}_1(x) = \psi_1(x)$ for all $x \in (a_\psi, b_\psi)$, $\sum_{i=1}^N \psi_1(\hat{\eta} + \hat{\lambda}'\hat{q}_i)/N = 1$, and $\sum_{i=1}^N \psi_1(\hat{\eta} + \hat{\lambda}'\hat{q}_i)\hat{q}_i/N = 0$ give $\widehat{P}_N = -\widehat{\Gamma}_N \geqslant -\sum_{i=1}^N \gamma(N\hat{p}_i)/N$. We need to prove that $\hat{\theta}$ is optimal. Let $(\bar{\eta}, \bar{\lambda}')' = \arg\min_{(\eta,\lambda')' \in \Lambda_N(\bar{\theta})} P_N(\eta, \lambda, \bar{\theta})$ for $\bar{\theta} \in \Theta$. We then have that

$$P_N(\bar{\eta}, \bar{\lambda}, \bar{\theta}) = -\sum_{i=1}^N \gamma(N\tilde{\gamma}_1(\bar{\eta} + \bar{\lambda}'q_i(\bar{\theta})))/N.$$

But $\sum_{i=1}^N \gamma(N\tilde{\gamma}_1(\bar{\eta} + \bar{\lambda}'q_i(\bar{\theta})))/N \geqslant \widehat{\Gamma}_N$ and, thus, $\widehat{P}_N \geqslant P_N(\bar{\eta}, \bar{\lambda}, \bar{\theta})$, as required. *Q.E.D.*

Proof of Theorem 3

Apply Lemma 1 to obtain that, for $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$, $\gamma \in \mathcal{G}$, $\psi_1(x) = \tilde{\gamma}_1(x)$, $x \in D_\psi$, $\Lambda_N(\theta) = \Lambda_N^\dagger(\theta)$ for $\theta \in \Theta$. Let p_i, $(i = 1, \ldots, N)$, feasible for $\tilde{\theta} \in \Theta$:

$$Np_i \in (a_\gamma, b_\gamma), \quad \sum_{i=1}^N p_i \tilde{q}_i = 0.$$

Evaluating the Fenchel inequality at $s = \tilde{\tau}'\tilde{q}_i$ and $t = Np_i$ yields

$$\psi(\tilde{\tau}'\tilde{q}_i) = \tilde{\tau}'\tilde{q}_i\tilde{\gamma}_1(\tilde{\tau}'\tilde{q}_i) - \gamma(\tilde{\gamma}_1(\tilde{\tau}'\tilde{q}_i)) \geqslant \tilde{\tau}'\tilde{q}_i p_i - \gamma(Np_i).$$

Summing over $(i = 1, \ldots, N)$, using $\tilde{\gamma}_1(x) = \psi_1(x)$, and $\sum_{i=1}^N \psi_1(\hat{\eta} + \hat{\lambda}'\hat{q}_i)\hat{q}_i/N = 0$ give

$$\widetilde{P}_N = -\widetilde{\Gamma}_N \geqslant -\sum_{i=1}^N \gamma(Np_i)/N$$

The last inequality implies that $\widetilde{\Gamma}_N \leqslant \sum_{i=1}^N \gamma(Np_i)/N$ and, thus, $\tilde{\pi}_i$ is optimal among all the weights that do not impose $\sum_{i=1}^N p_i = 1$ and, hence, not necessarily feasible for 2. For any feasible weights, say ς_i $(i = 1, \ldots, N)$, $\sum_{i=1}^N \varsigma_i = 1$, $\sum_{i=1}^N \varsigma_i \tilde{q}_i = 0$, it must hold

$$\Gamma_N(\tilde{\eta}, \tilde{\lambda}, \tilde{\theta}) = -\min_{(\eta, \lambda') \in \Lambda_N^\dagger(\tilde{\theta})} P_N(\eta, \theta, \tilde{\theta}) \leqslant \sum_{i=1}^N \gamma(N\varsigma_i)/N.$$

By convexity of $\gamma(x)$, $\gamma(x) \geqslant \gamma(y) + \gamma_1(y)(x - y)$ for all $x, y \in (a_\gamma, b_\gamma)$. Hence,

$$\Gamma_N(\tilde{\eta}, \tilde{\lambda}, \tilde{\theta}) \geqslant \widetilde{\Gamma}_N^\dagger + \sum_{i=1}^N \gamma_1(N\tilde{\omega})(\tilde{\pi}_i - \tilde{\omega}_i).$$

Let $\delta = N/\sum_{i=1}^N \tilde{\pi}_i$. We have that $\gamma_1(N\tilde{\omega}_i) = a(\delta) + h(\delta)\gamma_1(N\tilde{\pi}_i)$. By feasibility of $\tilde{\pi}_i$ and $\tilde{\omega}_i$, it follows that $\sum_{i=1}^N (\tilde{\pi}_i - \tilde{\omega}_i) = 0$ and $\sum_{i=1}^N \tilde{q}_i(\hat{\pi}_i - \tilde{\omega}_i) = 0$. Thus,

$$\begin{aligned}\sum_{i=1}^N \gamma_1(N\tilde{\omega})(\tilde{\pi}_i - \tilde{\omega}_i) &= a(\delta)\sum_{i=1}^N (\tilde{\pi}_i - \tilde{\omega}_i) + h(\delta)\sum_{i=1}^N \gamma_1(\tilde{\gamma}_1(\tilde{\tau}'\tilde{q}_i))(\tilde{\pi}_i - \tilde{\omega}_i) \\ &= h(\delta)\tilde{\tau}'\sum_{i=1}^N \tilde{q}_i(\tilde{\pi}_i - \tilde{\omega}_i) = 0,\end{aligned}$$

Therefore, $\widetilde{\Gamma}_N^\dagger \leqslant \Gamma_N(\tilde{\eta}, \tilde{\lambda}, \tilde{\theta}) \leqslant \sum_{i=1}^N \gamma(N\varsigma_i)/N$. But since $\Gamma_N(\tilde{\eta}, \tilde{\lambda}, \tilde{\theta})$ is optimal at $\theta = \tilde{\theta}$ it must be that $\widetilde{\Gamma}_N^\dagger = \Gamma_N(\tilde{\eta}, \tilde{\lambda}, \tilde{\theta})$. To show that $\tilde{\theta}$ is optimal for the MD problem note that, for $P_N(\tau^*, \theta) = \min_{\tau \in T_N^\dagger(\theta)} P_N(\tau, \theta)$, $\pi_i^* = \gamma_1(\tau^{*\prime} q_i(\theta))/N$, $\omega_i^{*\prime} = \pi_i^*/\sum_{i=1}^N \pi_i^*$, we

have

$$-\widetilde{\Gamma}_N = -\widetilde{\Gamma}_N^{\dagger} = \widetilde{P}_N \geqslant P_N(\tau^*, \theta) = -\sum_{i=1}^{N} \gamma(N\pi_i^*)/N = -\sum_{i=1}^{N} \gamma(N\omega_i^*)/N,$$

from which the result follows. *Q.E.D.*

Proof of Theorem 4

Lemma 1 gives the three first three conclusions. Let ς_i, $(i = 1, \ldots, N)$, feasible for $\theta = \tilde{\theta}$:

$$N\varsigma_i \in (a_\gamma, b_\gamma), \quad \sum_{i=1}^{N} \varsigma_i \tilde{q}_i = 0.$$

For $s = \tilde{\tau}'\tilde{q}_i$ and $t = N\varsigma_i$, the Fenchel inequality gives

$$\psi(\tilde{\tau}'\tilde{q}_i) = \tilde{\tau}'\tilde{q}_i\tilde{\gamma}_1(\tilde{\tau}'\tilde{q}_i) - \gamma(\tilde{\gamma}_1(\tilde{\tau}'\tilde{q}_i)) \geqslant \tilde{\tau}'\tilde{q}_i\varsigma_i - \gamma(N\varsigma_i).$$

By summing over $(i = 1, \ldots, N)$, using feasibility of ς_i, $\psi_1(x) = \tilde{\gamma}_1(x)$ we obtain

$$\sum_{i=1}^{N} \gamma(N\tilde{p}_i)/N \leqslant \sum_{i=1}^{N} \gamma(N\varsigma_i)/N$$

The proof is completed by showing, as in the proof of Theorem 1 and Theorem 3 that $\tilde{\theta}$ is optimal. *Q.E.D.*

Lemma 3. *Suppose Assumption A1 holds. Let*

$$\Lambda_N^s = \left\{(\eta, \lambda') : |\eta| \leqslant N^{-1+\xi},\ \|\lambda\| < N^{-1/2+\zeta},\ (\xi, \zeta) > 0\right\}.$$

Then $\sup_{\theta\in\Theta,(\eta,\lambda')\in\Lambda_N^s, i\leqslant N} |\eta + \lambda' q_i(\theta)| = o_p(1)$.

Proof. Apply Lemma 3 in Owen (1990) to deduce that

$$b_N := \sup_{i\leq N,\theta\in\Theta} \|q_i(\theta)\| = o(N^{1/2})$$

w.p.1 and that there exists a $\delta > 0$ such that $b_N = O(N^{1/2-\delta})$ w.p.1. Then

$$\sup_{i\leqslant N,\theta\in\Theta,(\eta,\lambda')\in\Lambda_N^s,} |\eta + \lambda' q_i(\theta)| \leqslant N^{-\xi} + \|\lambda\| b_N = N^{-\xi} + N^{-1/2+\zeta} O(N^{1/2-\delta}) = O(N^{\zeta-\delta}),$$

with probability one. Since ζ is arbitrary, the result follows for $\zeta < \delta$.

Q.E.D.

Lemma 4. *If Assumption A1 holds,* $(\eta(\theta_0), \lambda(\theta_0)') := \arg\min_{(\eta,\lambda')\in\Lambda_N(\theta_0)} P_N(\eta, \lambda, \theta_0)$ *exists w.p.a.1,* $\eta(\theta_0) = O_p(N^{-1})$*,* $\lambda(\theta_0) = O_p(N^{-1/2})$*, and* $P_N(\eta(\theta_0), \lambda(\theta_0), \theta_0) = O_p(N^{-1})$*.*

Proof. Let Λ_N^s be as defined as in Lemma 3, $(\tilde{\eta}, \tilde{\lambda}') := \arg\min_{(\eta,\lambda')\in\Lambda_N^s} P_N(\eta, \lambda, \theta)$, $\tilde{\upsilon}_i = t\tilde{\eta} + t\tilde{\lambda}' q_i(\theta_0)$, some $t \in [0, 1]$. By Lemma A1 and continuous differentiability of ψ we have that $\max_{i\leqslant N} \psi_2(\tilde{\upsilon}_i) = 1$ for all $t \in [0, 1]$ w.p.a.1. Positive definitiveness of Ω and a Taylor expansion imply that

$$\begin{aligned} 0 \;\leqslant P_N(0, \tilde{\lambda}, \theta_0) \;&= \tilde{\lambda}' q_n(\theta_0) + \tilde{\lambda}' \Bigg(\sum_{i=1}^N \psi_2(\tilde{\upsilon}_i) q_i(\theta_0) q_i(\theta_0)'/N \Bigg) \tilde{\lambda} \\ &\leqslant \|\tilde{\lambda}\| \|q_n(\theta_0)\| - C\|\tilde{\lambda}\|^2, \quad \text{w.p.a.1,} \end{aligned}$$

where C is a strictly positive constant. The inequality $C\|\tilde{\lambda}\|^2 \leqslant \|\tilde{\lambda}\| \|q_n(\theta_0)\|$ and the CLT yield $\tilde{\lambda} = O_p(N^{-1/2}) = o_p(N^{-1/2+\zeta})$. By optimality of $(\tilde{\eta}, \tilde{\lambda}')$, $0 = P_N(0, 0, \theta_0) \leqslant P_N(\tilde{\eta}, \tilde{\lambda}, \theta_0)$. Notice that $P_N(\tilde{\eta}, \tilde{\lambda}, \theta_0) \geqslant \sum_{i=1}^N \tilde{\lambda}' q_i(\theta_0)/N$, since it holds that $\psi(x) \geqslant \psi(y) + \psi_1(y)(x - y)$ for all $(x, y) \in D_\psi$. Therefore, a Taylor expansion gives the following

$$\begin{aligned} 0 \;&\leqslant\; -\tilde{\lambda}' \Bigg(\sum_{i=1}^N \psi_2(\tilde{\upsilon}_i) q_i(\theta_0) q_i(\theta_0)'/N \Bigg) \tilde{\lambda} - \tilde{\eta}^2 \sum_{i=1}^N \psi_2(\tilde{\upsilon}_i)/N - \tilde{\eta}\tilde{\lambda} \sum_{i=1}^N \psi_2(\tilde{\upsilon}_i) q_i(\theta_0)/N \\ &\leqslant\; -\tilde{\eta}^2 - \tilde{\eta}\tilde{\lambda}' q_n(\theta_0) \leqslant -\tilde{\eta}^2 - |\tilde{\eta}| \|\tilde{\lambda}\| \|q_n(\theta_0)\| \leqslant -\tilde{\eta}^2 + |\tilde{\eta}| \|\tilde{\lambda}\| \|q_n(\theta_0)\|, \quad \text{w.p.a.1.} \end{aligned}$$

This implies that $\tilde{\eta} = O_p(N^{-1}) = o_p(N^{-1+\xi})$ for all $\xi < 1$. It follows that $(\tilde{\eta}, \tilde{\lambda}') \in \text{Int}(\Lambda_N^s)$ w.p.a.1 and by convexity of $\Lambda_N(\theta_0)$ we have that w.p.a.1

$$(\eta(\theta_0), \lambda(\theta_0)') = \arg \min_{(\eta,\lambda')\in\Lambda_N(\theta_0)} P_N(\eta, \lambda, \theta_0) = (\tilde{\eta}, \tilde{\lambda}') = \arg \min_{(\eta,\lambda')\in\Lambda_N^s} P_N(\eta, \lambda, \theta_0),$$

yielding the first and second assertions of the theorem. The third assertion follows by expanding $P_N(\eta(\theta_0), \lambda(\theta_0), \theta_0)$ around $(\eta(\theta_0), \lambda(\theta_0)') = (0, 0')$ to obtain $P_N(\eta(\theta_0), \lambda(\theta_0), \theta_0) = \lambda(\theta_0)' q_n(\theta_0) + O_p(N^{-1}) = O_p(N^{-1})$.

Q.E.D.

Proof of Theorem 5

The proof is based on Wald (1949) and Wolfowitz (1949). The basic argument goes as follows. Let $B(\delta, \theta_0)$ denote a ball of radius $\delta > 0$ around θ_0. Inside $\Theta \backslash B(\delta, \theta_0)$, the sample objective function is bounded away from the maximum of the population objective function evaluated at the true parameter value, w.p.a.1. The maximum of the sample

objective function is by definition not smaller than its value at the true parameter value. The latter converges—by LLN—to the population objective function evaluated at θ_0. Hence, the maximum of the sample objective function is unlikely to occur in $\Theta \backslash B(\delta, \theta_0)$ for large enough N. This is tantamount to consistency of the maximum of the sample objective function.

Let

$$C_N = \big\{ w : \sup_{\theta \in \Theta} \|q(w,\theta)\| \leqslant N^{1/2} \upsilon \text{ and } \sup_{\theta \in \Theta} -\|q(w,\theta)\| \geq N^{1/2} \ell \big\},$$

for some $\ell < a_\psi < \upsilon < b_\psi$. Let $u(\theta) = q_i(\theta)/(1 + \|q_i(\theta)\|)$. By optimality of $\eta(\theta)$ and $\lambda(\theta)$, we have that

$$P_N(\eta(\theta), \lambda(\theta), \theta) \leqslant Q_N(\theta) := \sum_{i=1}^{N} \psi(-N^{-1/2} u(\theta)' q_i(\theta) 1(w_i \in C_N))/N.$$

For some $t \in [0,1]$, the mean value theorem implies

$$N^{1/2} Q_N(\theta) = \sum_{i=1}^{N} -u(\theta)' q_i(\theta)/N + \sum_{i=1}^{N} R_i(\theta, t)/N, \tag{A.2}$$

where

$$\begin{aligned} R_i(\theta, t) &= u(\theta)' q_i(\theta)(1 - \mathbb{I}(w_i \in C_N)) \\ &\quad + N^{-1/2} \psi_2(-N^{-1/2} t u(\theta)' q_i(\theta) \, \mathbb{I}\,(w_i \in C_N)) u(\theta)' q_i(\theta) q_i(\theta)' u(\theta) \, \mathbb{I}\,(w_i \in C_N). \end{aligned}$$

Repeated application of the Cauchy-Schwartz inequality, convexity of ψ, $\sup_{\theta \in \Theta} \|u(\theta)\| \leqslant 1$, $\sup_{\theta \in \Theta} \|u(\theta)\|^2 \leqslant 1$ yields

$$|R_i(\theta, t)| \leqslant \sup_{\theta \in \Theta} \|q_i(\theta)\| (1 - \max_{i \leqslant N} \mathbb{I}(w_i \in C_N)) + N^{-1/2} \psi_2(m) \sup_{\theta \in \Theta} \|q_i(\theta)\|^2 \max_{i \leqslant N} \mathbb{I}(w_i \in C_N)$$

for some $m \in (a_\psi, b_\psi)$. Now, since $1 - \max_{i \leqslant N} \mathbb{I}(w_i \in C_N) = o_p(1)$, by Assumption A1 the remainder term in (A.2) is uniformly $O_p(N^{-1/2})$. Thus,

$$N^{1/2} Q_N(\theta) = -\sum_{i=1}^{N} u(\theta)' q_i(\theta)/N + o_p(1), \quad \text{uniformly in } \Theta. \tag{A.3}$$

It follows that

$$\sup_{\theta\in\Theta} N^{1/2}P_N \leqslant \sup_{\theta\in\Theta} N^{1/2}Q_N(\theta) = \sup_{\theta\in\Theta} N^{1/2}\sum_{i=1}^{N} -u(\theta)'q_i(\theta)/N + o_p(1).$$

Compactness of Θ, continuity $u(\theta)'q_i(\theta)$ at each $\theta \in \Theta$ w.p.1, $E\big[\sup_{\theta\in\Theta}\|q_i(\theta)\|\big] < \infty$, and $|N^{1/2}u(\theta)'q_i(\theta)| \leqslant N^{1/2}\sup_{\theta\in\Theta}\|q_i(\theta)\|$ imply

$$\sup_{\theta\in\Theta}\left\|-\sum_{i=1}^{N} u(\theta)'q_i(\theta)/N - E\big[-u(\theta)'q_i(\theta)\big]\right\| = o_p(1). \tag{A.4}$$

Since $-E[u(\theta)'q_i(\theta)] = -E\big[q_i(\theta)/(1+\|q_i(\theta)\|\big] < 0$, continuity of $E\big[-u(\theta)'q_i(\theta)\big]$ implies that there exists for every $\delta > 0$ a number $h(\delta) > 0$ such that $\sup_{\theta\in\Theta\setminus B(\theta,\delta_0)} E[-u(\theta)'q_i(\theta)] \leqslant -h(\delta)$ and

$$\sup_{\theta\in\Theta\setminus B(\theta_0,\delta)} N^{1/2}P_N(\eta(\theta),\lambda(\theta),\theta) \leqslant \sup_{\theta\in\Theta\setminus B(\theta_0,\delta)} E[-u(\theta)'q_i(\theta)] \leqslant -h(\delta),$$

which together with (A.3) and (A.4) yield

$$P\left\{\sup_{\theta\in\Theta\setminus B(\theta_0,\delta)} P_N(\eta(\theta),\lambda(\theta),\theta) > -N^{-1/2}h(\delta)\right\} < \delta/2. \tag{A.5}$$

From convexity of $\psi(x)$ and optimality of the Lagrange Multipliers, we have that

$$P_N(\eta(\theta_0),\lambda(\theta_0),\theta_0) \geqslant \eta(\theta_0) + \sum_{i=1}^{N}\lambda(\theta_0)'q_i(\theta_0)/N.$$

Apply Lemma 4 to deduce that $\lambda(\theta_0) = O_p(N^{-1/2})$. Therefore, by convexity of $\psi(x)$,

$$P_N(\eta(\theta_0),\lambda(\theta_0),\theta_0) \geqslant \sum_{i=1}^{N}\lambda(\theta_0)'q_i(\theta_0)/N = O_p(N^{-1/2})O_p(N^{-1/2}) = o_p(N^{-1/2}).$$

If $\hat\theta \in \Theta\setminus B(\theta_0,\delta)$, then

$$\sup_{\theta\in\Theta\setminus B(\theta_0,\delta)} N^{1/2}P_N(\eta(\theta),\lambda(\theta),\theta) = N^{1/2}P_N(\eta(\hat\theta),\lambda(\hat\theta),\hat\theta) \geqslant N^{1/2}P_N(\eta(\theta_0),\lambda(\theta_0),\theta_0) = o_p(1).$$

Therefore, eventually,

$$P\left\{P_N(\eta(\theta_0),\lambda(\theta_0),\theta_0) < -N^{-1/2}h(\delta)\right\} < \delta/2. \tag{A.6}$$

We have

$$\begin{aligned}\left\{\hat{\theta} \in \Theta \backslash B(\theta_0, \delta)\right\} &\subset \left\{\sup_{\Theta \backslash B(\theta_0,\delta)} P_N(\eta(\theta), \lambda(\theta), \theta) > P_N(\eta(\theta_0), \lambda(\theta_0), \theta_0)\right\} \\ &\subset \left\{\sup_{\Theta \backslash B(\theta_0,\delta)} P_N(\eta(\theta), \lambda(\theta), \theta) > -N^{-1/2} h(\delta)\right\} \\ &\quad \cup \left\{P_N(\eta(\theta_0), \lambda(\theta_0), \theta_0) < -N^{-1/2} h(\delta)\right\}.\end{aligned}$$

Thus, for all $\delta > 0$, there exists a N_δ such that for all $N \geqslant N_\delta$ such that

$$P\left\{\hat{\theta} \in \Theta \backslash B(\theta_0, \delta)\right\} \leqslant \delta.$$

Since δ is arbitrary, $\hat{\theta}$ is consistent. *Q.E.D.*

Proof of Theorem 6

From Theorem 5 and Lemma 3 the first order conditions $\sum_{i=1}^{N} Q_i(\hat{\beta}) = 0$, are satisfied w.p.a.1. A mean value expansion of the first order conditions around $\theta = \theta_0$, $\lambda = 0$, and $\eta = 0$ gives

$$0 = \begin{pmatrix} 0 \\ q_n \\ 0 \end{pmatrix} + \begin{pmatrix} \Omega & G' & 0 \\ G & 0 & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \hat{\lambda} \\ \hat{\theta} - \theta_0 \\ \hat{\eta} \end{pmatrix} + o_p\left(\frac{1}{\sqrt{N}}\right).$$

Solving for $(\hat{\lambda}', (\hat{\theta} - \theta_0)', \hat{\eta})'$ by using the formula for the inverse of a block matrix yields

$$\sqrt{N}\begin{pmatrix} \hat{\lambda} \\ \hat{\theta} - \theta_0 \\ \hat{\eta} \end{pmatrix} = \begin{pmatrix} P & H' & 0 \\ H & -\Sigma & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 0 \\ \sqrt{N} q_n \\ 0 \end{pmatrix} + o_p(1),$$

Thus,

$$\sqrt{N}\hat{\eta} = o_p(1), \qquad \sqrt{N}\hat{\lambda} = P\sqrt{N} q_n + o_p(1), \qquad \sqrt{N}(\hat{\theta} - \theta_0) = H\sqrt{N} q_n + o_p(1).$$

The result then follows from the asymptotic normality of $\sqrt{N} q_n$ implied by A1-A2 and the fact that $P\Omega P = P$ and $H\Omega H' = \Sigma$.

Proof of Theorem 7

The consistency part follows by noting that $\tilde{\gamma}_1(\hat{\eta}+\hat{\lambda}'\hat{q}_i)/N = \psi_1(\hat{\eta}+\hat{\lambda}'\hat{q}_i)/N = N^{-1} + O_p(N^{-1})$ and, thus, $\hat{p}_A = \sum_{i=1}^N 1(w \in A)\tilde{\gamma}_1(\hat{\eta}+\hat{\lambda}'\hat{q}_i)/N = \sum_{i=1}^N 1(w \in A)/N + o_p(1)$ and by the WLLN $\hat{p}_A \xrightarrow{p} E[1(w \in A)] = p_A$. First, notice that the MD estimator for the augmented parameter vector $\beta = (p_A, \theta)$ is the solution to

$$\min_{\beta,\pi} \sum_{i=1}^N \gamma(N\pi_i), \ s.t. \ \sum_{i=1}^N \pi_i q_i(\theta) = 0, \ \sum_{i=1}^N \pi_i 1(w_i \in A) - p_A = 0, \ \sum_{i=1}^N \pi_i = 1.$$

It is easy to verify that $\sum_{i=1}^N \pi_i 1(w_i \in A) - p_A = 0$ is not binding and, thus, the MD estimator of p_A is $\hat{p}_A = \sum_{i=1}^N 1(w \in A)\hat{\pi}_i$ where $\hat{\pi}_i$ $(i = 1, \ldots, N)$ are the solutions to the MD problem that does not impose the constraint and optimizes over θ and π_i $(i = 1, \ldots, N)$. Asymptotic normality and semiparametric efficiency follows from Theorem (6); the asymptotic variance of β is then given by

$$V(\beta) := \begin{pmatrix} 1 & 0 \\ 0 & G \end{pmatrix} \begin{pmatrix} p_A(1-p_A) & -E(1(w \in A)q(w,\theta)') \\ -E(1(w \in A)q(w,\theta)) & \Omega \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & G' \end{pmatrix}.$$

By simple algebra it can be show that the $(1,1)$ element of $V(\beta)$ is V_A. *Q.E.D.*

Proof of Proposition 1

Taylor expansion of the first order condition that determine the Lagrange multiplier $\hat{\eta}$, $\|\hat{\lambda}\| = O_p(N^{-1/2})$, uniform convergence of $\sum_{i=1}^N q_i(\theta)q_i(\theta)'$ to Ω, and Lemma 3 give

$$0 = \sum_{i=1}^N \psi_1(\hat{\eta}+\hat{\lambda}'\hat{q}_i) - 1 = \hat{\eta} + \hat{\lambda}'\hat{q} + (\psi_3/2)\hat{\lambda}'\Omega\hat{\lambda} + O_p(N^{-3/2}).$$

Substituting $\hat{\lambda} = -\Omega^{-1}\hat{q} + O_p(N^{-1})$—which is obtained by a similar expansion from the first order conditions for λ— we have

$$\hat{\eta} = (1 - \psi_3/2)\hat{q}'\Omega^{-1}\hat{q} + O_p(N^{-3/2}).$$

Thus, for $\psi_3 \neq 2$,

$$\frac{N\hat{\eta}}{(1-\psi_3/2)} = N\hat{q}'\Omega\hat{q} + o_p(1).$$

$GEL(\tilde{\theta})$ expands as

$$\widetilde{P}_N(\tilde{\theta},\tilde{\tau}) = \tilde{\tau}'\tilde{q} + \tilde{\tau}'\Omega\tilde{\tau}/2 + o_p(N^{-1}) = -q_n(\tilde{\theta})'\Omega^{-1}q_n(\tilde{\theta})/2 + o_p(N^{-1}).$$

$D(\hat{\theta})$ expands as

$$\widehat{P}_N(\hat{\theta},\bar{\eta},\hat{\lambda}) = \hat{\lambda}'\hat{q} + \hat{\lambda}'\Omega\hat{\lambda}/2 + O_p(N^{-2}) = -q_n(\hat{\theta})'\Omega q_n(\hat{\theta})/2 + o_p(N^{-1}).$$

Also, $LM(\hat{\theta}) = LM(\tilde{\theta}) + o_p(1)$ and $LM(\tilde{\theta}) = Nq_n(\tilde{\theta})'\Omega^{-1}q_n(\tilde{\theta}) + o_p(1)$. The result follows from from as in Hansen (1982) that $Nq_n(\hat{\theta})'\Omega^{-1}q_n(\hat{\theta}) \xrightarrow{d} \chi^2(M-K)$ for any consistent estimator of θ_0. *Q.E.D.*

B Asymptotic Expansions

For the sake of notational clarity, we use—through this appendix—the following conventions for the partial derivatives: $\nabla_r q_i^j$ denotes the partial derivatives of the j-th element of q with respect of the $r-M$ element of θ. That is, $\nabla_r q_i^j = \partial q_i^j/\partial\beta^r = \partial q_i^j/\partial\theta^{r-M}$. The first partial derivatives are given by

$$\frac{\partial Q_i^j}{\partial\beta^k} = q_i^j q_i^k, \qquad \frac{\partial Q_i^j}{\partial\beta^r} = \nabla_r q_i^j, \qquad \frac{\partial Q_i^r}{\partial\beta^j} = \nabla_r q_i^j, \qquad \frac{\partial Q_i^s}{\partial\beta^t} = 0.$$

The partial second null derivatives are:

$$\begin{array}{lll} \dfrac{\partial Q_i^j}{\partial\beta^k\partial\beta^\ell} = \psi_3 q_i^j q_i^k q_i^\ell, & \dfrac{\partial Q_i^j}{\partial\beta^k\partial\beta^r} = q_i^k\nabla_r q_i^j + q_i^j\nabla_r q_i^k, & \dfrac{\partial Q_i^j}{\partial\beta^r\partial\beta^s} = \nabla_{r,s} q_i^j, \\ \dfrac{\partial Q_i^r}{\partial\beta^k\partial\beta^\ell} = q_i^k\nabla_r q_i^\ell + q_i^\ell\nabla_r q_i^k, & \dfrac{\partial Q_i^r}{\partial\beta^k\partial\beta^s} = \nabla_{r,s} q_i^k, & \dfrac{\partial Q_i^r}{\partial\beta^s\partial\beta^t} = 0. \end{array} \quad \text{(A.7)}$$

The partial third null derivatives are:

$$\begin{aligned} \frac{\partial Q_i^j}{\partial\beta^k\partial\beta^\ell\partial\beta^m} &= \psi_4 q_i^j q_i^k q_i^\ell q_i^m, \\ \frac{\partial Q_i^j}{\partial\beta^k\partial\beta^\ell\partial\beta^r} &= \psi_3\left(q_i^k q_i^\ell\nabla_r q_i^j + q_i^j q_i^\ell\nabla_r q_i^k + q_i^j q_i^k\nabla_r q_i^\ell\right), \\ \frac{\partial Q_i^j}{\partial\beta^k\partial\beta^r\partial\beta^s} &= \nabla_s q_i^k\nabla_r q_i^j + \nabla_s q_i^j\nabla_r q_i^k + q_i^j\nabla_{r,s} q_i^k + q_i^k\nabla_{r,s} q_i^j, \\ \frac{\partial Q_i^j}{\partial\beta^r\partial\beta^s\partial\beta^t} &= \nabla_{r,s,t} q_i^j, \end{aligned} \quad \text{(A.8)}$$

$$
\begin{aligned}
\frac{\partial Q_i^r}{\partial\beta^k\partial\beta^\ell\partial\beta^m} &= \psi_3\left(q_i^\ell q_i^m\nabla_r q^k + q_i^k q_i^m\nabla_r q^\ell + q_i^k q_i^\ell\nabla_r q^m\right), \\
\frac{\partial Q_i^r}{\partial\beta^k\partial\beta^\ell\partial\beta^s} &= \psi_3\left(\nabla_s q_i^k\nabla_r q_i^\ell + \nabla_s q_i^\ell\nabla_r q_i^t + q_i^k\nabla_{r,s} q_i^\ell + q_i^\ell\nabla_{r,s} q_i^k\right), \\
\frac{\partial Q_i^r}{\partial\beta^k\partial\beta^s\partial\beta^t} &= \nabla_{r,s,t} q_i^k, \\
\frac{\partial Q_i^r}{\partial\beta^s\partial\beta^t\partial\beta^u} &= 0.
\end{aligned}
$$

Define the following quantities

$$\kappa_a = E(Z_a), \quad \kappa_{a,b} = E(Z_a Z_b), \quad \kappa_{a,b,c} = E(Z_a Z_b Z_c), \ldots$$

and, similarly,

$$\kappa_{ab,c} = E(Z_{ab} Z_c), \quad \kappa_{ab,cd} = E(Z_{ab} Z_{cd}), \quad \kappa_{abc,d} = E(Z_{abc} Z_d), \ldots$$

Recall μ^{jk}, μ^{jr}, and μ^{rs} denote the element of the inverse of the matrix of expected derivatives of the moment conditions evaluated at β_0, which is given in matrix form by

$$\begin{pmatrix} P & H' & 0 \\ H & -\Sigma & \vdots \\ 0 & \cdots & 1 \end{pmatrix}.$$

Given the index convention $j, k, \ell, m, n, o \in \{1, \ldots, M\}$, $r, s, t, u, v, w \in \{M+1, \ldots, M+K\}$, we have that $\{\mu^{jk}\}$ denotes the element of P, $\{\mu^{jr}\}$ and $\{\mu^{rj}\}$ denote the elements of H' and H, respectively, and $\{\mu^{rs}\}$ denote the element of Σ. Also, since $\kappa_{j,k} = E[q^j q^k]$, $\kappa_{j,k}$ denotes the element of Ω. The matrix identities

$$P\Omega P = P, \quad H\Omega P = 0_{k\times m}, \quad H\Omega H' = \Sigma,$$

can be rewritten element-wise in summation notation as

$$\mu^{jk}\mu^{\ell m}\kappa_{j,\ell} = \mu^{km}, \tag{A.9a}$$

$$\mu^{jr}\mu^{k\ell}\kappa_{j,k} = 0, \tag{A.9b}$$

$$\mu^{jr}\mu^{ks}\kappa_{j,k} = -\mu^{rs}. \tag{A.9c}$$

The identities (A.9a)-(A.9c), which are central in deriving the results of this appendix, also hold for certain permutations of the indexes because of the symmetry of P, Ω, and Σ. In particular, symmetry implies that $\mu^{jk} = \mu^{kj}$, $\mu^{\ell m} = \mu^{m\ell}$ and $\kappa_{j,k} = \kappa_{k,j}$.

Derivation of Equations (11)

Using (10) and expanding the sums we obtain $O_p(N^{-2})$ expansions for $\hat{\beta}^j - \beta^j$, $j \in \{1, \ldots, M\}$ and $\hat{\beta}^r - \beta^r$, $r \in \{M, \ldots, M+K\}$ are $\hat{\beta}^j - \beta^j = -\mu^{jk} Z_k / \sqrt{N} + b_N^j / N + O_p(N^{-3/2})$, and $\hat{\beta}^r - \beta^r = -\mu^{rj} Z_j / \sqrt{N} + b_N^r / N + O_p(N^{-3/2})$, where

$$b_N^j = \mu^{ja}\mu^{bk} Z_{ab} Z_k - \mu^{j,k,\ell} Z_k Z_\ell / 2, \quad \text{and} \quad b_N^r = \mu^{ra}\mu^{bj} Z_{ab} Z_j - \mu^{r,j,k} Z_j Z_k / 2.$$

Taking expectations, using the identities (A.9a)-(A.9c), and exploiting the structure of the null derivatives of $Q(\beta)$ we obtain

$$\begin{aligned} E[b_N^r] = {} & \mu^{rj}\mu^{k\ell}\kappa_{jk,\ell} + \mu^{rj}\mu^{sk}\kappa_{js,k} + \mu^{rs}\mu^{k\ell}\kappa_{sk,\ell} \\ & - (\mu^{rj}\mu^{k\ell}\mu_{jk\ell} + \mu^{rj}\mu^{st}\mu_{jst} + \mu^{rs}\mu^{jk}\mu_{rjk})/2, \end{aligned} \tag{A.10}$$

and

$$\begin{aligned} E[b_N^j] = {} & \mu^{jk}\mu^{\ell m}\kappa_{k\ell,m} + \mu^{jk}\mu^{s\ell}\kappa_{ks,\ell} + \mu^{js}\mu^{k\ell}\kappa_{sk,\ell} \\ & -(\mu^{jk}\mu^{\ell m}\mu_{k\ell m} + \mu^{jk}\mu^{st}\mu_{kst} + \mu^{js}\mu^{k\ell}\mu_{jk\ell})/2. \end{aligned}$$

The expressions for $\kappa_{..,.}$ and $\mu_{...}$ appearing in the previous expression are given by

$$\begin{aligned} \kappa_{jk,\ell} = E(q_i^j q_i^k q_i^\ell) \quad \kappa_{js,k} = E(\nabla_s q_i^j q_i^k) \quad \kappa_{sk,\ell} = E(\nabla_s q_i^k q_i^\ell) \\ \mu_{jk\ell} = \psi_3 E(q_i^j q_i^k q_i^\ell) \quad \mu_{jst} = E(\nabla_{s,t} q_i^j) \quad \mu_{rjk} = E(q_i^j \nabla_r q^k) + E(q_i^k \nabla_r q^j). \end{aligned}$$

Noting that by symmetry of μ^{jk} we have $\mu^{rs}\mu^{jk}\mu_{rjk} = \mu^{rs}\mu^{jk}E(q_i^j \nabla_r q^k) + \mu^{rs}\mu^{kj}E(q_i^k \nabla_r q^j) = 2 \times \mu^{rs}\mu^{jk}E(q_i^k \nabla_r q^j)$, it follows that

$$\begin{aligned} E[b_N^r] &= \mu^{rj}\mu^{k\ell}\mu_{jk,\ell} + \mu^{rj}\mu^{sk}\mu_{js,k} + \mu^{rs}\mu^{k\ell}\mu_{sk,\ell} \\ &\quad - (\mu^{rj}\mu^{k\ell}\mu_{jk\ell} + \mu^{rj}\mu^{st}\mu_{jst} + \mu^{rs}\mu^{jk}\mu_{rjk})/2 \\ &= (1 - \psi_3/2)\mu^{rj}\mu^{kl}E(q_i^j q_i^k q_i^\ell)/N + \mu^{rj}\left[\mu^{sk}E\left(q_i^k \nabla_s q_i^j\right) - \mu^{st}E\left(\nabla_{s,s} q_i^j\right)/2\right], \end{aligned}$$

giving, thus, the desired result. The expression for $E[b_N^j]$ can be derived similarly.

Q.E.D.

Proof of Theorem 8

The $O(N^{-2})$ MSE of the MD estimator obtained from the objective function ψ is given by

$$MSE_2^{r,s}(\hat{\beta}_\psi) = \frac{1}{N}E(i^r_{\psi,N}i^s_{\psi,N}) + \frac{1}{N\sqrt{N}}\left[\text{cov}(b^r_{\psi,N}, i^s_N) + \text{cov}(b^s_{\psi,N}i^r_N)\right] + \frac{1}{N^2}\left[\text{cov}\left(i^r_N, c^s_{\psi,N}\right) + \text{cov}\left(i^s_N, c^r_{\psi,N}\right)\right].$$

the difference between the MSE of the MD estimator obtained from ψ and the MSE of the MD estimator obtained from ψ' is thus given by the difference of the corresponding terms in the relative expansions:

$$\begin{aligned} MSE_2^{r,s}(\hat{\beta}_\psi) - MSE_2^{r,s}(\hat{\beta}_{\psi'}) &= \frac{1}{N\sqrt{N}}\left[\text{cov}(b^r_{\psi,N} - b^r_{\psi',N}, i^s_N) + \text{cov}(b^s_{\psi,N} - b^s_{\psi,N}i^s_N)\right] \\ &+\frac{1}{N^2}\left[\text{cov}\left(i^r_N, c^s_{\psi,N} - c^s_{\psi',N}\right) + \text{cov}\left(i^s_N, c^r_{\psi,N} - c^r_{\psi',N}\right)\right] \end{aligned}$$

The proof proceeds in two steps. In the first step it is shown that if $\psi_3 = \psi'_3$, $\text{cov}(i^s_N, b^r_{\psi,N} - b^r_{\psi,N}) = 0$, since $b^r_{\psi,N} - b^r_{\psi,N} = 0$. This implies that $MSE_2^{r,s}(\hat{\beta}_\psi) - MSE_2^{r,s}(\hat{\beta}_{\psi'}) = o(N^{-2})$ if i_N is uncorrelated with $(c^s_{\psi,N} - c^s_{\psi,N})$ up to terms $o(1)$. The second step of the proof demonstrates that this is indeed the case.

First Step:

The expression for $b^r_{\psi,N}$, $b^r_{\psi,N} = b^r_N = \mu^{ra}\mu^{bj}Z_{ab}Z_j - \mu^{r,j,k}Z_jZ_k/2$, depends on ψ_3 through $\mu^{r,j,k} = \mu^{ra}\mu^{jb}\mu^{kc}\mu_{abc}$ and in particular through $\mu_{abc} = \partial Q^a(\beta)/\partial\beta^b\partial\beta^c$. Inspection of the second null derivatives reveals that terms that differ between two MD estimators are those that involve ψ_3, that is, $\mu_{jk\ell} = \psi_3 E[q^j_i q^k_i q^\ell_i]$, for all $j, k, \ell \in \{1, \ldots, M\}$. Thus,

$$b^r_{\psi,N} - b^r_{\psi',N} = (\psi_3 - \psi'_3)\mu^{r\ell}\mu^{jm}\mu^{kn}E[q^\ell_i q^m_i q^n_i]/2.$$

It follows that when $\psi_3 = \psi'_3$, $b^r_{\psi,N} - b^r_{\psi',N} = 0$ giving the first conclusion of the proof. This also implies that $\text{cov}(i^r_N, b^s_{\psi,N} - b^s_{\psi',N})/N\sqrt{N} = 0$ and that the higher order MSE reduces then to

$$MSE_2^{r,s}(\hat{\beta}_\psi) - MSE_2^{r,s}(\hat{\beta}_{\psi'}) = \frac{1}{N^2}\left[\text{cov}\left(i^r_N, c^s_{\psi,N} - c^s_{\psi',N}\right) + \text{cov}\left(i^s_N, c^r_{\psi,N} - c^r_{\psi',N}\right)\right].$$

Second Step:

Using the fact that $\kappa_{ab,r,j} = \kappa_{ab,j,r} = \kappa_{ab,j,k} = 0$, the covariance between i^r and $c^r_{\psi,N}$ is given by

$$\begin{aligned} \text{cov}(i^r_N, c^s_{\psi,N}) &= \mu^{rj}\mu^{sa}\mu^{bc}\mu^{dk}\kappa_{ab,cd,j,k} - \mu^{rj}\mu^{s,k,b}\mu^{c\ell}\kappa_{j,bc,k,\ell} + \mu^{rj}\mu^{sa}\mu^{b,k,\ell}\kappa_{ab,j,k,\ell} \\ &-\mu^{rj}\mu^{s,k,\ell}\mu^{m,a,b}\mu_{ab}\kappa_{j,k,\ell,m} + \mu^{rj}\mu^{sa}\mu^{kb}\mu^{\ell c}\kappa_{abc,j,k,\ell}/2 + \mu^{rj}\mu^{s,k,\ell,m}\kappa_{j,k,\ell,m}/6. \end{aligned}$$

Terms entering $\text{cov}(i_N^r, c_{\psi,N}^s)$ that depend on ψ are $\mu^{b,k,\ell}$, $\mu^{m,a,b}$,$\kappa_{abc,j,k,\ell}$, and $\mu^{s,k,\ell,m}$. By inspection of the null derivatives, it is easy to see that the first two terms depend only on ψ_3; also,

$$\kappa_{abc,j,k,\ell} = E\left[\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial Q_i^a}{\partial\beta^b\partial\beta^c} - \sqrt{N}\mu_{abc}\right)\frac{1}{\sqrt{N}}\sum_{i=1}^{N}q_i^j\frac{1}{\sqrt{N}}\sum_{i=1}^{N}q_i^k\frac{1}{\sqrt{N}}\sum_{i=1}^{N}q_i^\ell\right],$$

only depends on ψ_3 for indexes $a, b, c \in (2, \ldots, M+1)$. Thus, when $\psi_3 = \psi_3'$,

$$\text{cov}(i_N^r, c_{\psi,N}^s - c_{\psi',N}^s) = (\mu^{rj}\mu^{sn} + \mu^{sj}\mu^{rn})\mu^{ko}\mu^{\ell p}\mu^{ml}(\mu_{nopl}^{\psi} - \mu_{nopl}^{\psi'})\kappa_{j,k,\ell,m}/6,$$

where μ_{nopl}^{ψ} and $\mu_{nopl}^{\psi'}$ denote the terms that differ in the expansion when $\psi_4 \neq \psi_4'$. Now,

$$\kappa_{j,k,\ell,m} \;=\; E\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}q_i^j\frac{1}{\sqrt{N}}\sum_{i=1}^{N}q_i^k\frac{1}{\sqrt{N}}\sum_{i=1}^{N}q_i^\ell\frac{1}{\sqrt{N}}\sum_{i=1}^{N}q_i^m\right],$$

which, by independence of w_i, can be rewritten as

$$\kappa_{j,k,\ell,m} = E(q_i^j q_i^k q_i^\ell q_i^m)/N + \kappa_{j,k}\kappa_{\ell,m}[3] + \kappa_{k,j,\ell}\kappa_m[4] + \kappa_j\kappa_k\kappa_{\ell m}[6] + \kappa_j\kappa_k\kappa_\ell\kappa_m,$$

where the notation $[\cdot]$ denotes the sum over partitions of four indexes, so that, for example,

$$\kappa_{j,k}\kappa_{\ell,m}[3] = \kappa_{j,k}\kappa_{\ell,m} + \kappa_{j,\ell}\kappa_{k,m} + \kappa_{j,m}\kappa_{\ell,m}.$$

Because $\kappa_j = 0$, the last four terms in the last expression for $\kappa_{j,k,\ell,m}$ are all 0. By A3, $E(q_i^j q_i^k q_i^\ell q_i^m) = O(1)$, and thus

$$\kappa_{j,k,\ell,m} = \kappa_{j,k}\kappa_{\ell,m}[3] + O(N^{-1}).$$

Substituting this into the expression for $\text{cov}(i_N^r, c_{\psi,N}^s - c_{\psi',N}^s)$ and using $\mu_{jk\ell m}^{\psi} = \psi_4 E[q_i^j q_i^k q_i^\ell q_i^m]$ yields

$$\text{cov}(i_N^r, c_{\psi,N}^s - c_{\psi',N}^s) = (\psi_4 - \psi_4')\Delta^{r,s} + O(N^{-1}),$$

where

$$\Delta^{r,s} = (\mu^{rj}\mu^{sn} + \mu^{sj}\mu^{rn})\mu^{ko}\mu^{\ell p}\mu^{ml}E[q_i^n q_i^o q_i^p q_i^l]\kappa_{j,k}\kappa_{\ell,m}[3]/6.$$

Applying the identity in (A.9b) we have that $\mu^{rj}\mu^{ko}\kappa_{j,k} = \mu^{rj}\mu^{\ell p}\kappa_{j,\ell} = \mu^{rj}\mu^{ml}\kappa_{j,m} = 0$. Therefore, we have that $\Delta^{r,s} = 0$ which gives the desired result that $\text{cov}(i_N^r, c_{\psi,N}^s - c_{\psi',N}^s) =$

$O(N^{-1})$ and, thus, $MSE_2^{r,s}(\hat{\beta}_\psi) - MSE_2^{r,s}(\hat{\beta}_{\psi'}) = o(N^{-2})$. *Q.E.D.*

Proof of Theorem 9

We need to show only the first assertion of the theorem, as the second follows from equation (9). Notice that the bias correction can be written as $\widehat{E[b^r_{\psi,N}]} = u^r(d_N(\hat{\beta}_\psi))$, where $d_N(\beta)$ is a vector collecting the sample moments constituting $\widehat{E[b^r_{\psi,N}]}$, that is,

$$d_N(\beta) = (\{\mu_{ab,N}\}, \{A_N^{j,k,\ell}(\beta)\}, \{B_N^{j,s,k}(\beta)\}, \{C_N^{j,k,\ell}(\beta)\}),$$

and $u^r(\cdot)$ is twice continuously differentiable function. Let $d_0 = E[d_N(\beta_0)]$. Note that $E[b^r_{\psi,N}] = u^r(d_0)$. A Taylor expansion of $u^r(d_N(\hat{\beta}_\psi))$ around $d_N(\hat{\beta}_N) = d_0$ gives, after using repeatedly bounding arguments allowed by Assumptions A3, that

$$\widehat{E[b^r_{\psi,N}]} = E[b^r_{\psi,N}] + \xi^r_{\psi,N}/\sqrt{N} + O_p(N^{-1}),$$

where

$$\xi^r_{\psi,N} = \frac{\partial u^r(d_0)}{\partial d}\sqrt{N}\left[d_N(\beta_0) - E[d_N(\beta_0)] - \sum_{a=1}^{M+K+1} E\left(\frac{\partial d_N(\beta_0)}{\partial \beta^j}\right) i_N^j\right].$$

Note that $\xi^R_{\psi,N}$ depends on ψ only through $\partial\mu_{ab,N}(\beta_0)/\partial\beta$, that is, through $\partial Q^a(\beta_0)/\partial\beta^b\partial\beta^c$. Inspection of the null derivatives revels that $\partial Q^a(\beta_0)/\partial\beta^b\partial\beta^c$ only depends on ψ_3 for $a, b, c \in (1, \ldots, M)$ and the result follows. *Q.E.D.*

References

Agmon, N., Alhassid, Y., Levine, R., 1979. An algorithm for finding the distribution of maximal entropy. Journal of Computational Physics 30, 250–258.

Back, K., Brown, D., 1993. Implied Probabilities in GMM Estimators. Econometrica 61 (4), 971–975.

Bickel, P., 1998. Efficient and adaptive estimation for semiparametric models. Springer Verlag, Berlin, Germany.

Broniatowski, M., Keziou, A., 2004a. Estimation and tests for models satisfying linear constraints with unknown parameter. Prépublication de LSTA, Université Paris VI, Jussieu.

Broniatowski, M., Keziou, A., 2004b. Parametric estimation and tests through divergences. Prépublication de LSTA, Université Paris VI, Jussieu.

Brown, B., Newey, W., 1998. Efficient Semiparametric Estimation of Expectations. Econometrica 66 (2), 453–464.

Brown, B., Newey, W., 2002. Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference. Journal of Business & Economic Statistics 20 (4), 507–517.

Chen, S., Cui, H., 2007. On the second-order properties of empirical likelihood with moment restrictions. Journal of Econometrics 141 (2), 492–516.

Corcoran, S., 1998. Bartlett adjustment of empirical discrepancy statistics. Biometrika 85, 967–972.

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

Gallant, R. A., White, H., 1988. A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models. Basil Blackwell, New York.

Ghosh, J., Sinha, B., Wieand, H., 1980. Second Order Efficiency of the MLE with Respect to any Bounded Bowl-Shape Loss Function. The Annals of Statistics, 506–521.

Golan, A., 2002. Information and entropy econometrics-editor's view. Journal of Econometrics 107 (1), 1–16.

Golan, A., 2008. Information and Entropy Econometrics: A Review and Synthesis. Now Publisher, Hanover, MA, USA.

Golan, A., Judge, G., Miller, D., 1996. Maximum entropy econometrics: Robust estimation with limited data. John Wiley & Sons Inc, New York, USA.

Guggenberger, P., 2008. Finite Sample Evidence Suggesting a Heavy Tail Problem of the Generalized Empirical Likelihood Estimator. Econometric Reviews 27 (4-6), 526–541.

Guggenberger, P., Smith, R. J., July 2005. Generalized empirical likelihood estimators and tests under partial, weak, and strong identification. Econometric Theory 21 (04), 667–709.

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

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

Hansen, L. P., Heaton, J., Yaron, A., July 1996. Finite-sample properties of some alternative gmm estimators. Journal of Business and Economic Statistics 14 (3), 262–80.

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

Imbens, G. W., October 2002. Generalized method of moments and empirical likelihood. Journal of Business and Economic Statistics 20 (4), 493–506.

Imbens, G. W., Spady, R. H., Johnson, P., March 1998. Information theoretic approaches to inference in moment condition models. Econometrica 66 (2), 333–57.

Jaynes, E., 1984. Prior information and ambiguity in inverse problems. In: SIAM-AMS Proceedings. Vol. 14. pp. 151–66.

Jaynes, E., 2003. Probability theory: the logic of science. Cambridge Univ Press.

Kitamura, Y., 2001. Asymptotic optimality of empirical likelihood for testing moment restrictions. Econometrica 69, 1661–1672.

Kitamura, Y., 2007. Empirical Likelihood Methods in Econometrics: Theory and Practice. In: Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress. Cambridge University Press, p. 174.

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

Kitamura, Y., Tripathi, G., Ahn, H., 2004. Empirical Likelihood-Based Inference in Conditional Moment Restriction Models. Econometrica 72 (6), 1667–1714.

Komunjer, I., Ragusa, G., 2009. Existence and Uniqueness of Semiparametric Projections.

Kunitomo, N., Matsushita, Y., 2003. Finite sample distributions of the empirical likelihood estimator and the GMM estimator, university of Tokyo.

Maasoumi, E., 2007. 5 On Econometric Methodology. Economic Record 64 (4), 340–343.

McCullagh, P., 1987. Tensor Methods in Statistics. Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Mebane, Jr., W. R., Jr., Sekhon, J. S., 2009. Genetic optimization using derivatives: The rgenoud package for R. Journal of Statistical SoftwareForthcoming.
URL `http://www.jstatsoft.org/`

Michalewicz, Z., Swaminathan, S., Logan, T., 1993. GENOCOP (version 2.0). C language computer program source code. http://www. cs. adelaide. edu. au/˜ zbyszek/EvolSyst/genocop2. 0. tar. Z.

Mittelhammer, R., Judge, G., Schoenberg, R., 2005. Empirical Evidence Concerning the Finite Sample Performance of EL-Type Structural Equation Estimation and Inference Methods. Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg, Andrews and Stock (eds).

Newey, W. K., McFadden, D., 1994. Estimation and inference in large samples. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics. North-Holland, Amsterdam, pp. 2113–2245.

Newey, W. K., Smith, R. J., 2004. Higher order properties of GMM and Generalized Empirical Likelihood estimators. Econometrica 72 (1), 219–55.

Otsu, T., 2008. Conditional empirical likelihood estimation and inference for quantile regression models. Journal of Econometrics 142 (1), 508–538.

Owen, A., 2001. Empirical Likelihood. Chapman & Hall/CRC, Boca Raton, Florida.

Owen, A. B., 1990. Empirical likelihood ratio confidence regions. The Annals of Statistics 18, 90–120.

Pfanzagl, J., 1979. Nonparametric Minimum Contrast Estimators. Selecta Statistica Canadiana, 105–140.

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

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. Annals of Statistics 22, 300–325.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL `http://www.R-project.org`

Ramalho, J. J., Smith, R. J., 2005. Goodness of fit tests for moment condition models. Economics Working Papers 5-2005, available at http://ideas.repec.org/p/evo/wpecon/5_2005.html.

Rockafellar, T. R., 1970. Convex Analysis. Princeton University Press.

Rothenberg, T. J., 1984. Approximating the distributions of econometric estimators and test statistics. In: Griliches, Z., Intriligator, M. D. (Eds.), Handbook of econometrics. Vol. 2. North-Holland, Amsterdam, New York and Oxford, pp. 882–935.

Sargan, J., 1974. The Validity of Nagar's Expansion for the Moments of Econometric Estimators. Econometrica 42 (1), 169–176.

Schennach, S., 2007. Point Estimation with Exponentially Tilted Empirical Likelihood. The Annals of Statistics 35 (2), 634–672.

Smith, R., 1997. Alternative Semi-parametric Likelihood Approaches to Generalised Method of Moments Estimation. The Economic Journal 107 (441), 503–519.

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

Ullah, A., 2004. Finite sample econometrics. Oxford University Press, Oxford, UK.

Wald, A., 1949. Note on the Consistency of the Maximum Likelihood Estimate. The Annals of Mathematical Statistics 20 (4), 595–601.

Whang, Y., 2006. Smoothed empirical likelihood methods for quantile regression models. Econometric Theory 22 (02), 173–205.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica: Journal of the Econometric Society 50 (1), 1–25.

Wolfowitz, J., 1949. On Wald's Proof of the Consistency of the Maximum Likelihood Estimate. The Annals of Mathematical Statistics 20 (4), 601–602.