Minimum Divergence, Generalized Empirical Likelihoods, and Higher Order Expansions

Giuseppe Ragusa[*]

May 19, 2008

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 computationally tractable 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, 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.

[*] Department of Economics, University of California, Irvine, 3151 Social Science Plaza, 92697 Irvine, CA. email: gragusa@uci.edu

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 overidentified 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 overidentified 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 **?** 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 higher order bias, and it 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 one: 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 perfect equivalence between MD estimators and solutions to a saddle point problem has interesting implications. 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. Specifically, 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 efficiency properties of MD estimators. We show that MD estimators with the same higher order bias as EL also share the same higher order Mean Square Error (MSE). In light of the EL efficiency result in NS, this implies that there are many higher order efficient estimators in the MD class.

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 robustness to misspecification 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. 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 and behave well under misspecification.

A word on notation. If A is a matrix, $\|A\| = \sqrt{\operatorname{Tr} AA'}$ 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

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

where $\mathcal{F}_\gamma$ 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 $\mathrm{int}(D_\gamma) = (a_\gamma, b_\gamma)$, $a_\gamma < 1 < b_\gamma$; $\gamma(1) = 0$, $\gamma_1(1) = 0$, $\gamma_2(x) > 0$ 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 $\gamma \in \mathcal{F}_\gamma$ 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 $x \in \mathrm{int}(D_\rho)$. 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 $\bar{\rho} \in \mathcal{F}_\gamma$.

The MD problem in (2) defines a collection of estimators indexed by γ ranging in $\mathcal{F}_\gamma$. 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 $\gamma \in \mathcal{F}_\gamma$, the specific member of $\mathcal{F}_\gamma$ 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.

Corcoran (1998) analyzes problem (2) when the moment function does not depend on θ, with $q(w, \theta) = w$. In this case, the optimization takes place only over the weights $(\pi_1, \ldots, \pi_N)$.

[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 $\gamma^{cr}(\cdot, 2) : D_\gamma \mapsto \mathbb{R}$, $D_\gamma = [0, +\infty)$.

A precursor of the MD class estimators is the generalized minimum contrast class of estimators studied by Bickel (1998, Ch. 7) and Pfanzagl (1979).

In the exactly identified case, $M = K$, 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 $\pi_i = N^{-1}$ $(i = 1, \dots N)$. Thus, in this case, the MD estimator coincides with the Method of Moment estimator.

Problem (2) is feasible if the set $C(\theta) = \{y_i \in D_\gamma,\, i = 1 \dots, N : \sum_{i=1}^{N} y_i q_i(w_i, \theta) = 0\}$ 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 $\theta \in \Theta$.

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 $\pi_i(\ddot{\theta})$, $(i = 1, \dots, N)$, 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 $0 \le \zeta \le 1$. 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, $\pi_i(\dot{\theta}) = \pi_i(\ddot{\theta})$ $(i = 1, \dots, N)$.

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,

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

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) . Under (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 close form expression for some $\gamma \in \mathcal{F}_\gamma$. Even if $\tilde{\gamma}_1$ has a close 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] 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.

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

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

where $\Lambda^\dagger_N(\theta) = \left\{(\eta,\lambda') \in \mathbb{R}^{M+1} : \eta + \lambda' q_i(\theta) \in D_\psi, \text{ for all } i \leqslant N\right\}$. If $q(\cdot,\theta)$ is differentiable on Θ, a solution $\hat\theta \in int(\Theta)$ and $(\hat\eta, \hat\lambda') \in \Lambda^\dagger_N(\hat\theta)$ must satisfy the 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 $\widehat P_N := \sum_{i=1}^N \psi(\hat\eta + \hat\lambda' \hat q_i)/N - \hat\eta$.

Suppose $\hat\theta \in \Theta$ and $(\hat\eta, \hat\lambda') \in \Lambda_N(\hat\theta)$ solve (4) for some $\psi \in \mathcal{F}_\psi$. 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: $\gamma \in \mathcal{F}_\gamma$, $\psi_1(x) = \tilde\gamma_1(x)$ 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 2.2: for any divergence $\gamma \in \mathcal{F}_\gamma$, there exists a function $\psi \in \mathcal{F}_\psi$ 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).

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 $\gamma \in \mathcal{F}_\gamma$. 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: $\psi \in \mathcal{F}_\psi$, $\gamma_1(x) = \tilde\psi_1(x)$ for $x \in D_\gamma$, $\mathcal{A}_\gamma = D_\psi$, and $\widehat P_N = -\widehat\Gamma_N$.

Proof. See Appendix A.

Theorem 2.2 and Theorem 2.2 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 $\psi \in \mathcal{F}_\psi$; for any criterion function $\psi \in \mathcal{F}_\psi$, there exists an underlying MD problem whose solution is the same as the one to the saddle point problem.

If $q(\cdot, \theta)$ is differentiable on Θ, Theorems 2.2-2.2 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 2.2-2.2 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 $(\hat{\theta}', \hat{\eta}, \hat{\lambda}')$.

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 $\mathcal{F}_\gamma$, 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 2.2 is that it shows that any MD estimator can be defined from the "bottom-up" as the solution to (4) for a given $\psi \in \mathcal{F}_\psi$. 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.

[Table 1 about here.]

When the divergence implied by a given $\psi \in \mathcal{F}_\psi$ 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 $\psi \in \mathcal{F}_\psi$ with attractive statistical properties. We then compare them to the divergences of EL and ET.

Theorem 2.2 does not make any uniqueness claim about the solution. Uniqueness of $\hat{\theta}$ as a solution to (4) is not guaranteed because the function $\min_{(\eta,\lambda')\in\Lambda_N^\dagger(\theta)} \sum_{i=1}^N \psi(\eta + \lambda' q_i(\theta))/N - \eta$ is not necessarily (strictly) concave in θ. Theorem 2.2 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.

The GEL problem

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

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

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 $\gamma\in\mathcal{F}_\gamma$. First, we introduce the concept of generalized homogeneous functions.

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 $\widetilde P_N := \sum_{i=1}^N\psi(\tilde\tau'\tilde q_i)/N$.

Suppose $(\tilde\theta',\tilde\tau')$ solves (6) for some $\psi\in\mathcal{F}_\psi$. If $\tilde\psi_1$ is generalized homogeneous, then $\tilde\theta$ and $\tilde\omega_i$ $(i=1,\dots,N)$ solve (2) when $\gamma(x)=x\tilde\psi_1(x)-\psi(\tilde\psi_1(x))$. For this choice of γ it holds: $\gamma\in\mathcal{F}_\gamma$, $D_\psi=\mathcal{A}_\gamma$, $\psi_1(x)=\tilde\gamma_1(x)$, $x\in D_\psi$, and $\widetilde P_N=-\widetilde\Gamma_N^\dagger=\widehat P_N$.

Proof. See Appendix A □

If the inverse function of the first derivative of $\psi\in\mathcal{F}_\psi$ is generalized homogeneous, GEL estimators and MD estimators coincide for γ given in Theorem 2.2. Therefore, from Theorem 2.2, 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 $\hat\pi_i=\tilde\omega_i$ $(i=1,\dots,N)$.

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:

$$\min_{\theta\in\Theta,p_1,\dots,p_N}\sum_{i=1}^{N}\gamma(Np_i), \quad \gamma\in\mathcal{F}_\gamma$$

$$s.t.\ \sum_{i=1}^{N}p_i q_i(\theta)=0,\ \ Np_i\in(a_\gamma,b_\gamma),\ i=1,\dots,N. \tag{7}$$

Let $\tilde p_i=\psi_1(\tilde\tau'\tilde q_i)/N$.

[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 $\gamma^{cr}\in\mathcal{F}_\gamma$.

Suppose $(\tilde{\theta}', \tilde{\tau}')$ solves (6) for some $\psi \in \mathcal{F}_\psi$. Then $\tilde{\theta}$ and $\tilde{\pi}_i$ $(i = 1, \dots, N)$ solve (7) when $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$. For this choice of γ it holds: $\gamma \in \mathcal{F}_\gamma$, $\mathcal{A}_\gamma = D_\psi$, $\psi_1(x) = \tilde{\gamma}_1(x)$, $x \in D_\psi$, 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).

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) $E\big[\sup_{\theta\in\Theta} \|q(w, \theta)\|^2\big] < \infty$; (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) $\text{Rank}(G) = K$, $G = E\left[G_i(\theta_0)\right]$.

If A1 holds, $\hat{\theta} \xrightarrow{p} \theta_0$, $\hat{\eta} = O_p(N^{-1})$, 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.

[4] Since by Theorems 2.2 and 2.2 $\tilde{\psi}_1(x) = \gamma_1(x)$, $\tilde{\psi}_1$ does not have a closed form expression any time the corresponding divergence does not have a closed form expression.

If A1 and A2 hold,

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

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

Proof. See Appendix A □

The weights $\hat{\pi}_i$ $(i = 1, \dots, N)$ 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.

If A1 and A2 hold,

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

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 by Back and Brown (1993) and Brown and Newey (1998) in the GMM context. 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 $\tilde{\omega}_i = \tilde{\pi}_i / \sum_{i=1}^N \tilde{\pi}_i$ $(i = 1, \dots, N)$.

4 Testing overidentified 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 overidentication 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 2.2, 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 the EL and the ET problems. 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.[5] 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

[5] The Lagrange multipliers can be scaled by any consistent estimator of Ω, for instance by $\tilde{\Omega} = \sum_{i=1}^{N} q_i(\tilde{\theta})q_i(\tilde{\theta})'/N$ or $\hat{\Omega} = \sum_{i=1}^{N} q_i(\hat{\theta})q_i(\hat{\theta})'/N$ without affecting the validity of the asymptotic calibration. Imbens et al. (1998) also consider scaling the Lagrange multipliers by a robust weighting matrix given by $\hat{\Omega}_r = \hat{\Omega}_{\pi}\Big[\sum_{i=1}^{N} \hat{\pi}_i^2 q_i(\hat{\theta})q_i(\hat{\theta})'\Big]^{-1}\hat{\Omega}_{\pi}$.

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,

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

The above 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 $\psi_3(0) \neq 2$, the Lagrange multiplier $\hat{\eta}$ can be used to test the overidentified 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 $LM_\eta(\hat{\theta}) = -2\sum_{i=1}^N \psi(\hat{\lambda}' q_i(\hat{\theta}))$.

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 $2\sum_{i=1}^N \gamma(N\hat{\pi}_i)$ 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.

Any higher order asymptotic analysis begins with an expansion of $\sqrt{N}(\hat{\theta}-\theta_0)$. This expansion usually takes the following form

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

The terms appearing in (8) are tractable being sums and products of sample averages. Assumptions are necessary to guarantee that the remainder of the expansion, the last term in (8), is bounded in probability up to the required order. In our case, we will require that $r_N = O_p(N^{-3/2})$.

We can easily derive two properties of estimators that possess an expansion as in (8): the asymptotic bias of order N^{-1} and the asymptotic MSE of order N^{-2}. They are obtained by taking the expectation of the corresponding terms in the expansion (8).

If an estimator $\sqrt{N}(\hat{\theta}-\theta_0)$ admits an asymptotic expansion as in (8), its $O(N^{-1})$ bias is given by

$$E\big[i_N\big] + E\big[b_N\big]/N.$$

Often, $E\big[i_N\big]= 0$ and the asymptotic bias reduces to $E\big[b_N\big]/N$.

If an estimator $\sqrt{N}(\hat{\theta}-\theta_0)$ admits an expansion as in (8), the $O(N^{-1})$ MSE is given by

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

The asymptotic moments of Definition 2 and 3 are equivalent to those obtained by replacing the actual distribution of $\sqrt{N}(\hat{\theta}-\theta_0)$ with its $o(N^{-1})$ Edgeworth approximation, when the latter exists (see Rothenberg, 1984). Sargan (1974) shows that moments obtained from taking term-by-term expectations of the stochastic expansion (8) coincide with the moments of the finite sample distribution, when these moments are finite. As pointed out by Srinivasan (1970), it is possible that an estimator whose finite sample distribution does not have finite moments admits an asymptotic expansion. Kunitomo and Matsushita (2003) and Guggenberger (2004) suggest that EL estimators do not have finite moments in a linear simultaneous equations setting. These findings seem to question comparisons of MD estimators based on moments of terms in their asymptotic expansion. However, we interpret the moments based on (8) as the moments of an approximating distribution, and, as pointed out by Rothenberg (1984), with this interpretation it is not unreasonable to compare estimators in terms of higher order bias and MSE.

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 component-wise 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) $j, k, \ell, m, n, o \in \{2, \ldots, M + 1\}$, (iii) $r, s, t, u, v, w \in \{M + 2, \ldots, M + K + 1\}$. Let $\beta = (\eta, \lambda', \theta')'$ and define

$$\begin{aligned}
Q_{i,1}(\beta) &:= \psi_1(\eta + \lambda' q_i(\theta)) - 1 \\
Q_{i,2}(\beta) &:= \psi_1(\eta + \lambda' q_i(\theta)) q_i(\theta) \\
Q_{i,3}(\beta) &:= \psi_1(\eta + \lambda' q_i(\theta)) G_i(\theta)' \lambda.
\end{aligned}$$

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

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

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

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

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

and so forth.

The estimating equation of MD, $\sum_i Q_i(\hat{\beta})/N = 0$, is formally equivalent to the score equation of the MLE. We can then use the results in McCullagh (1987) and expand $\sum_i Q(\hat{\beta})/N = 0$

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 $O_p(N^{-3/2})$ expansion for $\hat{\delta}^a$ of the type in (8), one proceeds by telescopic substitution of lower order expansions to obtain

$$\hat{\delta}^a = i_N^a + b_N^a/\sqrt{N} + c_N^a/N + O_p(N^{-3/2}), \tag{9}$$

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

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

The bias of the MD estimator can be easily found. Here we give an expression for the bias in which the expectations of higher order derivatives of Q_i are substituted with expectations of higher order derivatives of q_i.

$$\begin{aligned} E(\hat{\delta}^r) = (1 - \psi_3(0)/2)\mu^{rj}\mu^{kl} E\left[q_i^j q_i^k q_i^\ell\right] / \sqrt{N} \\ + \mu^{rj}\Big\{\mu^{sk} E\big[(\partial q_i^j / \partial\beta^s) q_i^k\big] - \mu^{st} E\big[\partial q_i^j / \partial\beta^s \partial\beta^t\big]/2\Big\}/\sqrt{N}. \end{aligned} \tag{10}$$

Only the third derivative of ψ evaluated at 0 affects the magnitude of the higher order bias. When q_i has non zero generalized third moments, all MD estimators with $\psi_3(0) = 2$ have the same N^{-1} bias. EL has $\psi_3(0) = 2$.

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 $\hat{\delta}_{\psi}$ and $\hat{\delta}_{\psi'}$ denote MD estimators obtained from $\psi \in \mathcal{F}_{\psi}$ and $\psi' \in \mathcal{F}_{\psi}$ respectively.

If A1-A3 hold and $\psi_3(0) = \psi_3'(0)$, then $E(\hat{\delta}^r_{\psi}\hat{\delta}^s_{\psi}) - E(\hat{\delta}^r_{\psi'}\hat{\delta}^s_{\psi'}) = o(N^{-1})$.

Proof. See Appendix B □

The theorem states that two MD estimators obtained from two divergences such that $\psi_3(0) = \psi_3'(0)$ have the same higher order MSE. This result has an interesting implication. Adapting an argument of Pfanzagl and Wefelmeyer (1979), NS show that the bias corrected EL estimator is higher order efficient, having the lowest $O(N^{-2})$ MSE among all the bias corrected estimators based on the same moment conditions. Since two MD estimators with $\psi_3(0) = 2$ have the same $O(N^{-1})$ bias and the same higher order MSE, it follows that they also have the same higher order efficiency.

All the MD estimators obtained from an objective function with the property $\psi_3(0) = 2$ are higher order efficient.

This result has a substantive implication: higher order efficiency is an inadequate criterion for prescribing which MD estimator should be used in practice. If one aims at selecting an estimator among those have the same bias as the EL, then another criterion must supplement higher order efficiency. In the next section, we propose to use an estimator's behavior under misspecification as an additional criterion.

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, mispecification 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,

$$\inf_{G \in \mathcal{G}} \int \gamma(dG/dF)dF, \quad \gamma \in \mathcal{F}_{\gamma}$$

where

$$\mathcal{G} = \bigcup_{\theta \in \Theta} \mathcal{G}(\theta), \quad \mathcal{G}(\theta) = \left\{ G : \int q(w,\theta) \mathrm{d}G = 0, \ \int dG = 1, \quad G \ll F \right\}.$$

If the model is correctly specified, $F \in \mathcal{G}$ and $F = \inf_{G \in \mathcal{G}} \int \gamma(dG/dF)dF$. When the model is misspecified, $F \notin \mathcal{G}$ and $F^*_\gamma := \inf_{G \in \mathcal{G}} \int \gamma(dG/dF)dF$ can be interpreted as the pseudo-true probability measure. Likewise, the value θ^*_γ that corresponds to F^*_γ can be regarded as the pseudo-true parameter for the misspecified model. We index the solutions by γ to stress the fact that under misspecification the pseudo-true probability and parameter depend on the particular divergence that defines the MD problem.

Under misspecification, not only the solutions to the population problem depend on γ: also the behavior of the MD estimators does. Under MS, the equivalence of the MD problem (2) and the saddle point problem (4) may fail to hold for some $\gamma \in \mathcal{F}_\gamma$, rendering estimation of θ^*_γ and F^*_γ unfeasible.

The equivalence, in terms of solutions, between the MD problem and the saddle point problem is based entirely on Lagrangian type arguments: the equivalence holds if the optimal solutions to (2) can be expressed as a particular function of $M + K + 1$ parameters. As seen in Section 2, Lagrangian type arguments can be used if there exist $(\hat{\eta}, \hat{\lambda}') \in \mathbb{R}^{M+1}$ and $\hat{\theta} \in \Theta$ such that

$$\hat{\eta} + \hat{\lambda}' q_i(\hat{\theta}) \in \mathcal{A}_\gamma, \text{ for all } i \leqslant N, \tag{11}$$

the constraints are satisfied for $\hat{\pi}_i = \tilde{\gamma}_1(\hat{\eta} + \hat{\lambda}' q_i(\hat{\theta}))/N$, and $\sum_{i=1}^N \gamma(N\hat{\pi}_i)/N \leqslant \sum_{i=1}^N \gamma(N\bar{\pi}_i)/N$ for all feasible $\bar{\pi}_i$ $(i = 1, \ldots, N)$. Condition (11) is not binding for MD problems when $\mathcal{A}_\gamma = (-\infty, +\infty)$. When $\mathcal{A}_\gamma$ does not span all $\mathbb{R}$, however, the MD solution may not be characterized by Lagrangian arguments.

If the model is correctly specified, as $N \to \infty$, condition (11) will be satisfied w.p.a.1. Under misspecification, (11) may instead fail to hold even when $N \to \infty$. 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 2.2), condition (11) becomes

$$\max_{i \leq N} \tau' q_i(\theta) < 1. \tag{12}$$

We show now that under misspecification there does not exist a $\sqrt{N}-$consistent 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. $\sup_{w\in\mathcal{W}} v'q(w,\theta) = +\infty$ for all $\|v\| = 1$ 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 $\xi \in \mathbb{R}^M$, $\|\xi\| = 1$. But then, 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 (12), 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 MS, $\pi_i = N^{-1} + o_p(1)$ $(i = 1, \ldots, N)$ and $\dot{\theta}$ are not asymptotic solutions to the MD problem.

In an interesting paper, Schennach (2007) shows that calculating the EL estimator by solving

$$\arg\max_{\theta\in\Theta} \min_{\tau\in\Lambda^{\dagger}(\theta)} \sum_{i=1}^{N} \ln(1 - \tau'q_i(\theta))/N,$$

is not a $\sqrt{N}$ convergent procedure for the pseudo true value under MS.

As should be clear from the previous discussion, 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 whose underlying $\mathcal{A}_\gamma$ is the real line and that are higher order efficient. We proceed by first deriving functions $\psi \in \mathcal{F}_\psi$ with full domain, $D_\psi = \mathbb{R}$ and such that $\psi_3(0) = 2$. We then use Theorem 2.2 to derive the underlying divergences.

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

$$\psi(x) = \exp h(x) - xC_1 - C_2,$$

$$C_1 = \frac{h_1(0)}{h_1(0) + h_2(0)}, \quad C_1 = \frac{1}{h_1(0) + h_2(0)}$$

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 $\mathcal{F}_\psi$ 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.

[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 $\nu < 0$, $c_1 = h_1(\nu)/(h_1(\nu)^2/h_2(\nu) - h(\nu) + h(\nu))$, $c_2 = e^{c_1\nu}/(h(\nu) + h_1(\nu)^2/h_2(\nu) - h(\nu))$.

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

It is easy to verify that $\psi_3^{qt}(0) = \psi_3^{ht}(0) = 2$. 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 2.2 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}.

[Figure 1 about here.]

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 2.2. Notice that its first derivatives is

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

and the inverse of it is

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

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

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

[Figure 2 about here.]

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.

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.

Notice that $\psi^{ht}(x)$, $\psi^{qt}(x;\nu)$ and $\psi^{el}(x;\bar{\varepsilon})$ do not satisfy the generalized homogeneity conditions of Theorem 2.2 and, thus, the estimators obtained from solving the GEL problem with these functions do not correspond to minimum divergence estimators.

Numerical Example

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 model is given by

$$E\big[q(w,\theta_0)\big] = 0, \quad q(w_i,\theta) = \begin{bmatrix} w_i - \theta \\ (w_i-\theta)^2 - 1 \end{bmatrix}. \tag{13}$$

In each Monte Carlo replication, w is drawn from $w \sim N(0, 0.64)$. Under this distribution, the moment condition is misspecified. In each replications, we solve the saddle point problem in (4) with $\psi^{el}(x,\bar{\varepsilon})$, $\psi^{el}(x;\varepsilon_N)$, $\psi^{ht}(x)$, and $\psi^{qt}(x;\nu)$, with $\bar{\varepsilon} = 0.99$, $\varepsilon_N = 1 - N^{-1}$, and $\nu = -1.5$. We consider three sample sizes, $N = \{1000, 2500, 5000\}$ and we run 1000 replications for each sample size.

Figure 3 plots the sampling distributions of the four estimators considered. Each panel plots the sampling distribution of each estimator for the three sample sizes considered. The sampling distribution of $\hat{\theta}_{\varepsilon_N}^{el}$ (upper left panel) shows clear signs of non-normality; departures from normality are exacerbated as the sample size increases. The (empirical) sampling distributions of $\hat{\theta}^{el}_{\bar{\varepsilon}}$, $\hat{\theta}^{qt}_{\nu}$ and $\hat{\theta}^{ht}$ are in line with the sampling distribution of a $\sqrt{N}$ consistent estimator and no departures from normality can be detected.

[Figure 3 about here.]

7 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 overidentified test statistics 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 third 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 overidentified 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 overidentified 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. Finally, we note that in Monte Carlo simulations not reported here tests of overidentified 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

Suppose $\psi \in \mathcal{F}_\gamma$. Then the function $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$ belongs to $\mathcal{F}_\gamma$, 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 $\{y : \gamma_1(x) = y, x \in S\} = \mathrm{dom}\psi_1$. *Q.E.D.*

Suppose $\gamma \in \mathcal{F}_\gamma$. Then the function $\psi(x) = x\tilde{\gamma}_1(x) - \gamma(\tilde{\gamma}_1(x))$ belongs to $\mathcal{F}_\psi$, 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 $\{y : \gamma_1(x) = y, x \in S\} = \mathrm{dom}\psi_1$. *Q.E.D.*

Proof of Theorem 2.2

Apply Lemma A to obtain that, for $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$, $\gamma \in \mathcal{F}_\gamma$, $\psi_1(x) = \tilde{\gamma}_1(x)$, $x \in D_\psi$, $\Lambda_N(\theta) = \Lambda_N^\dagger(\theta)$ 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, $i = 1, \dots, N$. 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 $(i = 1, \dots, N)$, 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

$$\widehat{\Gamma}_N = -\widehat{P}_N. \tag{A.1}$$

Let $(\bar{\eta}, \bar{\lambda}')' = \arg\min_{(\eta,\lambda')' \in \Lambda_N^\dagger(\bar{\theta})} P_N(\eta, \lambda, \bar{\theta})$ and $\bar{\pi}_i = \psi_1(\bar{\eta} + \bar{\lambda}'q_i(\bar{\theta}))/N$, $(i = 1, \dots, N)$. 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 $(i = 1, \dots, N)$ 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

$$-\widehat{P}_N = \widehat{\Gamma}_N \leqslant \sum_{i=1}^N \gamma(N\bar{\pi}_i)/N = -P_N(\bar{\eta} + \bar{\lambda}'q_i(\bar{\theta})). \tag{A.2}$$

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

Proof of Theorem 2.2

Apply Lemma A to obtain that, for $\psi(x) = x\tilde{\gamma}_1(x) - \gamma(\tilde{\gamma}_1(x))$, $\gamma \in \mathcal{F}_\gamma$, $\psi_1(x) = \tilde{\gamma}_1(x)$, $x \in D_\psi$, $\Lambda_N(\theta) = \Lambda_N^\dagger(\theta)$ 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 $\hat{p}_i$, $(i = 1, \dots, N)$, 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 $(i = 1, \dots, N)$, 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

$$\widehat{P}_N = -\widehat{\Gamma}_N \geqslant -\sum_{i=1}^N \gamma(N\hat{p}_i)/N. \tag{A.3}$$

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 2.2

Apply Lemma A to obtain that, for $\gamma(x) = x\tilde\psi_1(x) - \psi(\tilde\psi_1(x))$, $\gamma \in \mathcal{F}_\gamma$, $\psi_1(x) = \tilde\gamma_1(x)$, $x \in D_\psi$, $\Lambda_N(\theta) = \Lambda_N^\dagger(\theta)$ for $\theta \in \Theta$. Let p_i, $(i = 1, \dots, N)$, 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 $(i = 1, \dots, N)$, 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 $(i = 1, \dots, N)$, $\sum_{i=1}^{N} \varsigma_i = 1$, $\sum_{i=1}^{N} \varsigma_i \tilde q_i = 0$, 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 $P_N(\tau^*, \theta) = \min_{\tau \in T_N^\dagger(\theta)} P_N(\tau, \theta)$, $\pi_i^* = \gamma_1(\tau^{*'} q_i(\theta))/N$, $\omega_i^{*'} = \pi_i^*/\sum_{i=1}^N \pi_i^*$, 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 2.2

Lemma A gives the three first three conclusions. Let ς_i, $(i = 1, \dots, N)$, 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 $(i = 1, \dots, N)$, 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 2.2 and Theorem 2.2 that $\tilde{\theta}$ is optimal. *Q.E.D.*

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.

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 A, $(\tilde\eta,\tilde\lambda') := \arg\min_{(\eta,\lambda')\in\Lambda_N^s} P_N(\eta,\lambda,\theta)$, $\tilde\upsilon_i = t\tilde\eta + t\tilde\lambda' q_i(\theta_0)$, 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

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

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

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

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 3

The proof is based on the ideas of 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 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

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

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

$$N^{1/2}Q_N(\theta) = \sum_{i=1}^{N} -u(\theta)'q_i(\theta)/N + \sum_{i=1}^{N} R_i(\theta,t)/N, \tag{A.4}$$

where

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

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.4) is uniformly $O_p(N^{-1/2})$ and, therefore,

$$N^{1/2} Q_N(\theta) = -\sum_{i=1}^{N} u(\theta)' q_i(\theta)/N + o_p(1), \quad \text{uniformly in } \Theta. \tag{A.5}$$

Therefore,

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

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

$$\sup_{\theta \in \Theta} \left\| -\sum_{i=1}^{N} u(\theta)' q_i(\theta)/N - E\big[-u(\theta)' q_i(\theta)\big] \right\| = o_p(1). \tag{A.6}$$

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.5) and (A.6) yield

$$P\left\{ \sup_{\theta \in \Theta \setminus B(\theta_0, \delta)} P_N(\eta(\theta), \lambda(\theta), \theta) > -N^{-1/2} h(\delta) \right\} < \delta/2. \tag{A.7}$$

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 A 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 $\hat{\theta} \in \Theta \backslash B(\theta_0, \delta)$, then

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

Therefore, eventually,

$$P\left\{ P_N(\eta(\theta_0), \lambda(\theta_0), \theta_0) < -N^{-1/2} h(\delta) \right\} < \delta/2. \tag{A.8}$$

Noting that

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

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

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

giving consistency of $\hat{\theta}$. *Q.E.D.*

Proof of Theorem 3

From Theorem 3 and Lemma 3 the first order conditions

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

are satisfied w.p.a.1. A mean value expansion of the first order conditions around $\theta = \theta_0$, $\eta = 0$ and $\lambda = 0$ gives

$$0 = \begin{pmatrix} 0 \\ \sqrt{N} q_n \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & \Omega & G' \\ 0 & G & 0 \end{pmatrix} \begin{pmatrix} \hat{\eta} \\ \hat{\lambda} \\ \hat{\theta} \end{pmatrix} + o_p(1).$$

Using the formula for the inverse of a block matrix yields

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

as desired.

Proof of Theorem 3

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

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

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 $\hat{\pi}_i$ $(i = 1, \dots, N)$ are the solutions to the MD problem that does not impose the constraint and optimizes over θ and π_i $(i = 1, \dots, N)$. Asymptotic normality and semiparametric efficiency follows from Theorem (3); 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 4

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 $r - M - 1$ element of θ. That is, $\nabla_r q_i^j = \partial q_i^j / \partial \beta^r = \partial q_i^j / \partial \theta^{r-M-1}$. 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:

$$\begin{aligned}
\frac{\partial Q_i^j}{\partial\beta^k\partial\beta^\ell} &= \psi_3 q_i^j q_i^k q_i^\ell, & \frac{\partial Q_i^j}{\partial\beta^k\partial\beta^r} &= q_i^k \nabla_r q_i^j + q_i^j \nabla_r q_i^k, & \frac{\partial Q_i^j}{\partial\beta^r\partial\beta^s} &= \nabla_{r,s} q_i^j, \\
\frac{\partial Q_i^r}{\partial\beta^k\partial\beta^\ell} &= q_i^k \nabla_r q_i^\ell + q_i^\ell \nabla_r q_i^k, & \frac{\partial Q_i^r}{\partial\beta^k\partial\beta^s} &= \nabla_{r,s} q_i^k, & \frac{\partial Q_i^r}{\partial\beta^s\partial\beta^t} &= 0.
\end{aligned}$$

The partial third null derivatives are:

$$\begin{aligned}
\frac{\partial Q_i^j}{\partial\beta^k\partial\beta^\ell\partial\beta^m} &= \psi_4 q_i^j q_i^k q_i^\ell q_i^m, \\
\frac{\partial Q_i^j}{\partial\beta^k\partial\beta^\ell\partial\beta^r} &= \psi_3 \left(q_i^k q_i^\ell \nabla_r q_i^j + q_i^j q_i^\ell \nabla_r q_i^k + q_i^j q_i^k \nabla_r q_i^\ell \right), \\
\frac{\partial Q_i^j}{\partial\beta^k\partial\beta^r\partial\beta^s} &= \nabla_s q_i^k \nabla_r q_i^j + \nabla_s q_i^j \nabla_r q_i^k + q_i^j \nabla_{r,s} q_i^k + q_i^k \nabla_{r,s} q_i^j, \\
\frac{\partial Q_i^j}{\partial\beta^r\partial\beta^s\partial\beta^t} &= \nabla_{r,s,t} q_i^j, \\
\frac{\partial Q_i^r}{\partial\beta^k\partial\beta^\ell\partial\beta^m} &= \psi_3 \left(q_i^\ell q_i^m \nabla_r q^k + q_i^k q_i^m \nabla_r q^\ell + q_i^k q_i^\ell \nabla_r q^m \right), \\
\frac{\partial Q_i^r}{\partial\beta^k\partial\beta^\ell\partial\beta^s} &= \psi_3 \left(\nabla_s q_i^k \nabla_r q_i^\ell + \nabla_s q_i^\ell \nabla_r q_i^t + q_i^k \nabla_{r,s} q_i^\ell + q_i^\ell \nabla_{r,s} q_i^k \right), \\
\frac{\partial Q_i^r}{\partial\beta^k\partial\beta^s\partial\beta^t} &= \nabla_{r,s,t} q_i^k, \\
\frac{\partial Q_i^r}{\partial\beta^s\partial\beta^t\partial\beta^u} &= 0.
\end{aligned}$$

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 so forth. Since μ^{jk}, μ^{jr}, and μ^{rs} represent the (j,k) elements, the (j,r) elements, and the (r,s) elements of the inverse of Jacobian of the moment conditions, respectively, the following identities hold:

$$\begin{aligned}
&\mu^{js}\mu^{km}\kappa_{j,k} = 0 \\
&\mu^{jk}\mu^{\ell r}\kappa_{j,\ell} = 0 \\
&\mu^{jk}\mu^{\ell m}\kappa_{j,\ell} = \mu^{km} \\
&\mu^{jr}\mu^{ks}\kappa_{j,k} = -\mu^{rs}.
\end{aligned} \tag{A.9}$$

The above identities, which are central in deriving the results of this appendix, also hold for permutations of the indexes in lieu of the symmetry of the inverse of Jacobian of the moment conditions.

Derivation of Equation (10)

The $O_p(N^{-1})$ expansion for $\hat{\delta}^r$, $r \in \{M+2, M+K+1\}$ given in equation (9) reduces to

$$\hat{\delta}^r = -\mu^{rj} Z_j + T^r/\sqrt{N} + O_p(N^{-1}),$$

where

$$\begin{aligned} T^r &= \mu^{rj}\mu^{k\ell} Z_{jk} Z_\ell + \mu^{rj}\mu^{sk} Z_{js} Z_k + \mu^{rs}\mu^{k\ell} Z_{sk} Z_\ell \\ &\qquad - (\mu^{rj}\mu^{k\ell} Z_j Z_k Z_\ell + \mu^{rj}\mu^{st} Z_j Z_s Z_t + \mu^{rs}\mu^{jk} Z_r Z_j Z_k)/2. \end{aligned}$$

The asymptotic bias of $\sqrt{N}(\hat{\theta} - \theta_0)$ is given, after simplifications implied by (A.9) and by the form of the null partial derivatives, by

$$\begin{aligned} E(\hat{\delta}^r) &= \mu^{rj}\mu^{k\ell}\kappa_{jk,\ell} + \mu^{rj}\mu^{sk}\kappa_{js,k} + \mu^{rs}\mu^{k\ell}\kappa_{sk,\ell} \\ &\qquad - (\mu^{rj}\mu^{k\ell}\mu_{jk\ell} + \mu^{rj}\mu^{st}\mu_{jst} + \mu^{rs}\mu^{jk}\mu_{rjk})/2 + O(N^{-1}). \end{aligned}$$

The expressions for the expected values of combination of the $Z's$ appearing in the previous expression are given by

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

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

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

giving, thus, the desired result.

Q.E.D.

Proof of Theorem 5

Using the index convention, the MSE of the MDE/GEL estimator obtained from the objective function ψ is given by

$$\begin{aligned}E(\hat{\delta}^r_\psi \hat{\delta}^s_\psi) = E(i^r_{\psi,N} i^s_{\psi,N}) + E\big[(b^r_{\psi,N}/\sqrt{N} + c^r_{\psi,N}/N) i^s_{\psi,N}\big] \\ + E\big[i^r_{\psi,N}(b^s_{\psi,N}/\sqrt{N} + c^s_{\psi,N}/N)'\big] + o(N^{-1}).\end{aligned}$$

The difference between the MSE of the MD/GEL estimator obtained from ψ and the MSE of the MD/GEL estimator obtained from ψ' is thus given by the difference of the corresponding terms in the relative expansions. Since $\hat{\delta}_\psi$ and $\hat{\delta}_{\psi'}$ are first order equivalent, the difference in MSE reduces to

$$\begin{aligned}&E(\hat{\delta}^r_\psi \hat{\delta}^s_\psi) - E(\hat{\delta}^r_{\psi'} \hat{\delta}^s_{\psi'}) \\ &\quad = E\big[(b^r_{\psi,N}/\sqrt{N} + c^r_{\psi,N}/N) i^s_{\psi,N} - (b^r_{\psi',N}/\sqrt{N} + c^r_{\psi',N}/N) i^s_{\psi',N}\big] \\ &\qquad + E\big[i^r_{\psi,N}(b^s_{\psi,N}/\sqrt{N} + c^s_{\psi,N}/N) - i^r_{\psi',N}(b^s_{\psi',N}/\sqrt{N} + c^s_{\psi',N}/N)\big] + o(N^{-1}).\end{aligned}$$

Now we inspect the terms involved in the previous expression to conclude that if $\psi_3 = \psi'_3$ the only terms that differ in the expansion of the MSE of $\hat{\delta}^r_\psi$ and $\hat{\delta}^r_{\psi'}$ are the expectations of the product of the score and the $O_p(N^{-1})$ term $c_{\psi,N}/N$. Note that,

$$E(i^r_{\psi,N} b^s_{\psi,N}) = -\mu^{rj}\mu^{sa}\mu^{bk}\kappa_{ab,j,k} + \mu^{rj}\mu^{s,k,\ell}\kappa_{j,k,\ell}/2.$$

Since $\mu_{ab,j,k} = E(q^j_i q^k_i \partial Q^a_i/\partial\beta^b)$, $\mu_{jk\ell} = \psi_3 E(q^j_i q^k_i q^\ell_i)$, and $\psi_3 = \psi'_3$ we have that $E(i^r_{\psi,N} b^s_{\psi,N}/\sqrt{N}) - E(i^r_{\psi',N} b^s_{\psi',N}/\sqrt{N}) = 0$. Thus,

$$\begin{aligned}E(\hat{\delta}^r_\psi \hat{\delta}^s_\psi) - E(\hat{\delta}^r_{\psi'} \hat{\delta}^s_{\psi'}) = E\big[(c^r_{\psi,N}/N) i^s_{\psi,N} - (c^r_{\psi',N}/N) i^s_{\psi',N}\big] \\ + E\big[i^r_{\psi,N}(c^s_{\psi,N}/N) - i^r_{\psi',N}(c^s_{\psi',N}/N)\big] + O(N^{-2}).\end{aligned}$$

Further,

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

The only terms that enters $E(i^r_{\psi,N} c^s_{\psi,N})$ that does depend on ψ_4 is $\mu^{rj}\mu^{s,k,\ell,m}\mu_{j,k,\ell,m}/6$. This last term is in turn equal to

$$\mu^{rj}\mu^{sa}\mu^{kb}\mu^{\ell c}\mu^{md}\mu_{abcd}\kappa_{j,k,\ell,m},$$

and will depend on ψ_4 only for $a, b, c, d \in (2, \ldots, M+1)$. Thus,

$$E(\hat{\delta}^r_\psi \hat{\delta}^s_\psi) - E(\hat{\delta}^r_{\psi'} \hat{\delta}^s_{\psi'}) = (\mu^{rj}\mu^{sn} + \mu^{sj}\mu^{rn})\mu^{ko}\mu^{\ell p}\mu^{ml}\mu_{nopl}\kappa_{j,k,\ell,m}/N,$$

where

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

which is equivalent to

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

where, for example,

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

Here, the notation [3] denotes the sum over the three partitions of four indexes. Since $\kappa_j = 0$ and, by assumption, $E(q_i^j q_i^k q_i^\ell q_i^m) = O(1)$, the expression for $\kappa_{j,k,\ell,m}$ simplify to

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

The difference in the (r, s) element of the MSE of the two estimators reduces to

$$E(\hat{\delta}^r_\psi \hat{\delta}^s_\psi) - E(\hat{\delta}^r_{\psi'} \hat{\delta}^s_{\psi'}) = (\mu^{rj}\mu^{sn} + \mu^{sj}\mu^{rn})\mu^{ko}\mu^{\ell p}\mu^{ml}\mu_{nopl}\kappa_{j,k}\kappa_{\ell,m}[3]/N + O(N^{-2}).$$

Applying the identities in (A.9) yields

$$(\mu^{rj}\mu^{sn} + \mu^{sj}\mu^{rn})\mu^{ko}\mu^{\ell p}\mu^{ml}\mu_{nopl}\kappa_{j,k}\kappa_{\ell,m}[3]/N = 0,$$

giving, thus, the desired result. *Q.E.D.*

References

Back, K. and D. Brown (1993): "Implied Probabilities in GMM Estimators," *Econometrica*, 61, 971–975.

Bickel, P. (1998): *Efficient and Adaptive Estimation for Semiparametric Models*, Springer.

Brown, B. and W. Newey (1998): "Efficient Semiparametric Estimation of Expectations," *Econometrica*, 66, 453–464.

——— (2002): "Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference," *Journal of Business & Economic Statistics*, 20, 507–517.

Chen, S. and H. Cui (2007): "On the second order properties of empirical likelihood for generalized estimation equations," *Journal of Econometrics, to appear.*

Corcoran, S. (1998): "Bartlett Adjustment of Empirical Discrepancy Statistics," *Biometrika*, 85, 967–972.

Cressie, N. and T. Read (1984): "Multinomial Goodness-of-Fit Tests," *J. R. Statist. Soc. B*, 46, 440–464.

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

Guggenberger, P. (2004): "Finite-Sample Evidence Suggesting a heavy Tail Problem of the Generalized Empirical Likelihhood Estimator," *Unpublished Manuscript, Department of Economics, UCLA.*

Guggenberger, P. and R. J. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests under Partial, Weak, and Strong Identification," *Econometric Theory*, 21, 667–709.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–54.

Hansen, L. P., J. Heaton, and A. Yaron (1996): "Finite-Sample Properties of Some Alternative Gmm Estimators," *Journal of Business and Economic Statistics*, 14, 262–80.

Imbens, G. W. (1997): "One-Step Estimators for over-Identified Generalized Method of Moments Models," *Review of Economic Studies*, 64, 359–83.

——— (2002): "Generalized Method of Moments and Empirical Likelihood," *Journal of Business and Economic Statistics*, 20, 493–506.

Imbens, G. W., R. H. Spady, and P. Johnson (1998): “Information Theoretic Approaches to Inference in Moment Condition Models,” *Econometrica*, 66, 333–57.

Kitamura, Y. (2001): “Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions,” *Econometrica*, 69, 1661–1672.

Kitamura, Y. and M. Stutzer (1997): “An Information-Theoretic Alternative to Generalized Method of Moments Estimation,” *Econometrica*, 65, 861–74.

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

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

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

Mittelhammer, R., G. Judge, and R. Schoenberg (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. and D. McFadden (1994): “Estimation and Inference in Large Samples,” in *Handbook of Econometrics*, ed. by R. Engle and D. McFadden, Amsterdam: North-Holland, 2113–2245.

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

Owen, A. (2001): *Empirical Likelihood*, Chapman & Hall/CRC.

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.

Pfanzagl, J. and W. Wefelmeyer (1979): “A Third-Order Optimum Property of the Maximum Likelihood Estimator,” *Journal of Multivariate Analysis*, 8, 1–29.

Qin, J. and J. Lawless (1994): “Empirical Likelihood and General Estimating Equations,” *Annals of Statistics*, 22, 300–325.

Ramalho, J. J. and R. J. Smith (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 *Handbook of econometrics*, ed. by Z. Griliches and M. D. Intriligator, Amsterdam, New York and Oxford: North-Holland, vol. 2, 882–935.

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

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

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

Srinivasan, T. N. (1970): "Approximations to Finite Sample Moments of Estimators whose Exact Sampling Distributions Are Unknown," *Econometrica*, 38, 533–41.

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

Whang, Y. (2006): "Smoothed Empirical Likelihood Methods for Quantile Regression Models," *Econometric Theory*, 22, 173–205.

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

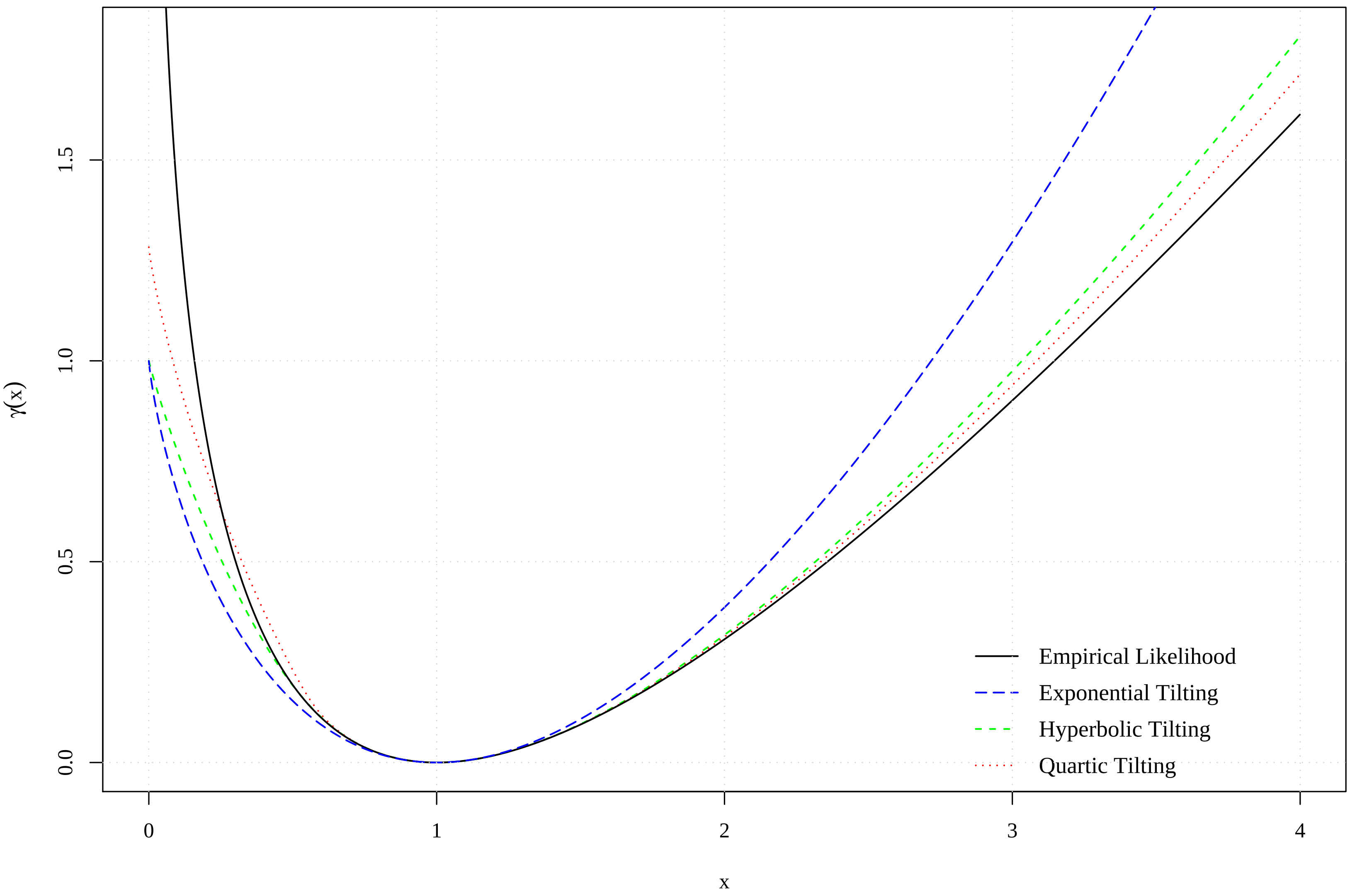

Figure 1: Implied divergence functions. For Quartic Tilting and Hyperbolic Tilting the divergences are obtained by numerically inverting the first derivative of ψ^{qt} and ψ^{ht}on a grid of points covering $(0,4)$ to obtain $\tilde{\psi}_1^{qt}$ and $\tilde{\psi}_1^{ht}$ and then by calculating $\gamma^{qt}(x)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

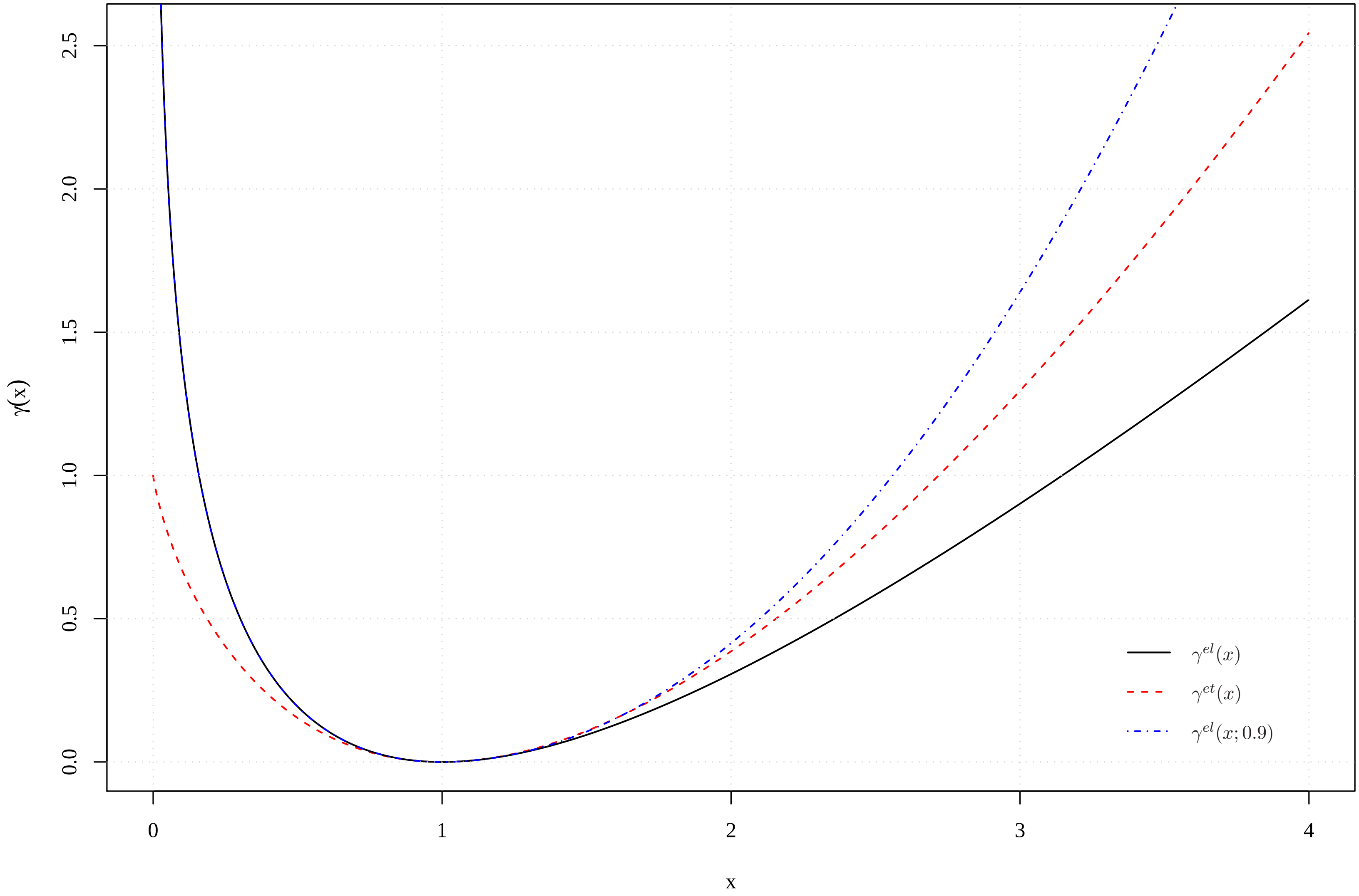

Figure 2: Empirical Likelihood, Exponential Tilting and Modified Empirical Likelihood divergences.

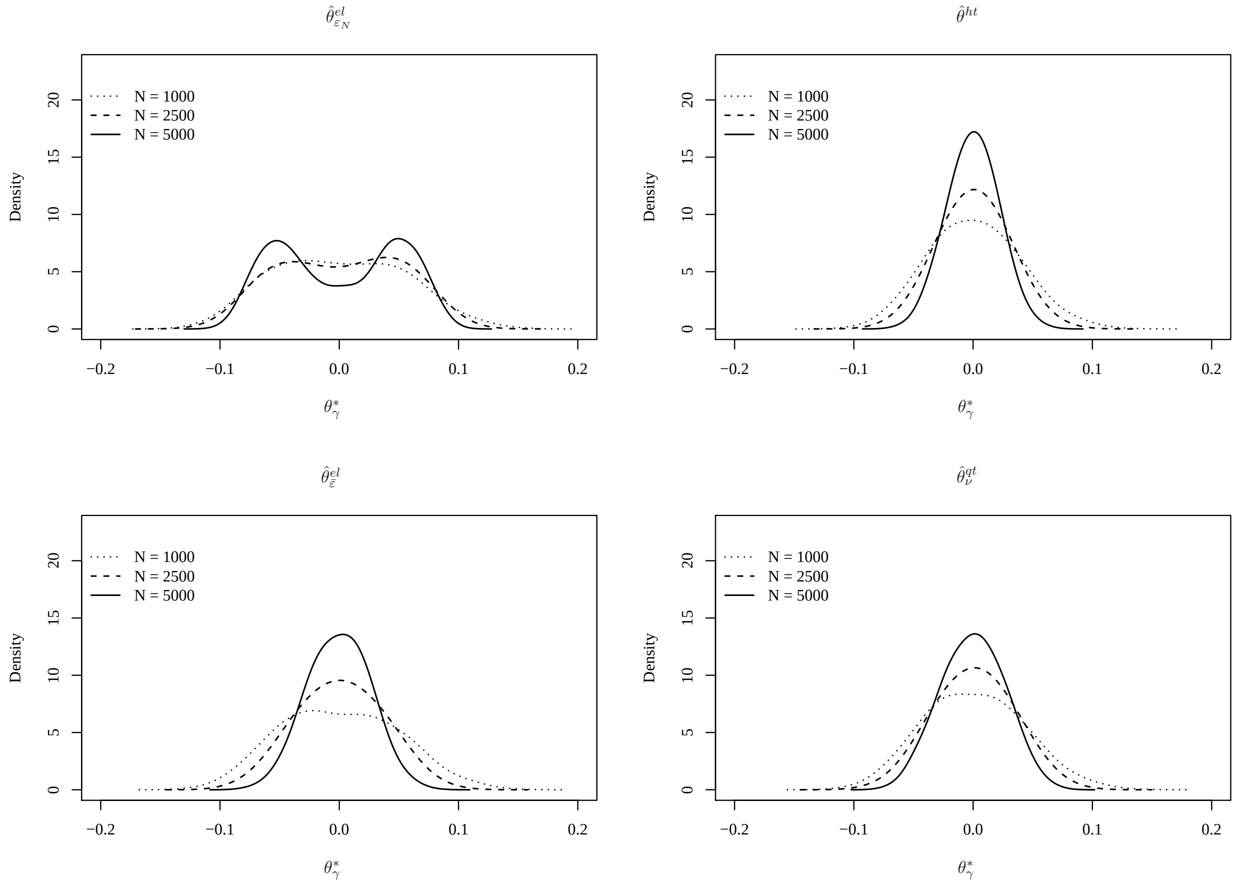

Figure 3: Empirical sampling distributions of the estimators based on $\gamma^{el}(x, \varepsilon_N)$, $\gamma^{el}(x, \bar{\varepsilon})$, $\gamma^{ht}(x)$ and $\gamma^{qt}(x; \nu)$, $\varepsilon = 0.99$ and $\varepsilon_N = 1 - N^{-1}$.

Name	γ	γ_1	$\tilde{\gamma}_1$	ψ	ψ_1	$\tilde{\psi}_1$	$\mathcal{A}$	$\psi_3(1)$
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	$\begin{cases} h(x) & x > x_0 \\ \frac{e^{c_1 x}}{c_2} - c_3 & x \leqslant x_0 \end{cases}$ †	$\begin{cases} h_1(x) & x > x_0 \\ \frac{c_1}{c_2} e^{c_1 x} & x \leqslant x_0 \end{cases}$ †	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 $\mathcal{A} = (-\infty, +\infty)$.

† $h(x) = e^{((1+x)^4 - 4x - 1)/12} + x - 1$, $x_0 < 0$, $c_1 = h_1(x_0)/(c_3 + h(x_0))$, $c_2 = e^{c_1 x_0}/(h(x_0) + c_3)$, and $c_3 = h_1(x_0)^2/h_2(x_0) - h(x_0)$.