

# Point Estimation with Exponentially Tilted Empirical Likelihood

S. M. Schennach\*  
University of Chicago

## Abstract

Parameters defined via General Estimating Equations (GEE) can be estimated by maximizing the Empirical Likelihood (EL). Newey and Smith (2004) have recently shown that this EL estimator exhibits desirable higher-order asymptotic properties, namely, that its  $O(n^{-1})$  bias is small and that bias-corrected EL is higher-order efficient. Although EL possesses these properties when the model is correctly specified, this paper shows that, in the presence of model misspecification, EL may cease to be root  $n$  convergent when the functions defining the moment conditions are unbounded (even when their expectations are bounded). In contrast, the related Exponential Tilting (ET) estimator avoids this problem. This paper shows that the ET and EL estimators can be naturally combined to yield an estimator called Exponentially Tilted Empirical Likelihood (ETEL) exhibiting the same  $O(n^{-1})$  bias and the same  $O(n^{-2})$  variance as EL, while maintaining root  $n$  convergence under model misspecification.

## 1 Introduction

Statistical models defined via General Estimating Equations (GEE) of the form  $E[g(x, \theta)] = 0$ , where  $g(x, \theta)$  is a vector-valued nonlinear function of a random

---

\*AMS 2000 subject classifications. Primary-62F10; secondary-62F12.  
Keywords: entropy, higher-order asymptotics, misspecified models.  
Supported by NSF grant SES-0214068.

vector  $x$  and of a parameter vector  $\theta$ , are very common in statistics. In such models, the parameter vector  $\theta$  is traditionally estimated using two-step efficient Generalized Method of Moments estimators (GMM) [21]. Over the last two decades, various one-step alternatives to two-step GMM have been suggested. Perhaps the best-known estimators of this class are the Empirical Likelihood (EL), Exponential Tilting (ET) and GMM with Continuous Updating (CU) estimators, which have been previously studied in the econometrics [22, 26, 27, 35, 47] and statistics [37, 45, 48–50, 53] literatures. While all of these alternative estimators of  $\theta$  share the first-order efficiency of efficient two-step GMM, their one-step nature provides them with desirable properties not enjoyed by GMM. In addition to bypassing the arbitrariness in the choice of first-step estimate (since any consistent estimate of  $\theta$  can, in principle, be used as a first step and lead to slightly different second-step estimates in finite samples), these one-step estimators are also invariant under general parameter-dependent linear transformations of the vector of moment conditions [30, 50] and possess superior higher-order asymptotic properties [27–29, 47].

A considerable effort has been devoted to identify which of these alternative estimators, EL, ET or CU, is preferable. Since all of these estimators are asymptotically equivalent up to  $O_p(n^{-1/2})$  when the overidentifying restrictions are valid, differences must reside in their higher-order asymptotic properties or in their behavior under potential model misspecification. The CU estimator is generally regarded as less desirable than EL and ET because its objective function has often been observed to possess multiple modes [22, 30] and because it lacks the ability to generate likelihood ratio-based confidence regions whose shape adapts to the support of the data [4, 50]. Comparing ET and EL proves to be more difficult. On the one hand, based on a stochastic expansion argument, Newey and Smith [47] have established

that EL should typically have a lower finite-sample bias relative to both ET and CU. Also, they have shown that bias-corrected EL is higher-order efficient relative to any other regular method of moment estimator. On the other hand, Imbens and co-workers [27, 30] have indicated that EL, unlike ET, exhibits a singularity in its influence function, suggesting that ET should be better behaved than EL in the presence of model misspecification. In addition, ET admits a computationally convenient treatment of misspecified models [32].

Although it can be argued that model misspecification can always be avoided through the use of specification tests, an alternative view is that most models are only approximations to the underlying phenomena and are therefore intrinsically misspecified. Accordingly, there exists a growing literature devoted to the study of misspecified models. The classic theory of maximum likelihood estimators (MLE) when the distributional assumptions are misspecified can be found in [1, 25, 63, 64]. In this context, MLE consistently estimates the so-called *pseudo-true value* of the parameter of interest [56] which is defined as the parameter value associated with the distribution which is the closest to the true data generating process according to the so-called Kullback-Leibler Information Criterion (KLIC) discrepancy.

In recent years, the analysis of misspecified models has been actively extended to various extremum estimators [2, 13, 44, 51] and, in particular, to overidentified moment condition models [8, 18, 26, 32, 34, 41]. Overidentified models arise naturally in a number of applications. For instance, consider a regression model  $y = x'\theta + \varepsilon$  where  $\varepsilon$  is correlated with  $x$  (so that least squares cannot be used) but uncorrelated with a vector of so-called instruments (denoted  $z$ ). This leads to a vector of restrictions of the form  $E[(y - x'\theta)z] = 0$ , the dimension of which typically exceeds the dimension of  $\theta$ . Given the overidentified (i.e. overdetermined) nature of the restric-

tions, it is then possible that no value of  $\theta$  simultaneously satisfies all the moment restrictions exactly in the population, resulting in a misspecified model [41]. A more extensive discussion of misspecified models as well as many references to empirical studies that perform inference with models which fail standard specification tests can be found in [18].

The motivation behind this interest for misspecified models stems from two observations. First, the imperfections of a model, although statistically detectable, may nevertheless be small in absolute terms and have consequently little impact on the results [42, p. 1168–1169]. Second, a misspecified but parsimoniously parametrized model may have better predictive power than a more realistic complex model which passes all specification tests [9, p. 596]. At fixed sample size, as the number of parameters increases, their variances tend to increase as well, while the power of overidentification tests tends to decrease.

This paper is organized as follows. After briefly reviewing the properties of the EL, ET and CU estimators, we present a simple result that characterizes EL's poor behavior under misspecification in order to motivate the need for a new estimator. We then introduce an estimator called Exponentially Tilted Empirical Likelihood (ETEL) that naturally combines EL and ET, extending an approach previously considered in [10, 31, 40] for constructing likelihood-ratio confidence regions for the mean to the case of point estimation of parameters defined via general moment restrictions. The ETEL estimator is shown to be well-behaved under model misspecification, like ET, while preserving the desirable higher-order asymptotic properties of EL established in [47]. Finally, simulations are used to illustrate the usefulness of this estimator. All proofs can be found in the Appendix.

## 2 Existing one-step alternatives to GMM

### 2.1 Generalities

We first introduce our notation.

**Definition 1** Let  $\theta$  denote the parameter vector of interest belonging to a compact subset  $\Theta$  of  $\mathbb{R}^{N_\theta}$ . Let  $x_i$  be sequence of random vectors taking values in  $\mathcal{X} \subset \mathbb{R}^{N_x}$ . Let  $g(x_i, \theta)$  denote a vector of moment function taking value in  $\mathbb{R}^{N_g}$  and satisfying  $E[g(x_i, \theta^*)] = 0$  at  $\theta^* \in \Theta$ . Let  $n$  denote sample size and let all summations be over  $1, \dots, n$ . Let  $\|\cdot\|$  denote any convenient vector or matrix norm.

The simplest way to summarize the properties of the EL, ET and CU estimators is to embed them in more general families of estimators. All three estimators admit two convenient representations. They can first be interpreted as minimum empirical discrepancy (MED) estimators [10, 11]:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \left( n^{-1} \sum_i h(\hat{w}_i(\theta)) \right), \quad (1)$$

where  $\hat{w}_i(\theta)$  is the solution to

$$\min_{\{w_i\}_{i=1}^n} n^{-1} \sum_i h(w_i) \quad (2)$$

subject to moment and normalization constraints:

$$\sum_i w_i g(x_i, \theta) = 0 \text{ and } \sum_i w_i = 1. \quad (3)$$

(The term *empirical discrepancy* is used here to emphasize the fact that it is a discrepancy between measures supported on the sample rather than on a fixed discrete support.) Different choices of the discrepancy measure  $h(\cdot)$  yield distinct estimators, as given in Table 1. Specific choices of  $h(\cdot)$  have historically been given special

names. The discrepancy used in EL,  $h(w) = -\ln nw$ , is known as the Kullback-Leibler Information Criterion (KLIC). Also, rewriting the minimization problem into an equivalent maximization problem, EL can be thought of as maximizing the “likelihood”. In a similar fashion, ET, with  $h(w) = nw \ln(nw)$ , can be interpreted as maximizing a quantity known as *entropy*.

The minimum discrepancy formulation emphasizes that the estimator seeks to “reweight” the sample in order to satisfy the moment conditions exactly. The function  $h(w_i)$  quantifies the amount of reweighting taking place and penalizes values that differ from  $w_i = n^{-1}$ . The point estimate  $\hat{\theta}$  is the value that minimizes the discrepancy between  $\hat{w}_i(\theta)$  and uniform weights. The weights  $\hat{w}_i(\hat{\theta})$  are sometimes called *implied probabilities* because they can be used to construct more efficient empirical estimates of the data generating process [3, 26, 53].

EL, ET and CU can also be characterized as particular cases of the so-called Generalized Empirical Likelihood (GEL) family of estimators [61]:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \left( n^{-1} \sum_i \rho \left( \hat{\lambda}(\theta)' g(x_i, \theta) \right) \right), \quad (4)$$

where the  $N_g$  dimensional vector  $\hat{\lambda}(\theta)$  is given by

$$\hat{\lambda}(\theta) = \arg \max_{\lambda} \left( n^{-1} \sum_i \rho \left( \lambda' g(x_i, \theta) \right) \right). \quad (5)$$

The choice of the function  $\rho(\cdot)$  defines the estimator used, as described in Table 1. The advantage of the GEL formulation is the computational convenience of solving a  $N_g + N_\theta$  dimensional optimization problem rather than the  $n + N_\theta$  dimensional problem defining an MED estimator.

As pointed out in [47], only specific choices of the  $h(\cdot)$  lead to an estimator admitting an equivalent GEL representation. A particularly rich class of such dis-

crepancies are the Cressie-Read (CR) discrepancies [11],

$$h(w_i) = \frac{(nw_i)^{\gamma+1} - 1}{\gamma(\gamma+1)} \quad (6)$$

where  $\gamma$  is the parameter indexing the family. The corresponding  $\rho(\cdot)$  is given in Table 1. The GEL representation of an MED estimator is called a dual problem because it amounts to reformulating the optimization in terms of the Lagrange multiplier  $\lambda$  of the moment constraints. Newey and Smith [47] conjecture that Cressie-Read discrepancies may be the *only* discrepancies admitting a GEL representation. The weights attributed to the sample points in the original MED estimator can be recovered from

$$\hat{w}_i(\theta) = \frac{\tau(\hat{\lambda}(\theta)'g(x_i, \theta))}{\sum_j \tau(\hat{\lambda}(\theta)'g(x_j, \theta))}, \quad (7)$$

where  $\tau(\xi) = d\rho(\xi)/d\xi$ . We will refer to  $\tau(\xi)$  as the tilting function because, as seen in Equation (7), it indicates how the sample points are reweighted. The EL, ET and CU estimators are all members of this class (see Table 1) of Empirical Cressie-Read (ECR) estimators. For a more detailed description of these families of estimators, we refer the reader to the excellent discussions found in [47, 50].

## 2.2 Comparing the ECR estimators

Let us first give the properties shared by all ECR estimators. For just-identified models ( $N_g = N_\theta$ ), all of these estimators are trivially identical because the moment conditions can be satisfied exactly simply by choosing  $\hat{\theta}$  appropriately without the need for tilting ( $\hat{\lambda}(\hat{\theta}) = 0$ ). In over-identified models for which the over-identifying restrictions are valid, all ECR (and GEL) estimators possess the same asymptotic variance [47], which is equal to the asymptotic variance of the two-step efficient GMM estimator. All ECR estimators also enable the construction of confidence

regions for the mean ( $g(x_i, \theta) = x_i - \theta$ ) through convenient  $\chi^2$ -calibrated likelihood-ratio tests [4]. In light of the results in [53], Baggerly’s results should extend to general  $g(x_i, \theta)$ .

The similarities end at the level of first order asymptotic properties in correctly specified models, however. As noted in [4], the behavior of the implied probabilities  $\hat{w}_i(\hat{\theta})$  in finite samples differs markedly as a function of the sign of the parameter  $\gamma$ . For ECR with  $\gamma \leq 0$ , the implied probabilities  $\hat{w}_i(\hat{\theta})$  are positive by construction, while for  $\gamma > 0$ , they can take on negative values. In a correctly specified model (where the implied probabilities converge to  $n^{-1}$  for all ECR), negative weights become decreasingly likely as sample size grows and it is possible to entirely avoid negative weights via the use of a “shrinkage factor” correction (see [6]) that vanishes asymptotically and that has no effect on the limiting distribution. Nevertheless, under misspecification, the “shrinkage factor” correction does not vanish asymptotically since negative weights remain likely even asymptotically when  $\gamma > 0$ .

Positive implied probabilities are associated with likelihood-ratio confidence regions whose shape better adapts to the data [4, 50]. For instance, confidence regions for the mean then always lie within the convex hull of the support of the density of the corresponding random variable. Positive implied probabilities are also important in applications that require empirical estimates of the data generating process, as in the bootstrap, for instance [7]. These observations indicate that EL (with  $\gamma = -1$ ) and ET (with  $\gamma = 0$ ) should be preferable to CU (with  $\gamma = 1$ ). CU also suffers from a different problem, namely the potential presence of multiple local maxima in its objective function [22, 30].

Numerous authors have sought to further narrow down the choice of desirable ECR estimators. EL is often singled out among the ECR because it leads to likeli-

hood ratio tests that are often, though not always, Bartlett correctable [10, 14, 39]. Newey and Smith [47] have recently shown that EL generally exhibits a smaller  $O(n^{-1})$  bias than any other member of the ECR family (unless the centered third moments of the distribution of  $g(x_i, \theta^*)$  happen to all vanish, in which case all ECR estimators have the same  $O(n^{-1})$  bias). They have also shown that bias-corrected EL is higher-order efficient, possessing an  $O(n^{-2})$  variance that is no greater than that of any other bias-corrected regular method of moment estimator.

### 2.3 Behavior under misspecification

As mentioned in the introduction, in the presence of misspecification, the object of interest is the pseudo-true value of the parameter vector. In the case of MED estimators, the pseudo-true value is defined as the value of  $\theta$  which minimizes the population version of the empirical discrepancy used in the estimation procedure.

It is important to note that although two different estimators may consistently estimate the truth in a correctly specified model, they may converge in probability to different pseudo-true values in the presence of misspecification. These two pseudo-true values merely represent the minimizers of two different well-defined discrepancies. Even though it could be argued that pseudo-true values are generally “biased”, the literature on estimation under model misspecification considers estimators of pseudo-true values as valid statistics for the purpose of inference (see [56], as an early reference). Following the recent literature using various ECR estimators under model misspecification, we will not argue whether any ECR has a “better” pseudo-true value than another in a given context. Instead, we will compare the convergence of various ECR estimators towards their respective pseudo-true values — a property that will be relevant regardless of the context of interest.

Imbens, Spady and Johnson [30] have informally argued that EL may be ill-

behaved under model misspecification due to the fact that its influence function [20] is proportional to

$$\frac{1}{1 - \lambda' g(x_i, \theta^*)} \frac{\partial g'(x_i, \theta^*)}{\partial \theta} \lambda,$$

where the denominator  $(1 - \lambda' g(x_i, \theta^*))$  can approach zero. We formalize this concern by showing that EL suffers from a dramatic degradation of its asymptotic properties under even the slightest amount of misspecification.

**Theorem 1** *Let  $x_i$  be an i.i.d. sequence and assume  $g(x, \theta)$  is twice continuously differentiable in  $\theta$  for all  $x$  and all  $\theta \in \Theta$  and such that  $\sup_{\theta \in \Theta} E \left[ \|g(x_i, \theta)\|^2 \right] < \infty$ . If  $\inf_{\theta \in \Theta} \|E[g(x_i, \theta)]\| \neq 0$  and  $\sup_{x \in \mathcal{X}} u' g(x, \theta) = \infty$  for any  $\theta \in \Theta$  and any unit vector  $u$ , then there exists no  $\theta_{EL}^* \in \Theta$  such that  $\|\hat{\theta}_{EL} - \theta_{EL}^*\| = O_p(n^{-1/2})$ .*

This theorem can be extended to the case where the moment function  $g(x_i, \theta)$  diverges only along some directions  $u$  but not others. In that case, the lack of root  $n$  consistency is avoided only when  $E[g(x_i, \theta^*)]$  happens to be orthogonal to the hyperplane along which  $g(x_i, \theta)$  diverges.

While Theorem 1 does not prevent  $\hat{\theta}_{EL}$  from being a consistent estimator of the pseudo-true value  $\theta_{EL}^*$ , it does preclude  $\hat{\theta}_{EL}$  from being root  $n$  consistent, under the assumptions of the theorem. The proof of this result, which can be found in the Appendix, is somewhat involved, because standard asymptotics break down for EL under misspecification with unbounded  $g(x, \theta)$ . The following heuristic argument illustrates the nature of the problem: First note that the EL implied probabilities are given by

$$\hat{w}_i = n^{-1} (1 - \hat{\lambda}' g(x_i, \theta^*))^{-1} \tag{8}$$

and must be positive [49]. This implies that  $\hat{\lambda} \xrightarrow{p} 0$  for otherwise,  $\max_{i \leq n} \hat{\lambda}' g(x_i, \theta^*)$  would become unbounded as  $n \rightarrow \infty$ , causing some  $\hat{w}_i$  to become negative. Now, the

population version of the first order condition for  $\hat{\lambda}$  is  $E [g(x_i, \theta^*) / (1 - \lambda^{*'} g(x_i, \theta^*))] = 0$  where  $\lambda^*$  and  $\theta^*$  denote pseudo-true values. Yet, at the pseudo-true value  $\lambda^* \equiv \text{plim } \hat{\lambda} = 0$ , this expectation takes the value  $E [g(x_i, \theta^*)]$ , which is not zero, by the assumption of misspecification. Hence, the asymptotics of EL cannot be determined from a standard expansion of the first-order conditions around the pseudo-true values that satisfy the first-order conditions in the population. The limit as  $n \rightarrow \infty$  and as  $\hat{\lambda} \xrightarrow{P} 0$  cannot be freely exchanged, indicating that the moments entering the first-order conditions violate the standard dominance regularity conditions used to establish the asymptotics of  $M$ -estimators [46].

Theorem 1 indicates that, unless one is willing to solely use moment functions that take values in a compact set (so that  $\sup_{x \in \mathcal{X}} u' g(x, \theta^*)$  is bounded for any  $u$ ), the slightest amount of misspecification can cause the first-order asymptotic properties of EL to degrade catastrophically. It is important to realize that it is very common that the *function*  $g(x, \theta)$  itself is unbounded even when  $E [g(x, \theta)]$  is finite. For instance, if  $g(x, \theta) = (x_1 - \theta, x_2 - 1)'$  and  $x = (x_1, x_2)$  is drawn from a bivariate normal,  $g(x, \theta)$  is unbounded even though  $E [g(x, \theta)]$  exists.

Of course, when the main hypothesis of Theorem 1 ( $\sup_{x \in \mathcal{X}} \|g(x, \theta)\| < \infty$ ) does not hold, root  $n$  consistency becomes possible. For instance, the type moment conditions advocated in the robustness literature (e.g. [20, 24]) involve bounded functions and root  $n$  consistent estimation under misspecification therefore possible using EL. Nevertheless, Theorem 1 does rule out moment conditions such as a simple average of random variables drawn from a distribution with an unbounded support.

Theorem 1 is especially important given the growing literature on minimum empirical discrepancy estimators in misspecified models [8, 23, 26, 32, 34, 54, 61]. In the non-nested model selection literature using minimum discrepancies, it is of-

ten assumed that the competing models may be all misspecified and one is merely concerned with choosing the *least* misspecified model (e.g. [8, 32, 34]). Since the model that is eventually used for inference may then be misspecified, Theorem 1 is particularly relevant in this context and indicates that EL may not be well-suited to these applications — unless the assumption of bounded  $g(x_i, \theta)$  is made, which is precisely the assumption that the model selection literature using EL has so far relied upon [8, 23, 34].

EL’s implied probability weights also exhibit a questionable behavior under misspecification with unbounded  $g(x_i, \theta)$ . Since the EL implied probabilities  $w_i = n^{-1} \left(1 - \hat{\lambda}' g(x_i, \theta)\right)^{-1}$  must be positive [4], it is straightforward to see that  $\hat{\lambda} \xrightarrow{p} 0$  when  $g(x_i, \theta)$  is unbounded. Then note that the implied probabilities associated with all points  $x_i$  such that  $g(x_i, \hat{\theta}) \in C$  for a given compact set  $C$  converge to  $n^{-1}$  uniformly. Since this result holds for any compact set, this shows that, as sample size grows, all the adjustments to the implied probabilities become concentrated on the extreme observations. This would be desirable if the weights of these extreme observations were always decreased to ensure that the moment conditions are satisfied, but this is not the case. In fact, due to the convexity of EL’s tilting function  $\tau(\xi) = 1/(1 - \xi)$ , the reweighting of the sample in order to satisfy the misspecified moment conditions will be achieved by placing a large weight on a few extreme observations, while slightly reducing the weights (relative to  $n^{-1}$ ) of the bulk of the observations. Note that this problem is exacerbated by the fact that the weights can become extremely large as the singularity in the tilting function is approached. This feature will be visible in our simulations below.

We conjecture that any ECR estimator with  $\gamma < 0$  exhibits the same problems as EL under misspecification due to the presence of a ratio in the tilting function. Thus,

if we solely focus on ECR which preclude negative implied probabilities ( $\gamma \leq 0$ ), we are left with ET (corresponding to  $\gamma = 0$ ) as the only candidate ECR whose behavior might not degrade dramatically under misspecification. This is precisely the choice made in [32] for the analysis of misspecified moment restriction models: The asymptotic variance of ET under misspecification is finite under reasonable assumptions, the most restrictive of which is slightly stronger than the requirement of the existence of the moment generating function  $M_\theta(\lambda) = E[\exp(\lambda'g(x_i, \theta))]$  for  $\theta$  and  $\lambda$  in some bounded sets.

### 3 Exponentially tilted empirical likelihood

Higher-order asymptotic properties in correctly specified models point to EL, while good behavior under misspecification points toward ET. There thus appear to be significant benefits to be able to combine EL and ET into a single estimator exhibiting the advantages of both.

It has been suggested [10, 47, 50] that other GEL estimators that exhibit the same higher-order properties as EL can be devised by simply employing a tilting function  $\tau(\xi)$  which admits the same Taylor expansion as the tilting function of EL in the vicinity of  $\xi = 0$  up to sufficiently high order. The behavior of  $\tau(\xi)$  further away from  $\xi = 0$  could then be independently set to match the behavior of ET under misspecification. This option is not particularly attractive because (i) the estimator completely loses its interpretation as a minimum empirical discrepancy estimator, (ii) the estimator can no longer be seen as either a maximum likelihood or a maximum entropy estimator, concepts that initially motivated the form of the EL or ET estimators, and (iii) there still exist an infinite number of ways to interpolate between EL and ET in order to construct  $\tau(\xi)$ , making the procedure highly non

unique. For these reasons we focus on a different approach.

### 3.1 The estimator

We propose to combine the EL and ET estimators in the following fashion.

**Definition 2** (*ETEL estimator*)

$$\hat{\theta} = \arg \min_{\theta} \left( n^{-1} \sum_i \tilde{h}(\hat{w}_i(\theta)) \right), \quad (9)$$

where  $\hat{w}_i(\theta)$  is the solution to

$$\min_{\{w_i\}_{i=1}^n} n^{-1} \sum_i h(w_i) \quad (10)$$

subject to

$$\sum_i w_i g(x_i, \theta) = 0 \text{ and } \sum_i w_i = 1, \quad (11)$$

and where

$$\tilde{h}(w_i) = -\ln(nw_i) \quad (12)$$

$$h(w_i) = nw_i \ln(nw_i). \quad (13)$$

The discrepancies used in the above optimization problem correspond to using ET to find  $\hat{w}_i(\theta)$  and using EL to find  $\hat{\theta}$ . Since  $h(\cdot)$  belongs to the family of ECR discrepancies, this type of estimator still admits a  $N_g + N_\theta$  dimensional dual optimization problem of the form:

$$\hat{\theta} = \arg \min_{\theta} n^{-1} \sum_i \tilde{h}(\hat{w}_i(\theta)), \quad (14)$$

where  $\hat{w}_i(\theta)$  is given by Equation (7) with

$$\hat{\lambda}(\theta) = \arg \max_{\lambda} \left( n^{-1} \sum_i \rho(\lambda' g(x_i, \theta)) \right) \quad (15)$$

and  $\rho(\xi) = -\exp(\xi)$ . This approach yields a unique estimator that combines the likelihood form of EL (Equation (9)) while incorporating the concept of entropy

characterizing ET (Equation (10)). For these reasons, we call this estimator Exponentially Tilted Empirical Likelihood (ETEL). Other authors [10, 31, 40] have considered this combination of EL and ET for the purpose of constructing likelihood-ratio confidence regions for the mean. It has also been shown that a nonparametric Bayesian procedure based on a prior on the space of distributions that favors distributions having a large entropy yields a posterior whose maximum would define the ETEL estimator [58]. This paper’s contribution will be to identify the numerous desirable asymptotic properties of ETEL point estimates in the case of general moment functions  $g(x_i, \theta)$  in the context of *overidentified* and possibly misspecified models.

The fact that the ETEL point estimate is the solution to two nested optimization problems (one of dimension  $N_g$  and one of dimension  $N_\theta$ ) instead of a single saddle-point problem does not complicate the implementation of the estimator. Indeed, ECR estimators are often implemented as two nested optimization problems despite their saddle point form, because it is easier to design robust numerical methods for locating either a maxima or a minima that do not break down near inflection points of the objective function [43].

ETEL represents only one of the many possible combinations between two different discrepancies (one to find the  $\hat{w}_i(\theta)$  and one to find  $\hat{\theta}$ ). However, using the EL discrepancy to find  $\hat{\theta}$  stands out as a particularly attractive choice because the optimization problem defining  $\hat{\theta}$  maintains the maximum likelihood form of EL, thus making it more likely that EL’s higher-order properties will be preserved, an issue that will be investigated below. The use of the ET discrepancy to find the weights  $\hat{w}_i(\theta)$  is also natural. Since the objective function for  $\hat{\theta}$  contains  $\ln(\hat{w}_i(\theta))$ , it is imperative that the weights  $\hat{w}_i(\theta)$  be positive by construction and not only

asymptotically in correctly specified models. As noted earlier, if we focus on weights obtained from the ECR family, in order to maintain the low-dimensional dual formulation, only ECR with  $\gamma \leq 0$  provide positive weights by construction [4]. However, any ECR with  $\gamma < 0$  contains a singularity in its influence function, leaving  $\gamma = 0$ , or ET, as the only sensible choice to find the weights in the presence of potential model misspecification.

From a conceptual point of view, one may wonder about the interpretation of the ETEL estimator, since its definition combines two different discrepancies. It is often pointed out that in the case of discrete distributions, EL provides maximum likelihood estimates of both  $\theta^*$ , the true value of the parameter vector of interest, *and* the weights. Since ET weights are used in ETEL, ETEL weights are not maximum likelihood estimates, but in itself, this is not a great concern since the weights are nuisance parameters and inference focuses on  $\theta^*$ . Indeed, after solving for all the parameters in terms of  $\theta$ , both the ETEL and EL estimators of  $\theta^*$  can be cast into the familiar form of a maximum likelihood estimator of  $\theta^*$  (as opposed to both  $\theta^*$  and the weights, as in EL):

$$\hat{\theta} = \arg \max_{\theta} \left( n^{-1} \sum_i \ln (\hat{w}_i(\theta)) \right), \quad (16)$$

where  $\hat{w}_i(\theta)$  is given by Equations (7) and (5). Of course, such an estimator can only formally be identified to a maximum likelihood estimator in the special case of a discrete distribution having a support consisting of a finite number of points. More generally, for continuous distributions, we can nevertheless refer to  $\hat{\theta}$  as a MED estimator of  $\theta^*$  using the KLIC discrepancy (as for maximum likelihood estimators), an interpretation that will be relevant under model misspecification. The distinction between EL and ETEL lies in how the estimate of the distribution of the data generating process  $\hat{w}_i(\theta)$  given  $\theta$  is constructed. In a parametric likelihood,  $\hat{w}_i(\theta)$

would be uniquely given by the distributional assumptions of the model. When moment conditions replace distributional assumptions, however, there exists no such unique choice of  $\hat{w}_i(\theta)$ , due to the nonparametric nature of the problem. Both EL and ETEL replace parametric distributional assumptions by a so-called *least favorable family* of distributions (see, for instance, [15]), that is, a parametric family of distributions (indexed by  $\theta$ ) for which the estimation problem is as difficult as the original nonparametric problem. In other words, for each  $\theta$ , there exists an infinite number of distributions satisfying the moment conditions, and the specific discrete distribution defined by  $\hat{w}_i(\theta)$  represents a worst-case scenario among them. As pointed out in [15], there exist numerous least favorable families; EL and ETEL merely employ different ones and, a priori, there is no reason to favor one over the other.

In the case of ETEL, the least favorable family chosen is the class of distributions obtained by maximizing entropy under the  $\theta$ -dependent moment constraints imposed by the model. Entropy maximization has a long history as a device to construct distributions which properly model lack of prior information under a set of known constraints (see, for instance, [12, 17, 36, 38, 60]). ETEL thus combines the well-established concept of entropy maximization to handle the nonparametric part of the estimation problem, while using likelihood maximization to deal with the parametric part of the problem. The idea of substituting nonparametric non-maximum likelihood estimates (here, the  $\hat{w}_i(\theta)$ ) into a likelihood-type objective function to avoid the pathological behavior of an approach based solely on maximum likelihood also parallels the work of Fan and co-workers [16].

One may have preferences regarding which estimator of the distribution is the more appealing, but the choice between EL and ETEL should ultimately be based

on the comparison of the actual asymptotic properties of each estimator and their performance in simulation experiments, which is what the remainder of this article is devoted to.

## 3.2 Properties

### 3.2.1 First-order properties

To simplify the notation, we make the dependence of all quantities on  $\theta$  implicit and introduce the following definitions.

**Definition 3**  $\hat{w}_i = \hat{w}_i(\theta)$ ,  $\hat{\lambda} = \hat{\lambda}(\theta)$ ,  $g_i = g(x_i, \theta)$ ,  $\hat{g} = n^{-1} \sum_i g(x_i, \theta)$ ,  $G_i = \partial g(x_i, \theta) / \partial \theta'$ ,  $G = E[G_i]$ ,  $\hat{G} = n^{-1} \sum_i G_i$ ,  $\tilde{G} = \sum_i \hat{w}_i G_i$ ,  $\hat{\Omega} = n^{-1} \sum_i g_i g_i'$ ,  $\Omega = E[g_i g_i']$ ,  $\tilde{\Omega} = \sum_i \hat{w}_i g_i g_i'$ . Quantities evaluated at  $\theta = \theta^*$  are denoted by  $*$ .

Simple algebraic manipulations yield the following.

**Theorem 2** *The ETEL estimator  $\hat{\theta}_{ETEL}$  maximizes the objective function*

$$\ln \hat{L}(\theta) = -\ln \left( n^{-1} \sum_i \exp(\hat{\lambda}'(g_i - \hat{g})) \right) \quad (17)$$

where  $\hat{\lambda}$  is such that

$$n^{-1} \sum_i \exp(\hat{\lambda}' g_i) g_i = 0. \quad (18)$$

The first-order conditions for  $\hat{\theta}_{ETEL}$  can be written as

$$n^{-1} \sum_i (1 - n \hat{w}_i) \frac{d(\hat{\lambda}' g_i)}{d\theta'} = 0, \quad (19)$$

where the total derivative indicates that  $\hat{\lambda}$  is allowed to vary with  $\theta$ .

We then establish that ETEL is at least as good as any ECR estimator both in terms of its first-order asymptotic properties and in terms of its invariance properties.

**Assumption 1** (*Regularity conditions*)

1.  $x_i$  forms an i.i.d. sequence,
2.  $\theta^* \in \text{int}(\Theta)$  is the unique solution to  $E[g(x_i, \theta)] = 0$ , where  $\Theta$  is compact,
3.  $g(x_i, \theta)$  is continuous (in  $\theta$ ) at each  $\theta \in \Theta$  with probability one,
4.  $E\left[\sup_{\theta \in \Theta} \|g_i\|^{2+\delta}\right] < \infty$  for some  $\delta > 0$  and  $E[\sup_{\theta \in \mathcal{N}} \|G_i\|] < \infty$ ,
5.  $\Omega^*$  is nonsingular and finite and  $\text{rank}(G^*) = N_\theta$ ,
6.  $g(x_i, \theta)$  is continuously differentiable (in  $\theta$ ) in a neighborhood  $\mathcal{N}$  of  $\theta^*$ .

These assumptions match those of Theorem 3.2 in [47] and include those of Theorem 3.4 in [50].

**Theorem 3** (*First-order properties*) Under Assumptions 1, the ETEL estimator (i) has the same limiting distribution as efficient two-step GMM,

$$n^{1/2} \left( \hat{\theta} - \theta^* \right) \xrightarrow{d} N(0, \Sigma)$$

where  $\Sigma = \left( G^{*'} (\Omega^*)^{-1} G^* \right)^{-1}$  (ii) ETEL enables the construction of  $\chi^2$ -calibrated likelihood-ratio confidence regions for  $\theta$ :

$$-2n \ln \left( \hat{L}(\theta) / \hat{L}(\hat{\theta}) \right) \xrightarrow{d} \chi_{N_\theta}^2$$

and (iii) of  $\chi^2$ -calibrated test of the validity of overidentifying restrictions

$$-2n \ln \left( \hat{L}(\hat{\theta}) \right) \xrightarrow{d} \chi_{N_g - N_\theta}^2.$$

**Theorem 4** (*Implied probabilities and invariance properties*) Whenever the ETEL estimator is defined, (i) it yields positive implied probabilities ( $\hat{w}_i(\theta) \geq 0$ ) (ii)

it is invariant under arbitrary one-to-one differentiable re-parametrization  $\theta = T(\beta)$  of the moment conditions (the estimate  $\hat{\beta}$  obtained from the reparametrized moment conditions satisfies  $\hat{\theta} = T(\hat{\beta})$ ) and (iii) it is invariant under general parameter-dependent nonsingular linear transformation  $A(\theta)$  of the vector of moment conditions (using  $E[A(\theta)g(x_i, \theta)] = 0$  or  $E[g(x_i, \theta)] = 0$  as moment conditions gives the same  $\hat{\theta}$ ).

### 3.2.2 Higher-order asymptotic properties

Estimators having the same (first-order) asymptotic variance can be compared on the basis of their higher-order ( $o_p(n^{-1/2})$ ) asymptotic properties [55]. While it has been established that likelihood-ratio confidence regions of the mean constructed using ETEL do not share EL’s Bartlett correctability [10, 31], another type of analytic higher-order correction permits the same improvement in the order of the coverage accuracy [40]. Moreover, it has been observed in simulation studies [50, 62] that the Bartlett correction is often ineffective in practice because the “QQ” plots for the EL likelihood ratio test statistics are typically curved, making it unlikely that a linear correction such as Bartlett’s would improve coverage accuracy. Finally, given that ETEL’s objective function can be interpreted as a posterior for the parameter  $\theta$  obtained via a nonparametric Bayesian procedure [58], it may be a more relevant and interesting topic of future research to verify whether a Bayesian Bartlett correction [5], which differs from the usual frequentist Bartlett correction, would be applicable to ETEL.

More importantly, we can show that the ETEL point estimate  $\hat{\theta}_{ETEL}$  shares all of the other higher-order properties of EL established in [47]. Higher-order asymptotic properties of an estimator  $\hat{\theta}$  are defined through a stochastic expansion (see, for

instance, [52, 55]) of the form

$$\left(\hat{\theta} - \theta^*\right) = n^{-1/2}\bar{\psi} + n^{-1}\bar{q} + n^{-3/2}\bar{r} + O_p(n^{-2}) \quad (20)$$

where  $\bar{\psi}$ ,  $\bar{q}$ , and  $\bar{r}$  are  $O_p(1)$  and where  $\bar{\psi}$ , and  $\bar{r}$  have zero mean. Within this framework, the  $O(n^{-1})$  bias is defined as  $E[\bar{q}]$  and represents the most important correction to standard first-order asymptotic based solely on the influence function  $\bar{\psi}$ . Another important correction to first-order asymptotics is the  $O(n^{-2})$  variance, defined as

$$\text{Var}[\bar{q}] + \text{Covar}[\bar{r}, \bar{\psi}] + \text{Covar}[\bar{\psi}, \bar{r}]. \quad (21)$$

This expression can be informally obtained by computing the variance of Equation (20). In general, it is not meaningful to compare the  $O(n^{-2})$  variance of two estimators that possess a different  $O(n^{-1})$  bias and bias-corrected estimators should be used to compare efficiency.

We now proceed to compare the stochastic expansions of  $\hat{\theta}_{ETEL}$  and  $\hat{\theta}_{EL}$ , using assumptions found in [47]. Our approach consists of establishing that the difference  $\hat{\theta}_{ETEL} - \hat{\theta}_{EL}$  is such that the Newey and Smith's results for  $\hat{\theta}_{EL}$  carry over to  $\hat{\theta}_{ETEL}$ .

**Theorem 5 (Higher-order equivalence to EL)** *Under Assumption 1 and if  $E[\sup_{\theta \in \mathcal{N}} \|g_i\|^4] < \infty$ ,  $E[\sup_{\theta \in \mathcal{N}} \|G_i\|^2] < \infty$  and for  $\theta \in \mathcal{N}$ ,  $G(x_i, \theta)$  is Lipschitz in  $\theta$  with prefactor  $b(x_i)$  such that  $E[b(x_i)] < \infty$ , then  $\hat{\theta}_{ETEL} - \hat{\theta}_{EL} = O_p(n^{-3/2})$ .*

A consequence of this result is that the ETEL estimator has the same  $O(n^{-1})$  bias as the EL estimator obtained in [47], under their assumptions. (As shown in [59], this result in fact extends to all estimators constructed by substituting GEL weights into the EL objective function.)

**Assumption 2** *There exists a function  $b(x_i)$  with  $E[(b(x_i))^6] < \infty$  such that, in a neighborhood  $\mathcal{N}$  of  $\theta^*$ , all partial derivatives of  $g(x_i, \theta)$  with respect to  $\theta$  up to order 4 exist, are bounded by  $b(x_i)$  and are Lipschitz in  $\theta$  with prefactor  $b(x_i)$*

**Theorem 6 (Small bias property)** *Under Assumptions 1 and 2, ETEL's  $O(n^{-1})$  bias is*

$$n^{-1}H(-a + E[G_i H g_i])$$

where  $H = \Sigma G' \Omega^{-1}$  and  $a$  is a vector whose elements are  $a_j = \text{tr}(\Sigma E[\partial^2 g_j(x_i, \theta^*) / \partial \theta \partial \theta']) / 2$  where  $g_j(x_i, \theta^*)$  is the  $j$ -th element of  $g(x_i, \theta^*)$ .

A simple intuition for the small bias of EL is that the EL first-order condition resembles the first order condition of GMM ( $\hat{g}'\hat{\Omega}^{-1}\hat{G} = 0$ ) except for the fact that the Hessian term  $\hat{\Omega}$  and the Jacobian term  $\hat{G}$  are replaced by efficient averages that are weighted by the EL implied probabilities [47]. This reweighting removes the  $O(n^{-1})$  correlation between the different sample averages entering the first-order condition, thus reducing the bias. As shown in the Appendix, ETEL also efficiently weights the Hessian and the Jacobian terms, only using the ET weights. Since ET and EL implied probabilities are equivalent to a sufficiently high order, using the ET instead of the EL weights only contributes to a negligible  $O_p(n^{-3/2})$  remainder.

The fact that ETEL and EL are equivalent up to  $O_p(n^{-1})$  leads to two important simplifications in the comparison of their  $O(n^{-2})$  variances. First, since their  $O(n^{-1})$  bias are the same, the moments entering the expression for the bias correction of EL and ETEL are the same. If these moments were estimated in the same way for EL and for ETEL, then comparing the  $O(n^{-2})$  variance of EL and ETEL with or without performing a bias correction would obviously give the same answer. This conclusion remains unchanged if the bias correction is applied using

the EL estimate of  $\theta$  for the EL bias correction and the ETEL estimate of  $\theta$  for the ETEL bias correction since these estimates differ by  $O_p(n^{-3/2})$ , which would give rise to a difference of only  $O_p(n^{-1}n^{-3/2})$  in the bias correction. Moreover, as pointed out in [47], whether the moments entering the bias correction are estimated by sample averages or averages weighted by implied probabilities has no effect on the higher-order variance of the resulting bias-corrected estimator. Hence, using EL weights for the EL bias correction and ETEL weights for the ETEL bias correction makes no difference either. In conclusion, we can meaningfully compare the  $O(n^{-2})$  variances of EL and ETEL before performing a bias correction.

The second simplification made possible by the equivalence of the  $O_p(n^{-1})$  terms of the EL and ETEL stochastic expansions, is that the differences in their  $O(n^{-2})$  variance must take the form

$$\text{Covar}[\bar{r}^{ETEL} - \bar{r}^{EL}, \bar{\psi}] + \text{Covar}[\bar{\psi}, \bar{r}^{ETEL} - \bar{r}^{EL}], \quad (22)$$

as seen in Equation (21). Hence, it is possible for ETEL and EL to differ by  $O_p(n^{-3/2})$ , while still sharing the same  $O(n^{-2})$  variance, as long as that difference is uncorrelated with their (identical) influence function  $\bar{\psi}$ . In fact, this is precisely the case, as shown in the Appendix.

**Theorem 7 (*Higher-order efficiency*)** *Under Assumptions 1 and 2, the  $O(n^{-2})$  variances of ETEL and EL are equal.*

Maintaining the maximum likelihood form for the optimization problem defining  $\hat{\theta}_{ETEL}$  thus achieves the desired goal, namely, to maintain the higher-order asymptotic properties of EL found in [47]. It is the fact that (22) vanishes that enables ETEL to be higher-order efficient even though it differs sufficiently from EL to fail to be Bartlett correctable in the frequentist sense.

### 3.2.3 Behavior under misspecification

While in the previous section, we have seen that ETEL inherits the higher-order properties of EL, we will now show that it also exhibits some of the desirable properties of ET that EL lacks under model misspecification.

Following the discussion of Section 3.1, ETEL's pseudo-true value  $\theta^*$  minimizes the KLIC discrepancy between the true data generating process and an entropy maximizing least favorable family of distributions parametrized by  $\theta$  (which replaces the distributional assumptions in parametric maximum likelihood).

We will now study the first-order asymptotic properties of ETEL under misspecification.

**Theorem 8** *For a given  $\theta$ , assume that  $E[\exp(\lambda'g(x_i, \theta))]$  exists in a neighborhood of its minimum. If a subvector of  $g(x_i, \theta)$  is statistically independent from the remaining elements of  $g(x_i, \theta)$ , then the empirical cdf obtained from ETEL (or ET) implied probabilities at  $\theta$  converges pointwise (at every point of continuity of the true cdf) to a cdf that maintains this independence, even under misspecification. EL achieves this only in the absence of misspecification.*

This indicates the possibility that using an empirical cdf obtained from the implied probability weights of EL in the hope of improving accuracy could actually result in the introduction of a spurious dependence among variables. ETEL avoids this unappealing eventuality. This property could be helpful when the implied probabilities are employed to improve the efficiency of the bootstrap, as in [7], when the model happens to be misspecified.

A more important quality that ETEL shares with ET is the nonsingular behavior of its influence function. As noted by [30], an estimator's influence function  $\psi(x_i)$

is proportional to its first-order conditions. By inspection of ETEL’s first-order condition (Equation (19)) it is clear ETEL’s influence function will not contain any singularity, unlike EL’s influence function. It will therefore not be surprising that ETEL avoids EL’s undesirable behavior under misspecification, under regularity conditions similar to the ones made by [32] for ET, as shown more formally below.

Let  $\lambda^*(\theta)$  denote the solution to  $E[\exp(\lambda'g(x_i, \theta))g(x_i, \theta)] = 0$ , which is unique, by the strict convexity of  $E[\exp(\lambda'g(x_i, \theta))]$  in  $\lambda$ .

**Assumption 3** (*Regularity conditions under misspecification*)

1.  $x_i$  forms an *i.i.d.* sequence.
2. The function  $\ln L(\theta) \equiv -\ln(E[\exp(\lambda^{*'}(\theta)(g(x_i, \theta) - E[g(x_i, \theta)]))])$  is maximized at a unique “pseudo-true” value  $\theta^* \in \text{int}(\Theta)$ , where  $\Theta$  is compact.
3.  $g(x_i, \theta)$  is continuous (in  $\theta$ ) at each  $\theta \in \Theta$  with probability one.
4.  $E[\sup_{\theta \in \Theta} \sup_{\lambda \in \Lambda(\theta)} \exp(\lambda'g(x_i, \theta))] < \infty$  where  $\Lambda(\theta)$  is a compact set such that  $\lambda^*(\theta) \in \text{int}(\Lambda(\theta))$ .
5.  $S_{jl}(x_i, \theta) = \partial^2 g(x_i, \theta^*) / \partial \theta_j \partial \theta_l$  is continuous (in  $\theta$ ) for  $\theta \in \mathcal{N}$ , a neighborhood of  $\theta^*$ .
6. There exists  $b(x_i)$  satisfying  $E[\sup_{\theta \in \mathcal{N}} \sup_{\lambda \in \Lambda(\theta)} \exp(k_1 \lambda'g(x_i, \theta)) (b(x_i))^{k_2}] < \infty$  for  $k_1 = 1, 2$  and  $k_2 = 0, 1, 2, 3, 4$  such that  $\|G(x_i, \theta)\| \leq b(x_i)$  and  $\|S_{jl}(x_i, \theta)\| \leq b(x_i)$  for  $j, l = 1, \dots, N_\theta$  for any  $x_i \in \mathcal{X}$  and for all  $\theta \in \mathcal{N}$ .

The simplest way to describe the asymptotics of ETEL under misspecification is to introduce an equivalent just-identified GMM estimator involving an augmented parameter vector  $\beta = (\tau, \kappa', \lambda', \theta')'$ . The vector  $\theta \in \mathbb{R}^{N_\theta}$  is the parameter vector

of interest, while  $(\tau, \kappa', \lambda')' \in \mathbb{R}^{1+2N_g}$  are auxiliary parameters to be estimated jointly with  $\theta$ . The dimension of this augmented parameter vector is higher than in the case of GEL estimators under misspecification ( $1 + 2N_g + N_\theta$  instead of  $N_g + N_\theta$ ). This is due to the fact that the first-order conditions for  $\hat{\theta}$  in ETEL involve a few additional terms taking the form of a product of sample moments that are absent in GEL estimators. Each of these products of sample moments can be linearized by introducing the additional parameters  $\kappa$  and  $\tau$ . Note that these additional parameters are merely a device used to simplify the construction of the covariance matrix of the estimator. The point estimate  $\hat{\theta}$  can be obtained without introducing  $\kappa$  and  $\tau$ , as seen in Theorem 2.

**Lemma 9** *The ETEL point estimate  $\hat{\theta}$  is given by the appropriate subvector of the vector  $\hat{\beta} = (\hat{\tau}, \hat{\kappa}', \hat{\lambda}', \hat{\theta}')'$ , the solution to*

$$n^{-1} \sum_i \phi(x_i, \hat{\beta}) = 0$$

where, letting  $\hat{\tau}_i = \exp(\hat{\lambda}' g_i)$ ,

$$\phi(x_i, \hat{\beta}) = \begin{bmatrix} \hat{\tau}_i - \hat{\tau} \\ \frac{\partial}{\partial \kappa} (\hat{\tau}_i g_i' \hat{\kappa} + \hat{\tau} g_i' \hat{\lambda} - \hat{\tau}_i) \\ \frac{\partial}{\partial \lambda} (\hat{\tau}_i g_i' \hat{\kappa} + \hat{\tau} g_i' \hat{\lambda} - \hat{\tau}_i) \\ \frac{\partial}{\partial \theta} (\hat{\tau}_i g_i' \hat{\kappa} + \hat{\tau} g_i' \hat{\lambda} - \hat{\tau}_i) \end{bmatrix} = \begin{bmatrix} \hat{\tau}_i - \hat{\tau} \\ \hat{\tau}_i g_i \\ (\hat{\tau} - \hat{\tau}_i) g_i + \hat{\tau}_i g_i g_i' \hat{\kappa} \\ \hat{\tau}_i G_i' \hat{\kappa} + \hat{\tau}_i G_i' \hat{\lambda} g_i' \hat{\kappa} - \hat{\tau}_i G_i' \hat{\lambda} + \hat{\tau} G_i' \hat{\lambda} \end{bmatrix}. \quad (23)$$

Given the just-identified nature of the estimator defined in Lemma 9, its asymptotic distribution follows quite directly.

**Theorem 10 (Asymptotics under misspecification)** *Let  $\Gamma = E[\partial \phi(x_i, \beta) / \partial \beta']|_{\beta=\beta^*}$  and  $\Phi = E[\phi(x_i, \beta^*) \phi'(x_i, \beta^*)]$ . Under Assumption 3, if  $\Gamma$  is nonsingular, then  $n^{1/2}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, \Gamma^{-1} \Phi (\Gamma')^{-1})$ .*

## 4 Simulations

We first illustrate the fact that ETEL has the same  $O(n^{-1})$  bias as EL. We use the simple experimental design suggested in [19] and subsequently used in [30, 33], slightly expanded to have  $K$  moment conditions rather than 2. The moment conditions are

$$g(x_i, \theta) = [ r(x_i, \theta) \quad r(x_i, \theta) x_{i2} \quad r(x_i, \theta) (x_{i3} - 1) \quad \cdots \quad r(x_i, \theta) (x_{iK} - 1) ]'$$

where  $r(x_i, \theta) = \exp(-0.72 - (x_{i1} + x_{i2})\theta + 3x_{i2}) - 1$ . These restrictions are satisfied at  $\theta^* = 3$  when  $(x_{i1}, x_{i2})' \sim N(0, (0.16)I)$  and  $x_{ik} \sim \chi_1^2$ , for  $k = 3, \dots, K$ . Note that the third moments of all elements of  $g(x_i, \theta)$  are non-zero and that  $g(x_i, \theta)$  is nonlinear in  $\theta$ , so that the  $O(n^{-1})$  bias does not trivially vanish. Figure 1 shows the cdf of the EL, ET and ETEL estimators of  $\theta$  obtained from 10000 replicated samples of the above design (with  $K = 4$  and  $K = 10$ ), each containing 200 observations. (Samples for which at least one of the three estimators considered failed to converge were discarded. This happened 14 times for  $K = 4$  and 32 times for  $K = 10$ . The most frequent reason for failure of convergence was that the origin was not contained within the convex hull of the values of  $g(x_i, \theta)$  for any  $\theta$ , in which case none of the estimators are even defined. The number of nondiscarded samples is 10000.) It is apparent that the ETEL and EL point estimates have very similar distributions, as expected from their equivalence up to the  $O_p(n^{-1})$  term of their stochastic expansion. The distribution of the ET point estimates differs noticeably from the ones of EL and ETEL and the main difference takes the form of a bias, which is reported in Table 2. The bias of ET increases more rapidly with the number of moment conditions than the biases of both EL and ETEL, as the higher-order asymptotics analyses given in [47] and in the present work would suggest.

Our next simulation compares the behavior of EL, ET and ETEL under misspecification. We consider a simple case where we wish to estimate the mean while imposing a known variance. In this example, the moment conditions are

$$g(x_i, \theta) = [ x_i - \theta \quad (x_i - \theta)^2 - 1 ]$$

where  $x_i$  is drawn either from a correctly specified model C or a misspecified model M:

$$\begin{aligned} x_i &\sim N(0, 1) && \text{(for Model C)} \\ x_i &\sim N(0, (0.8)^2) && \text{(for Model M)} \end{aligned}$$

Note that this experiment is specifically designed so that the pseudo-true value ( $\theta^* = 0$ ) for the misspecified model is the same for EL, ET and ETEL, thus enabling a meaningful comparison of the variance of these estimators.

Figure 2 shows the cdf of the EL, ET and ETEL estimators of  $\theta$  for a sample size of 1000 and a sample size of 5000, evaluated with 10000 and 2000 replications, respectively. The variability of the EL estimate is clearly larger than the one of ET and ETEL, as confirmed by the calculated standard deviations given in Table 3. Interestingly, the distributions (and the standard deviations) of the ET and ETEL estimators are quite similar. While the ET and ETEL standard deviations shrink by the expected factor of  $\sqrt{5}$  as the sample size is increased from 1000 to 5000, the standard deviation of EL barely changes, which is not surprising given the results of Theorem 1. Note that the difference between the distribution of EL and the two other estimators can be made arbitrarily large either by increasing the amount of misspecification or by increasing the sample size.

We can also use simulations to illustrate the source of EL's poor behavior under misspecification. Figure 3 shows the implied probabilities for EL and ETEL in two simulated samples of size  $n = 1000$  and  $n = 5000$  drawn from the misspecified

Model M. It is apparent that the EL implied probabilities attribute an excessive weight to the extreme observations. As sample size grows, this trend worsens: The upper right graph exhibits an extremely large weight at  $x_i \approx -3$  and  $nw_i \approx 95$ . In contrast, the ETEL implied probabilities distribute the weight more uniformly over the whole sample and, even more importantly, the weights do not become increasingly concentrated in the tails as sample size grows.

These examples, although simple and perhaps not realistic, illustrate how ETEL matches the low-bias property of the EL estimator and shares the reasonable behavior of ET under misspecification.

## 5 Conclusion

Our first important result is to show that although empirical likelihood (EL) is known to exhibit numerous desirable higher-order asymptotic properties in correctly specified models, its first-order asymptotic properties can degrade catastrophically in the presence of the slightest amount of misspecification, causing the loss of root  $n$  consistency. Although the use of only bounded functions  $g(x_i, \theta)$  in the moment conditions  $E[g(x_i, \theta)] = 0$  avoids this problem, this is a rather strong constraint. In contrast, exponential tilting (ET) is known to be inferior to EL in terms of its higher-order properties, but remains well behaved in the presence of misspecification under relatively weak regularity conditions [32].

Our second main contribution is to show that EL and ET can be combined to yield an estimator that exhibits the advantages of both. This so-called Exponentially Tilted Empirical Likelihood (ETEL) has the same low  $O(n^{-1})$  bias and the same  $O(n^{-2})$  variance as EL in correctly specified models, and yet avoids EL's pitfalls in misspecified models.

## A Proofs

The quantities given in Definitions 1 and 3 will be used throughout the Appendix. Let  $C$  denote a generic constant which may take distinct values in different contexts. Let CSI stand for Cauchy-Schwartz Inequality and let w.p.a. 1 stand for the phrase “with probability approaching one”.

**Proof of Theorem 1.** The proof proceeds by constructing a triangular array of estimators  $\hat{\theta}_{k,n}$  indexed by sample size  $n$  and by an auxiliary truncation parameter  $k$ . To define this array, let  $\mathcal{G}_k$  be increasing sequence of nested compact subsets of  $\mathbb{R}^{N_g}$  such that  $\cup_{k=1}^{\infty} \mathcal{G}_k = \mathbb{R}^{N_g}$ . Then, let  $\mathcal{C}_k = \{x \in \mathcal{X} : g(x, \theta) \in \mathcal{G}_k \text{ for all } \theta \in \Theta\}$ . Note that  $\mathcal{C}_k$  is nonempty for  $k$  sufficiently large.

Let  $F_{\infty}(x)$  denote the distribution of  $x$  and let  $\hat{\theta}_{\infty,n}$  denote the EL estimator obtained from a sample of size  $n$  and let  $\theta_{\infty}^*$  denote EL’s pseudo-true value, assuming it exists (for otherwise,  $\hat{\theta}_{\infty,n}$  could not even be consistent).

Let  $F_k(x)$  be a sequence of distributions indexed by  $k \in \mathbb{N}$ , each having support  $\mathcal{C}_k$ . We choose  $F_k(x)$  so that, for all  $k$  sufficiently large, the moment conditions are uniformly misspecified ( $\inf_{k \geq \bar{k}} \inf_{\theta \in \Theta} \|E_{F_k}[g(x, \theta)]\| > 0$  for some  $\bar{k} \in \mathbb{N}$ ). Let  $\hat{\theta}_{k,n}$  denote the EL estimator in a sample size of  $n$  when the true data generating process is  $F_k(x)$  and let  $\theta_k^* \in \Theta$  denote the corresponding pseudo-true value. We then note that it is also always possible to choose a distribution  $F_k(x)$  with support  $\mathcal{C}_k$  such that  $P \left[ \left| u' \left( \hat{\theta}_{k,n} - \theta_k^* \right) \right| \geq \varepsilon \right] \leq P \left[ \left| u' \left( \hat{\theta}_{\infty,n} - \theta_{\infty}^* \right) \right| \geq \varepsilon \right]$  for any  $\varepsilon > 0$ , any conformable unit vector  $u$  and all  $n$ . For instance, one could first construct a distribution  $\tilde{F}_k(x)$  equal to  $F_{\infty}(x)$  conditional on the event  $x \in \mathcal{C}_k$ . Let  $\theta_k^*$  denote the pseudo-true value associated with  $\tilde{F}_k(x)$ . Then set  $F_k(x)$  to be a mixture of  $\tilde{F}_k(x)$  and a degenerate distribution that would give  $\theta_k^*$  as an EL estimate with certainty. In this fashion,  $F_k(x)$  is a “truncated” version of  $F_{\infty}(x)$  designed to make

the estimation of  $\theta_k^*$  by  $\hat{\theta}_{k,n}$  easier than the estimation of  $\theta_\infty^*$  by  $\hat{\theta}_{\infty,n}$ . Obviously,  $\hat{\theta}_{k,n}$  is an infeasible estimator that uses out of sample information. It is introduced solely for the purpose of facilitating the proof. Note that  $\theta_k^* \neq \theta_\infty^*$  in general, but the proof will never require that  $\theta_k^* = \theta_\infty^*$ .

For a distribution  $F_k(x)$  having compact support, the EL estimator can be written as a just identified GMM estimator of an augmented parameter vector  $\hat{\beta} = (\hat{\theta}'_{k,n}, \hat{\lambda}'_{k,n})'$  satisfying the first-order conditions

$$n^{-1} \sum_i G' \left( x_i, \hat{\theta}_{k,n} \right) \hat{\lambda}_{k,n} / \left( 1 - \hat{\lambda}' g \left( x_i, \hat{\theta}_{k,n} \right) \right) = 0 \quad (24)$$

$$n^{-1} \sum_i g \left( x_i, \hat{\theta}_{k,n} \right) / \left( 1 - \hat{\lambda}' g \left( x_i, \hat{\theta}_{k,n} \right) \right) = 0, \quad (25)$$

Note that these first-order conditions form a just-identified system of equations, whether the model is correctly specified or not. Hence, in this formulation the standard asymptotic theory of just-identified GMM estimators applies [46] (see also [32] for the application of this idea to ET under misspecification). The asymptotic variance of a just-identified GMM of the form  $n^{-1} \sum_i \phi \left( x_i, \hat{\beta} \right) = 0$  is given by

$$\left( E \left[ \partial \phi' \left( x_i, \beta \right) / \partial \beta \right] \right)^{-1} \left( E \left[ \phi \left( \beta \right) \phi' \left( \beta \right) \right] \right) \left( E \left[ \partial \phi' \left( x_i, \beta \right) / \partial \beta \right] \right)^{-1}. \quad (26)$$

For  $k$  sufficiently large, we can always choose  $F_k(x)$  so as to satisfy the necessary regularity conditions for this expression to hold. In particular, the compact support of  $F_k(x)$  enables  $E \left[ g \left( x_i, \theta \right) / \left( 1 - \lambda' g \left( x_i, \theta \right) \right) \right]$  to exist for  $(\theta', \lambda)'$  in some neighborhood of the pseudo-true value  $(\theta_k^{*'}, \lambda_k^{*'})'$ . The asymptotic distribution of  $(\hat{\theta}'_{k,n}, \hat{\lambda}'_{k,n})'$  is then given by

$$n^{1/2} \left( (\hat{\theta}'_{k,n}, \hat{\lambda}'_{k,n}) - (\theta_k^{*'}, \lambda_k^{*'}) \right)' \xrightarrow{d} N \left( 0, H_k^{-1} S_k H_k^{-1} \right) \text{ as } n \rightarrow \infty \text{ for fixed } k \quad (27)$$

where

$$S_k = E \left[ \begin{pmatrix} \tau_i^2 G_i' \lambda_k^* \lambda_k^{*'} G_i & \tau_i^2 G_i' \lambda_k^* g_i' \\ \tau_i^2 g_i \lambda_k^{*'} G_i & \tau_i^2 g_i g_i' \end{pmatrix} \right] \quad (28)$$

$$H_k = E \left[ \begin{pmatrix} \dot{\tau}_i G_i' \lambda_k^* \lambda_k^{*'} G_i + \tau_i \frac{\partial(G_i' \lambda_k^{*'})}{\partial \theta'} & \dot{\tau}_i G_i' \lambda_k^* g_i' + \tau_i G_i' \\ \dot{\tau}_i g_i \lambda_k^{*'} G_i + \tau_i G_i & \dot{\tau}_i g_i g_i' \end{pmatrix} \right] \quad (29)$$

and where,  $\tau_i = \tau(\lambda_k^{*'} g_i) = (1 - \lambda_k^{*'} g_i)^{-1}$ ,  $\dot{\tau}_i = \frac{\partial \tau(\xi)}{\partial \xi} \Big|_{\xi = \lambda_k^{*'} g_i} = (1 - \lambda_k^{*'} g_i)^{-2} = \tau_i^2$  and where all moments are evaluated at  $\theta_k^*$  and  $\lambda_k^* = \text{plim}_{n \rightarrow \infty} \hat{\lambda}_{k,n}$  and assuming that  $x$  is drawn from  $F_k(x)$  (that is,  $E[\cdot] \equiv E_{F_k}[\cdot]$ ).

We focus on the upper left  $N_\theta \times N_\theta$  submatrix of  $H_k^{-1} S_k H_k^{-1}$ , denoted by  $\Sigma_k$ . For a given  $k$ , the submatrix  $\Sigma_k$  provides the asymptotic variance of  $\hat{\theta}_{k,n}$ . We will now analyze the behavior of  $\Sigma_k$  as  $k \rightarrow \infty$  (we are not claiming that this provides the asymptotic variance of EL for infinite support at this point). Since EL's implied probabilities must be positive (see, for instance, [4, 50]), it follows that  $(1 - \lambda_k^{*'} g(x, \theta_k^*))^{-1} > 0$  for all  $x \in \mathcal{C}_k$ , or

$$\max_{x \in \mathcal{C}_k} (\lambda_k^{*'} g(x, \theta_k^*)) < 1. \quad (30)$$

Since  $\{g(x, \theta_k^*) : x \in \mathcal{X}\}$  is unbounded in every direction, the set  $\{g(x, \theta_k^*) : x \in \mathcal{C}_k\}$  becomes unbounded in every direction as  $k \rightarrow \infty$ . Hence the only way to satisfy Equation (30) is to have  $\lambda_k^* \rightarrow 0$  as  $k \rightarrow \infty$ . Since  $\lambda_k^* \rightarrow 0$  as  $k \rightarrow \infty$ , the expressions for  $S_k$  and  $H_k$  can be simplified by noting that when the product  $H_k^{-1} S_k H_k^{-1}$  is calculated, any term containing  $\lambda_k^*$  will be dominated by terms not containing  $\lambda_k^*$ . We then obtain (keeping the  $\tau_i = \tau(\lambda_k^{*'} g_i)$  prefactors even though  $\lambda_k^* \rightarrow 0$  because the  $g(x, \theta_k^*)$  are unbounded and it is not clear whether we necessarily have  $\tau_i \rightarrow 1$ ):

$$S_k \rightarrow \begin{bmatrix} 0 & 0 \\ 0 & E[\tau_i^2 g_i g_i'] \end{bmatrix}$$

$$H_k^{-1} \rightarrow \begin{bmatrix} 0 & E[\tau_i G_i'] \\ E[\tau_i G_i] & E[\tau_i^2 g_i g_i'] \end{bmatrix}^{-1} \equiv \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \quad (31)$$

(Note that the sequence  $F_k$  can be easily chosen so that the smallest eigenvalue  $H_k$  remains bounded away from zero for all  $k$  sufficiently large, since the moment conditions remain the same over  $k$  and  $\mathcal{G}_k$  increases with  $k$ . Hence,  $\lim_{k \rightarrow \infty} H_k$  can be assumed nonsingular and interchanging the limit as  $k \rightarrow \infty$  and the matrix inversion operation is justified.) We then have that

$$\Sigma_k = B_{12}E[\tau_i^2 g_i g_i'] B_{21} + \rho_k, \quad (32)$$

where  $\rho_k$  is a remainder that vanishes as  $k \rightarrow \infty$  (its precise form has no bearing on the rest of the argument). By the partitioned inverse formula,

$$B_{21} = (E[\tau_i^2 g_i g_i'])^{-1} E[\tau_i G_i] \left( E[\tau_i G_i'] (E[\tau_i^2 g_i g_i'])^{-1} E[\tau_i G_i] \right)^{-1} = B'_{12}. \quad (33)$$

Substituting this expression for  $B_{21}$  into Equation (34) yields

$$\Sigma_k = \left( E[\tau_i G_i'] (E[\tau_i^2 g_i g_i'])^{-1} E[\tau_i G_i] \right)^{-1} + \rho_k. \quad (34)$$

We will now show that  $\Sigma_k$  diverges as  $k \rightarrow \infty$ . For EL,  $\lambda_k^*$  is such that  $E[g(x_i, \theta_k^*) / (1 - \lambda_k^* g(x_i, \theta_k^*))] = 0$ . Since  $E[g(x_i, \theta_k^*) / (1 - \lambda_k^* g(x_i, \theta_k^*))] = E[g(x_i, \theta_k^*)] + E[g(x_i, \theta_k^*) g'(x_i, \theta_k^*) / (1 - \lambda_k^* g(x_i, \theta_k^*)) \lambda_k^*]$ , we have

$$\Omega_k \lambda_k^* = -E[g(x_i, \theta_k^*)] \quad (35)$$

where  $\Omega_k = E[g(x_i, \theta_k^*) g'(x_i, \theta_k^*) / (1 - \lambda_k^* g(x_i, \theta_k^*))]$ . Since  $\inf_{k \geq \bar{k}} E[g(x_i, \theta_k^*)] > 0$  for some  $\bar{k} \in \mathbb{N}$  by construction, having  $\lambda_k^* \rightarrow 0$  as  $k \rightarrow \infty$  is only possible if at least one of the eigenvalues of  $\Omega_k$  diverges as  $k \rightarrow \infty$ . Let  $v$  be a (unit) eigenvector associated with one of these eigenvalues. Then, by the CSI,  $v' \Omega_k v$  equals

$$E \left[ \frac{v' g(x_i, \theta_k^*)}{(1 - \lambda_k^* g(x_i, \theta_k^*))} v' g(x_i, \theta_k^*) \right] \leq \left( E \left[ \frac{(v' g(x_i, \theta_k^*))^2}{(1 - \lambda_k^* g(x_i, \theta_k^*))^2} \right] E \left[ (v' g(x_i, \theta_k^*))^2 \right] \right)^{1/2} \quad (36)$$

Since  $E \left[ (v'g(x_i, \theta_k^*))^2 \right] \leq \sup_{\theta \in \Theta} E \left[ \|g(x_i, \theta)\|^2 \right] < \infty$ , Equation (36) therefore implies that  $E \left[ \frac{(v'g(x_i, \theta_k^*))^2}{(1 - \lambda_k^{*'} g(x_i, \theta_k^*))^2} \right] = E \left[ \tau_i^2 v' g_i g_i' v \right]$  diverges and thus that  $E \left[ \tau_i^2 g_i g_i' \right]$  has a divergent eigenvalue. Since  $E \left[ \tau_i^2 g_i g_i' \right]$  enters the expression of  $\Sigma_k$  (given by Equation (34)),  $\Sigma_k$  has at least one divergent eigenvalue as  $k \rightarrow \infty$ . Note that the other terms entering the expression of  $\Sigma_k$  cannot compensate for the explosive behavior of  $E \left[ \tau_i^2 g_i g_i' \right]$ , since a simple application of the CSI shows that, as  $k \rightarrow \infty$ ,  $\|E \left[ \tau_i G_i \right]\| = \|E \left[ (1 + \tau_i \lambda_k^{*'} g_i) G_i \right]\| = O \left( E \left[ \tau_i \|g_i\| \|\lambda_k^*\| \|G_i\| \right] \right) = O \left( \left( E \left[ \tau_i^2 \|g_i\|^2 \right] \right)^{1/2} \right) \left( E \left[ \|G_i\|^2 \right] \right)^{1/2} \|\lambda_k^*\| = o \left( \left( E \left[ \tau_i^2 \|g_i\|^2 \right] \right)^{1/2} \right) = o \left( \left( E \left[ \tau_i^2 \|g_i g_i'\| \right] \right)^{1/2} \right) = o \left( \left( E \left[ \tau_i^2 v' g_i g_i' v \right] \right)^{1/2} \right)$ .

We will now show that the divergent behavior of  $\Sigma_k$  implies that EL is not root  $n$  consistent. We start by calculating the probability that  $\hat{\theta}_{k,n}$  lies outside of a root  $n$  neighborhood of the pseudo-true value  $\theta_k^*$ . Let  $P_{k,n}$  be the finite sample distribution of  $n^{1/2} (u' \Sigma_k u)^{-1/2} u' \left( \hat{\theta}_{k,n} - \theta_k^* \right)$  for some conformable unit vector  $u$  such that  $u' \Sigma_k u \rightarrow \infty$  as  $k \rightarrow \infty$  ( $u$  is an eigenvector associated with one of the divergent eigenvalues of  $\Sigma_k$ ). Let  $P_{k,\infty}$  denote the corresponding asymptotic distribution, the cdf of a  $N(0, 1)$  for all  $k$ . For a given  $\xi < 0$ , the probability that  $u' \left( \hat{\theta}_{k,n} - \theta_k^* \right) \leq n^{-1/2} \xi$  is  $P_{k,n} \left( (u' \Sigma_k u)^{-1/2} \xi \right)$ .

Let  $n_k = \min \left\{ n : \sup_{m \geq n} \left| P_{k,m} \left( (u' \Sigma_k u)^{-1/2} \xi \right) - P_{k,\infty} \left( (u' \Sigma_k u)^{-1/2} \xi \right) \right| \leq k^{-1} \right\}$ .

This defines the sample size beyond which the difference (at  $(u' \Sigma_k u)^{-1/2} \xi$ ) between the finite-sample and asymptotic distribution is less than  $k^{-1}$ . Such a finite  $n$  can always be found, since  $P_{k,n}$  converges pointwise to  $P_{k,\infty}$ . Now define the ‘‘inverse’’ sequence  $k_n = \max \{ k : n_k \leq n \}$ . Note that  $k_n \rightarrow \infty$ , as  $n \rightarrow \infty$ , since  $n_k \rightarrow \infty$  as  $k \rightarrow \infty$ .

Since  $P \left[ \left| u' \left( \hat{\theta}_{\infty,n} - \theta_\infty^* \right) \right| \geq \varepsilon \right] \geq P \left[ \left| u' \left( \hat{\theta}_{k_n,n} - \theta_{k_n}^* \right) \right| \geq \varepsilon \right]$  for all  $\varepsilon > 0$  and any  $n$ , by the construction of  $F_k$ ,  $P \left[ u' \left( \hat{\theta}_{\infty,n} - \theta_\infty^* \right) \leq n^{-1/2} \xi \right] \geq P \left[ u' \left( \hat{\theta}_{k_n,n} - \theta_{k_n}^* \right) \leq n^{-1/2} \xi \right] =$

$P_{k,n} \left( (u' \Sigma_k u)^{-1/2} \xi \right)$  for any  $k$  and any  $n$  and any  $\xi < 0$ . In particular, for  $k = k_n$ ,

$$P \left[ u' \left( \hat{\theta}_{\infty,n} - \theta_{\infty}^* \right) \leq n^{-1/2} \xi \right] \geq P_{k_n,n} \left( (u' \Sigma_{k_n} u)^{-1/2} \xi \right) \quad (37)$$

$$\begin{aligned} &= P_{k_n,\infty} \left( (u' \Sigma_{k_n} u)^{-1/2} \xi \right) + \left( P_{k_n,n} \left( (u' \Sigma_{k_n} u)^{-1/2} \xi \right) - P_{k_n,\infty} \left( (u' \Sigma_{k_n} u)^{-1/2} \xi \right) \right) \\ &\geq P_{k_n,\infty} \left( (u' \Sigma_{k_n} u)^{-1/2} \xi \right) - k_n^{-1} \end{aligned} \quad (39)$$

by the definition of  $k_n$ . As  $n \rightarrow \infty$ ,  $k_n^{-1} \rightarrow 0$ . Since  $P_{k,\infty}$  is the same for all  $k$  and is continuous (it is the cdf of a  $N(0, 1)$ ), for any  $k$  we have  $\lim_{n \rightarrow \infty} P_{k_n,\infty} \left( (u' \Sigma_{k_n} u)^{-1/2} \xi \right) = \lim_{n \rightarrow \infty} P_{k,\infty} \left( (u' \Sigma_{k_n} u)^{-1/2} \xi \right) = P_{k,\infty} \left( \lim_{n \rightarrow \infty} (u' \Sigma_{k_n} u)^{-1/2} \xi \right) = P_{k,\infty}(0) = 1/2$ , where we have used the fact that  $(u' \Sigma_{k_n} u)^{-1/2} \rightarrow 0$  since  $u' \Sigma_k u$  diverges as  $k \rightarrow \infty$ . We then have,  $\lim_{n \rightarrow \infty} P \left[ u' \left( \hat{\theta}_{\infty,n} - \theta_{\infty}^* \right) \leq n^{-1/2} \xi \right] \geq 1/2$  for any  $\xi < 0$ . A similar reasoning for  $\xi > 0$  implies that  $\lim_{n \rightarrow \infty} P \left[ u' \left( \hat{\theta}_{\infty,n} - \theta_{\infty}^* \right) \geq n^{-1/2} \xi \right] \geq 1/2$ . It follows that  $\hat{\theta}_{\infty,n}$  lies outside a  $n^{-1/2}$  neighborhood of  $\theta_{\infty}^*$  with probability approaching  $1/2 + 1/2 = 1$  as  $n \rightarrow \infty$ , thus ruling out root  $n$  convergence.

To summarize, for any EL estimator  $\hat{\theta}_{\infty,n}$  based on a distribution  $F_{\infty}(x)$  with unbounded support, there exists a family of other estimators  $\hat{\theta}_{k,n}$  based on compactly supported distributions  $F_k(x)$  having all a narrower distribution than EL for each  $n$ . Yet the asymptotic variance of  $\hat{\theta}_{k,n}$  diverges as  $k \rightarrow \infty$ . By a standard diagonal argument, there exists an estimator sequence  $\hat{\theta}_{k_n,n}$  that is not root  $n$  consistent but whose distribution is narrower than the one of EL at each  $n$ . Hence EL is not root  $n$  consistent. ■

### Proof of Theorem 2.

$$\begin{aligned} \ln \hat{L} &\equiv n^{-1} \sum_i \ln n \hat{w}_i = n^{-1} \sum_i \ln \left( \exp \left( \hat{\lambda}' g_i \right) / \left( n^{-1} \sum_j \exp \left( \hat{\lambda}' g_j \right) \right) \right) \\ &= n^{-1} \sum_i \hat{\lambda}' g_i - \ln \left( n^{-1} \sum_j \exp \left( \hat{\lambda}' g_j \right) \right) = -\ln \left( n^{-1} \sum_j \exp \left( \hat{\lambda}' (g_j - \hat{g}) \right) \right) \end{aligned}$$

$$\begin{aligned}
\frac{d \ln \hat{L}}{d\theta'} &= n^{-1} \sum_i \frac{d(\hat{\lambda}' g_i)}{d\theta'} - \left( n^{-1} \sum_j \exp(\hat{\lambda}' g_j) \right)^{-1} n^{-1} \sum_i \exp(\hat{\lambda}' g_i) \frac{d(\hat{\lambda}' g_i)}{d\theta'} \\
&= n^{-1} \sum_i \frac{d(\hat{\lambda}' g_i)}{d\theta'} - n^{-1} \sum_i n \hat{w}_i \frac{d(\hat{\lambda}' g_i)}{d\theta'} = n^{-1} \sum_i (1 - n \hat{w}_i) \frac{d(\hat{\lambda}' g_i)}{d\theta'}
\end{aligned}$$

From Equation (15), the first order condition for  $\hat{\lambda}$  is  $\sum_i g_i \exp(\hat{\lambda}' g_i) = 0$ . ■

**Proof of Theorem 3.** Expanding the ETEL first-order conditions for  $\hat{\theta}$  and  $\hat{\lambda}$  around  $\theta = \theta^*$  and  $\lambda = 0$  reveals an expansion identical to the one of EL at least up to  $O_p(n^{-1/2})$  in an  $O(n^{-1/2})$  neighborhood of  $\theta = \theta^*$  and  $\lambda = 0$ :

$$n^{-1} \sum_i \begin{bmatrix} 0 \\ g(x_i, \theta^*) \end{bmatrix} + n^{-1} \sum_i \begin{pmatrix} 0 & G'(x_i, \theta^*) \\ G(x_i, \theta^*) & g(x_i, \theta^*) g'(x_i, \theta^*) \end{pmatrix} \begin{bmatrix} \theta - \theta^* \\ \lambda \end{bmatrix} = o_p(n^{-1/2}).$$

Calculational details can be found in [57]. In addition, in a  $O(n^{-1/2})$  neighborhood of  $\theta = \theta^*$ , both the ETEL and the EL objective functions for  $\hat{\theta}$  share the same expansion in  $(\theta - \theta^*)$  at least up to  $O_p(n^{-1})$ :

$$-\frac{1}{2} (\theta - \theta^*)' \left( n^{-1} \sum_i G'(x_i, \theta^*) \right) \left( \hat{\Omega}^* \right)^{-1} \left( n^{-1} \sum_i G(x_i, \theta^*) \right) (\theta - \theta^*) + o_p(n^{-1})$$

where  $\hat{\Omega}^* = n^{-1} \sum_i g(x_i, \theta^*) g'(x_i, \theta^*)$ . It is known (see, for instance, [47, 49]) that the EL estimator is asymptotically such that the solutions  $\hat{\lambda}_{EL}$  and  $\hat{\theta}_{EL}$  lie within the  $O(n^{-1/2})$  neighborhood where the remainder terms of these expansions are negligible. Hence, asymptotically,  $\hat{\lambda}_{EL}$  and  $\hat{\theta}_{EL}$  also solve the ETEL first-order conditions, apart from negligible remainders. Since  $\hat{\lambda}_{EL} \xrightarrow{p} 0$ , and since both the EL and the ETEL objective functions for  $\hat{\theta}$  converge to their maximum possible value when  $\hat{\lambda}_{EL} \xrightarrow{p} 0$  and  $\hat{\lambda}_{ETEL} \xrightarrow{p} 0$ , respectively, the existence of another solution outside of the neighborhood of validity of the above expansions can be ruled out. ETEL thus inherits all the first-order properties of EL established in [47, 49]. ■

**Proof of Theorem 4.** The first conclusion follows from the fact that the

implied probabilities are given by

$$\hat{w}_i(\theta) = \exp\left(\hat{\lambda}(\theta)' g(x_i, \theta)\right) / \left(\sum_j \exp\left(\hat{\lambda}(\theta)' g(x_j, \theta)\right)\right),$$

a necessarily positive quantity for any  $\hat{\lambda}$  and  $\theta$ . The second conclusion holds for any estimator where  $\theta$  extremizes a differentiable objective function:

$$\frac{\partial \ln \hat{L}(T(\beta))}{\partial \beta} = \frac{\partial T(\beta)'}{\partial \beta} \frac{\partial \ln \hat{L}(\theta)}{\partial \theta} = 0$$

if and only if  $\partial \ln \hat{L}(\theta) / \partial \theta = 0$  since  $\partial T(\beta)' / \partial \beta$  has full rank ( $T(\beta)$  being one-to-one). The third conclusion can be shown by noting that any invertible linear transformation of the moment function  $A(\theta) g(x_i, \theta)$  simply causes the Lagrange multiplier  $\hat{\lambda}(\theta)$  to become  $\left((A(\theta))^{-1}\right)' \lambda(\theta)$ . Indeed, under these two transformations, the first-order conditions for both  $\hat{\theta}$  and  $\hat{\lambda}(\hat{\theta})$  remain satisfied:

$$n^{-1} \sum_i \left(1 - n \hat{w}_i(\hat{\theta})\right) d\left(\hat{\lambda}'(A(\theta))^{-1} A(\theta) g_i\right) / d\theta = n^{-1} \sum_i \left(1 - n \hat{w}_i(\hat{\theta})\right) d\left(\hat{\lambda}' g_i\right) / d\theta = 0$$

where  $\hat{w}_i(\hat{\theta}) = \exp\left(\hat{\lambda}'(A(\theta))^{-1} A(\theta) g_i\right) / \left(\sum_j \exp\left(\hat{\lambda}'(A(\theta))^{-1} A(\theta) g_j\right)\right) = \exp\left(\hat{\lambda}' g_i\right) / \left(\sum_j \exp\left(\hat{\lambda}' g_j\right)\right)$  and  $n^{-1} \sum_i \exp\left(\hat{\lambda}'(A(\theta))^{-1} A(\theta) g_i\right) A(\theta) g_i = A(\theta) n^{-1} \sum_i \exp\left(\hat{\lambda}' g_i\right) g_i = 0$  if and only if  $n^{-1} \sum_i \exp\left(\hat{\lambda}' g_i\right) g_i = 0$  since  $A(\theta)$  is invertible. ■

**Proof of Theorem 5.** By Theorem 2, the first order condition for  $\hat{\theta}$  is

$$\begin{aligned} \frac{d \ln \hat{L}}{d\theta'} &= n^{-1} \sum_i (1 - n \hat{w}_i) \left( g_i' \frac{\partial \hat{\lambda}}{\partial \theta'} + \hat{\lambda}' G_i \right) \\ &= n^{-1} \sum_i g_i' \frac{\partial \hat{\lambda}}{\partial \theta'} + \hat{\lambda}' n^{-1} \sum_i G_i - \left( \sum_i \hat{w}_i g_i' \right) \frac{\partial \hat{\lambda}}{\partial \theta'} - \hat{\lambda}' \sum_i \hat{w}_i G_i \\ &= \hat{g}' \frac{\partial \hat{\lambda}}{\partial \theta'} + \hat{\lambda}' \tilde{G} - 0 - \hat{\lambda}' \tilde{G} \end{aligned} \tag{40}$$

To find  $\partial \hat{\lambda} / \partial \theta'$ , we note that a total differential of  $\sum_i \exp\left(\hat{\lambda}' g_i\right) g_i = 0$  yields

$$\begin{aligned} \sum_i \exp\left(\hat{\lambda}' g_i\right) g_i g_i' d\hat{\lambda} + \sum_i \exp\left(\hat{\lambda}' g_i\right) G_i d\theta + \sum_i g_i \exp\left(\hat{\lambda}' g_i\right) \hat{\lambda}' G_i d\theta &= 0 \\ \sum_i g_i g_i' \hat{w}_i d\hat{\lambda} + \sum_i \hat{w}_i \left( I + g_i \hat{\lambda}' \right) G_i d\theta &= 0 \end{aligned}$$

implying that

$$\frac{\partial \hat{\lambda}}{\partial \theta'} = -\tilde{\Omega}^{-1} \left( \sum_i \hat{w}_i (I + g_i \hat{\lambda}') G_i \right). \quad (41)$$

Substituting this result into Equation (40) gives

$$\begin{aligned} \frac{\partial \ln \hat{L}(\theta)}{\partial \theta'} &= -\hat{g}' \tilde{\Omega}^{-1} \left( \sum_i \hat{w}_i (I + g_i \hat{\lambda}') G_i \right) + \hat{\lambda}' \hat{G} - \hat{\lambda}' \tilde{G} \\ &= -\hat{g}' \tilde{\Omega}^{-1} \tilde{G} - \hat{g}' \tilde{\Omega}^{-1} \sum_i \hat{w}_i g_i \hat{\lambda}' G_i + \hat{\lambda}' \hat{G} - \hat{\lambda}' \tilde{G} \\ &= -\hat{g}' \tilde{\Omega}^{-1} \tilde{G} - \hat{g}' \tilde{\Omega}^{-1} \sum_i \hat{w}_i g_i \hat{\lambda}' G_i + n^{-1} \sum_i (1 - n \hat{w}_i) \hat{\lambda}' G_i. \end{aligned} \quad (42)$$

By the first-order equivalence between EL and ETEL established in Theorem 3 and using Theorem 3.1 in [47],  $\hat{\lambda}(\hat{\theta}) = O_p(n^{-1/2})$  and  $\hat{g} = O_p(n^{-1/2})$  for  $\hat{\theta}$  such that  $\|\theta - \theta^*\| = O_p(n^{-1/2})$ . These facts, along with the fact that  $\sup_{\theta \in \Theta} \max_{i \leq n} \|g_i\| = o_p(n^{1/2})$  by Assumption 1.4, provide us with asymptotic expansions for  $n \hat{w}_i$  and  $\hat{\lambda}$ :

$$\begin{aligned} n \hat{w}_i &= \frac{\exp(\hat{\lambda}' g_i)}{n^{-1} \sum_j \exp(\hat{\lambda}' g_j)} = \frac{1 + \hat{\lambda}' g_i + O\left(\left(\hat{\lambda}' g_i\right)^2\right)}{1 + \hat{\lambda}' \hat{g} + O_p(n^{-1})} \\ &= \frac{1 + \hat{\lambda}' g_i + O_p(n^{-1}) \|g_i\|^2}{1 + O_p(n^{-1}) + O_p(n^{-1})} = 1 + \hat{\lambda}' g_i + O_p(n^{-1}) \|g_i\|^2 \end{aligned} \quad (43)$$

An expansion for  $\hat{\lambda}$  is obtained by noting that the left-hand side of  $n^{-1} \sum_i g_i \exp(g_i' \hat{\lambda}) = 0$  can be written as

$$\begin{aligned} n^{-1} \sum_i g_i (1 + g_i' \hat{\lambda}) + R_0 &= n^{-1} \sum_i g_i + \left( n^{-1} \sum_i g_i g_i' \right) \hat{\lambda} + R_0 \\ &= n^{-1} \sum_i g_i + \left( n^{-1} \sum_i n \hat{w}_i g_i g_i' \right) \hat{\lambda} + R_0 + R_1 = \hat{g} + \tilde{\Omega} \hat{\lambda} + R_0 + R_1, \end{aligned}$$

implying that

$$\hat{\lambda} = -\tilde{\Omega}^{-1} \hat{g} - \tilde{\Omega}^{-1} (R_0 + R_1), \quad (44)$$

where the remainder terms  $R_0, R_1$  can be bounded using the assumption  $E \left[ \sup_{\theta \in \mathcal{N}} \|g_i\|^4 \right] < \infty$  and Equation (43):  $\|R_0\| = O_p(n^{-1}) n^{-1} \sum_i \|g_i\|^3 = O_p(n^{-1})$  and  $\|R_1\| \leq$

$$n^{-1} \sum_i (n\hat{w}_i - 1) \|g_i\|^2 \|\hat{\lambda}\| = n^{-1} \sum_i O\left(\|\hat{\lambda}\| \|g_i\|\right) \|g_i\|^2 \|\hat{\lambda}\| = O\left(\|\hat{\lambda}\|^2\right) n^{-1} \sum_i \|g_i\|^3 = O_p(n^{-1}).$$

Substituting expansion (43) into the last term of Equation (42) yields

$$\frac{\partial \ln \hat{L}(\theta)}{\partial \theta'} = -\hat{g}' \tilde{\Omega}^{-1} \tilde{G} - \hat{g}' \tilde{\Omega}^{-1} \sum_i \hat{w}_i g_i \hat{\lambda}' G_i + n^{-1} \sum_i \hat{\lambda}' g_i \hat{\lambda}' G_i + R_2 \quad (45)$$

where  $\|R_2\| \leq O_p(n^{-1}) n^{-1} \sum_i \|g_i\|^2 \|\hat{\lambda}\| \|G_i\| \leq O_p(n^{-3/2}) n^{-1} \sum_i \|g_i\|^2 \|G_i\| \leq O_p(n^{-3/2}) \left(n^{-1} \sum_i \|g_i\|^4\right)^{1/2} \left(n^{-1} \sum_i \|G_i\|^2\right)^{1/2} = O_p(n^{-3/2})$ , after using the CSI and the fact that  $E\left[\sup_{\theta \in \mathcal{N}} \|g_i\|^4\right] < \infty$  and  $E\left[\sup_{\theta \in \mathcal{N}} \|G_i\|^2\right] < \infty$ . Equation (45) then becomes

$$\frac{\partial \ln \hat{L}(\theta)}{\partial \theta'} = -\hat{g}' \tilde{\Omega}^{-1} \tilde{G} - \hat{g}' \tilde{\Omega}^{-1} \sum_j \hat{w}_j g_j \hat{\lambda}' G_j - \left(\hat{\lambda}\right)' n^{-1} \sum_j g_j \hat{\lambda}' G_j + O_p(n^{-3/2})$$

where the parenthesized  $\hat{\lambda}$  can be replaced by expansion (44)

$$\frac{\partial \ln \hat{L}(\theta)}{\partial \theta'} = -\hat{g}' \tilde{\Omega}^{-1} \tilde{G} - \hat{g}' \tilde{\Omega}^{-1} \sum_j \hat{w}_j g_j \hat{\lambda}' G_j + \hat{g}' \tilde{\Omega}^{-1} n^{-1} \sum_j g_j \hat{\lambda}' G_j + R_3 + O_p(n^{-3/2}) \quad (46)$$

where  $\|R_3\| = O_p(n^{-1}) n^{-1} \sum_j \|g_j\| \|\hat{\lambda}\| \|G_j\| = O_p(n^{-3/2}) n^{-1} \sum_j \|g_j\| \|G_j\| \leq O_p(n^{-3/2}) \left(n^{-1} \sum_j \|g_j\|^2\right)^{1/2} \left(n^{-1} \sum_j \|G_j\|^2\right)^{1/2} = O_p(n^{-3/2})$  by the CSI,  $E\left[\sup_{\theta \in \mathcal{N}} \|g_i\|^4\right] < \infty$  and  $E\left[\sup_{\theta \in \mathcal{N}} \|G_i\|^2\right] < \infty$ . Equation (46) then becomes

$$\begin{aligned} \frac{\partial \ln \hat{L}(\theta)}{\partial \theta'} &= -\hat{g}' \tilde{\Omega}^{-1} \tilde{G} + \hat{g}' \tilde{\Omega}^{-1} n^{-1} \sum_j (1 - n\hat{w}_j) g_j \hat{\lambda}' G_j + O_p(n^{-3/2}) \\ &= -\hat{g}' \tilde{\Omega}^{-1} \tilde{G} - \hat{g}' \tilde{\Omega}^{-1} n^{-1} \sum_j \left(\hat{\lambda}' g_j\right) g_j \hat{\lambda}' G_j + R_4 + O_p(n^{-3/2}) \end{aligned} \quad (47)$$

where we have used expansion (43) again and where  $\|R_4\| \leq \left\|\hat{g}' \tilde{\Omega}^{-1}\right\| n^{-1} \sum_j O((\hat{\lambda}' g_j)^2) \|g_j\| \|\hat{\lambda}\| \|G_j\| = \|\hat{g}\| \|\hat{\lambda}\|^3 \left\|\tilde{\Omega}^{-1}\right\| n^{-1} \sum_j \|g_j\|^2 \|g_j\| \|G_j\| \leq O_p(n^{-2}) (\max_{i \leq n} \|g_i\|) n^{-1} \sum_j \|g_j\|^2 \|G_j\| = O_p(n^{-2}) O_p(n^{1/2}) n^{-1} \sum_j \|g_j\|^2 \|G_j\| = O_p(n^{-3/2})$  by the CSI, the assumptions that  $E\left[\sup_{\theta \in \mathcal{N}} \|g_i\|^4\right] < \infty$ ,  $E\left[\sup_{\theta \in \mathcal{N}} \|G_i\|^2\right] < \infty$  and the fact

that  $E \left[ \|g_i\|^2 \right] < \infty \Rightarrow \max_{i \leq n} \|g_i\| = O_p(n^{1/2})$  (as in [47], Lemma A1). Equation (47) finally becomes

$$\frac{\partial \ln \hat{L}(\theta)}{\partial \theta'} = -\hat{g}' \tilde{\Omega}^{-1} \tilde{G} + O_p(n^{-3/2}).$$

Now, the term  $\hat{g}' \tilde{\Omega}^{-1} \tilde{G}$  is similar to the first-order conditions for EL, except that the weights used in  $\tilde{\Omega}$  and  $\tilde{G}$  are the ET rather than the EL weights. However, by Equation (43) and a similar expansion for the EL weights,  $n(\hat{w}_{i,ET} - \hat{w}_{i,EL}) = O_p(n^{-1}) \|g_i\|^2$ . That fact, along with  $\hat{g} = O_p(n^{-1/2})$ , implies that

$$\begin{aligned} & \hat{g}' \left( \sum_i \hat{w}_{i,ET} g_i g_i' \right)^{-1} \left( \sum_i \hat{w}_{i,ET} G_i \right) \\ &= \hat{g}' \left( n^{-1} \sum_i n \hat{w}_{i,EL} g_i g_i' + R_5 \right)^{-1} \left( n^{-1} \sum_i n \hat{w}_{i,EL} G_i + R_6 \right) \\ &= \hat{g}' \left( \sum_i \hat{w}_{i,EL} g_i g_i' \right)^{-1} \left( \sum_i \hat{w}_{i,EL} G_i \right) + O_p(n^{-1/2}) O_p(n^{-1}) \end{aligned}$$

by the differentiability of the inverse and the fact that  $\|R_5\| \leq n^{-1} \sum_i O_p(n^{-1}) \|g_i\|^2 \|g_i\|^2 = O_p(n^{-1}) n^{-1} \sum_i \|g_i\|^4 = O_p(n^{-1})$  and  $\|R_6\| \leq n^{-1} \sum_i O_p(n^{-1}) \|g_i\|^2 \|G_i\| = O_p(n^{-1})$ .

This implies that the first-order condition for ETEL is the same as the one of EL up to  $O_p(n^{-3/2})$ . The continuous differentiability of  $\hat{g}$  in  $\theta$  implies  $\left\| \hat{\theta}_{EDEL} - \hat{\theta}_{EL} \right\| = O_p(n^{-3/2})$  by a standard expansion of the first-order condition around  $\theta = \theta^*$ . ■

**Proof of Theorem 7.** Lemma A4 in [47] establishes that under regularity conditions implied by the ones given in the statement of the present Theorem, a just-identified GMM estimator  $\hat{\beta}$  defined by  $n^{-1} \sum_i \phi(x_i, \hat{\beta}) = 0$  admits a stochastic expansion of the form

$$\hat{\beta}_l - \beta_l^* = n^{-1/2} \bar{\Psi}_l + n^{-1} \bar{Q}_l + n^{-3/2} \bar{R}_l + O_p(n^{-2}) \quad (48)$$

where

$$\bar{Q}_l = \sum_j \bar{\Psi}_{l,j} \bar{\Psi}_j + \frac{1}{2} \sum_{j,k} \bar{\Psi}_{l,jk} \bar{\Psi}_j \bar{\Psi}_k$$

$$\begin{aligned}
\bar{R}_l &= \sum_j \bar{\Psi}_{l,j} \bar{Q}_j + \sum_{j,k} \Psi_{l,jk} \bar{\Psi}_j \bar{Q}_k + \frac{1}{2} \sum_{j,k} \bar{\Psi}_{l,jk} \bar{\Psi}_j \bar{\Psi}_k + \frac{1}{6} \sum_{j,k,h} \Psi_{l,jkh} \bar{\Psi}_j \bar{\Psi}_k \bar{\Psi}_h \\
\bar{\Psi}_l &= \sum_q \Phi_{lq}^{-1} \bar{\Phi}_q & \bar{\Psi}_{l,j} &= \sum_q \Phi_{lq}^{-1} \bar{\Phi}_{q,j} & \bar{\Psi}_{l,jk} &= \sum_q \Phi_{lq}^{-1} \bar{\Phi}_{q,jk} \\
\Psi_l &= \sum_q \Phi_{lq}^{-1} \Phi_q & \Psi_{l,j} &= \sum_q \Phi_{lq}^{-1} \Phi_{q,j} & \Psi_{l,jk} &= \sum_q \Phi_{lq}^{-1} \Phi_{q,jk} & \Psi_{l,jkh} &= \sum_q \Phi_{lq}^{-1} \Phi_{q,jkh} \\
\Phi^{-1} &= \left( E \left[ \partial \phi(x_i, \beta) / \partial \beta' \Big|_{\beta=\beta^*} \right] \right)^{-1} \\
\Phi_{l,j} &= E \left[ \frac{\partial \phi_l(x_i, \beta)}{\partial \beta_j} \Big|_{\beta=\beta^*} \right] & \Phi_{l,jk} &= \left[ \frac{\partial^2 \phi_l(x_i, \beta)}{\partial \beta_j \partial \beta_k} \Big|_{\beta=\beta^*} \right] & \Phi_{l,jkh} &= \left[ \frac{\partial^3 \phi_l(x_i, \beta)}{\partial \beta_j \partial \beta_k \partial \beta_h} \Big|_{\beta=\beta^*} \right] \\
\bar{\Phi}_l &= n^{-1/2} \sum_i \phi_l(x_i, \beta^*) & \bar{\Phi}_{l,j} &= n^{-1/2} \sum_i \left( \frac{\partial \phi_l(x_i, \beta)}{\partial \beta_j} \Big|_{\beta=\beta^*} - \Phi_{l,j} \right) \\
\bar{\Phi}_{l,jk} &= n^{-1/2} \sum_i \left( \frac{\partial^2 \phi_l(x_i, \beta)}{\partial \beta_j \partial \beta_k} \Big|_{\beta=\beta^*} - \Phi_{l,jk} \right)
\end{aligned}$$

(We have adapted Newey and Smith's result to follow our notation and slightly simplified it using the fact that  $\Psi_{l,jk} = \Psi_{l,kj}$ ). We now write the ETEL and EL estimators as just identified GMM estimators that can be easily compared. As shown in Lemma 9, and as discussed in Section 3.2.3 in the text, the ETEL estimator can be written as a subvector  $\hat{\theta}$  of an augmented parameter vector  $\hat{\beta} = (\hat{\tau}, \hat{\kappa}', \hat{\lambda}', \hat{\theta})'$  that solves a just-identified vector of moment conditions  $n^{-1} \sum_i \phi^{ETEL}(x_i, \hat{\beta}_{ETEL}) = 0$ , where  $\phi^{ETEL}(x_i, \hat{\beta})$  is given by Equation (23).

It is well-known that EL can also be written as a subvector  $\hat{\theta}$  of an augmented parameter vector  $(\hat{\kappa}', \hat{\theta})'$  that solves a just-identified vector of moment conditions

$$n^{-1} \sum_i \begin{bmatrix} \hat{\varepsilon}_i g_i \\ \hat{\varepsilon}_i G_i' \kappa \end{bmatrix} = 0 \quad (49)$$

where  $\hat{\varepsilon}_i = (1 - \hat{\kappa}' g_i)^{-1}$  and  $\hat{\kappa}$  is the Lagrange multiplier of the moment constraints, which has been relabelled  $\hat{\kappa}$  to simplify the comparison with ETEL. Once again, to further simplify the comparison, we augment the vector in Equation (49) by  $1 + \dim \kappa$  additional moment conditions and introduce the same number of additional parameters  $(\hat{\tau}, \hat{\lambda})$  where  $\tau \in \mathbb{R}$  and  $\lambda \in \mathbb{R}^{\dim \kappa}$ :

$$n^{-1} \sum_i \begin{bmatrix} (\hat{\tau}_i - \hat{\tau}) & \hat{\tau}_i g_i' & \hat{\varepsilon}_i g_i' & (\hat{\varepsilon}_i G_i' \hat{\kappa})' \end{bmatrix}' = 0$$

where  $\hat{\tau}_i = \exp(\hat{\lambda}' g_i)$ . In this fashion, the dimension of the vector of moment conditions and the number of parameters are the same in ETEL as in EL. The additional moment conditions merely define the values of the new parameters  $(\hat{\tau}, \hat{\lambda})$  and do not change the values of  $(\hat{\kappa}', \hat{\theta}')$ . Indeed, whenever  $(\hat{\kappa}', \hat{\theta}')$  are such that the bottom two subvectors are zero, one can always find a value of  $(\hat{\tau}, \hat{\lambda})$  that will make the top two subvectors vanish as well. (There exists  $\hat{\lambda}$  such that  $n^{-1} \sum_i \hat{\tau}_i g_i = 0$  w.p.a. 1. Then, we can just set  $\hat{\tau} = n^{-1} \sum_i \hat{\tau}_i$ .)

Finally, since just-identified GMM is invariant under linear transformations of the vector of moment conditions, the moment conditions for EL can equivalently be written as  $n^{-1} \sum_i \phi^{EL}(x_i, \hat{\beta}_{EL}) = 0$ , where

$$\phi^{EL}(x_i, \hat{\beta}) = \begin{bmatrix} \hat{\tau}_i - \tau \\ \hat{\tau}_i g_i \\ \hat{\varepsilon}_i g_i - \hat{\tau}_i g_i \\ \hat{\varepsilon}_i G_i' \hat{\kappa} \end{bmatrix}. \quad (50)$$

Equipped with Equations (23) and (50), we can construct a stochastic expansion of the form (48) for each estimator. The  $O(n^{-2})$  covariance between two elements of the parameter vector,  $\hat{\theta}_l$  and  $\hat{\theta}_m$ , is given by

$$W_{lm} \equiv \text{Covar} [\bar{Q}_{l_\theta+l}, \bar{Q}_{l_\theta+m}] + \text{Covar} [\bar{R}_{l_\theta+l}, \bar{\Psi}_{l_\theta+m}] + \text{Covar} [\bar{\Psi}_{l_\theta+l}, \bar{R}_{l_\theta+m}] \quad (51)$$

where  $l_\theta = 1 + 2 \dim \lambda$ . The quantities associated with each estimator will be distinguished by an ‘‘E TEL’’ or ‘‘EL’’ superscript.

We provide below the sequence of equalities that need to be established in order to show, as directly as possible, that ETEL and EL have the same  $O(n^{-2})$  variance. The tedious yet straightforward calculational details that prove each statement are omitted below but can be found in [57].

- 1)  $\bar{\Phi}_l^{E TEL} = \bar{\Phi}_l^{EL}$  and  $\Phi_{l,j}^{E TEL} = \Phi_{l,j}^{EL} \equiv \Phi_{l,j} \Rightarrow \bar{\Psi}_j^{E TEL} = \bar{\Psi}_j^{EL}$
- 2) (1)  $\Rightarrow \bar{Q}_l^{E TEL} - \bar{Q}_l^{EL} = \sum_j \left( \bar{\Psi}_{l,j}^{E TEL} - \bar{\Psi}_{l,j}^{EL} \right) \bar{\Psi}_j + \frac{1}{2} \sum_{j,k} \left( \Psi_{l,jk}^{E TEL} - \Psi_{l,jk}^{EL} \right) \bar{\Psi}_j \bar{\Psi}_k$

- 2a)  $(\bar{\Psi}_{l,j}^{E TEL} - \bar{\Psi}_{l,j}^{EL}) \bar{\Psi}_j = \sum_{q,j} \Phi_{l_q}^{-1} (\bar{\Phi}_{q,j}^{E TEL} - \bar{\Phi}_{q,j}^{EL}) \bar{\Psi}_j$  where  $\sum_j (\bar{\Phi}_{q,j}^{E TEL} - \bar{\Phi}_{q,j}^{EL}) \bar{\Psi}_j = 0$
- 2b)  $(\Psi_{l,jk}^{E TEL} - \Psi_{l,jk}^{EL}) \bar{\Psi}_j \bar{\Psi}_k = \sum_{q,j,k} \Phi_{l_q}^{-1} (\Phi_{q,jk}^{E TEL} - \Phi_{q,jk}^{EL}) \bar{\Psi}_j \bar{\Psi}_k$   
 where  $\sum_{j,k} (\Phi_{q,jk}^{E TEL} - \Phi_{q,jk}^{EL}) \bar{\Psi}_j \bar{\Psi}_k = 0$
- 3) (2), (2a) and (2b)  $\Rightarrow \bar{Q}_l^{E TEL} - \bar{Q}_l^{EL} = 0$
- 4) (1) and (3)  $\Rightarrow W_{lm}^{E TEL} - W_{lm}^{EL} = \text{Covar} [\bar{R}_{l_{\theta+l}}^{E TEL} - \bar{R}_{l_{\theta+l}}^{EL}, \bar{\Psi}_{l_{\theta+m}}] +$   
 $+\text{Covar} [\bar{\Psi}_{l_{\theta+l}}, \bar{R}_{l_{\theta+m}}^{E TEL} - \bar{R}_{l_{\theta+m}}^{EL}]$
- 5) (1) and (3)  $\Rightarrow \bar{R}_l^{E TEL} - \bar{R}_l^{EL} = \sum_j (\bar{\Psi}_{l,j}^{E TEL} - \bar{\Psi}_{l,j}^{EL}) \bar{Q}_{,j} + \sum_{j,k} (\Psi_{l,jk}^{E TEL} - \Psi_{l,jk}^{EL}) \bar{\Psi}_j \bar{Q}_k +$   
 $+\frac{1}{2} \sum_{j,k} (\bar{\Psi}_{l,jk}^{E TEL} - \bar{\Psi}_{l,jk}^{EL}) \bar{\Psi}_j \bar{\Psi}_k + \frac{1}{6} \sum_{j,k,h} (\Psi_{l,jkh}^{E TEL} - \Psi_{l,jkh}^{EL}) \bar{\Psi}_j \bar{\Psi}_k \bar{\Psi}_h$
- 5a)  $\sum_j (\bar{\Psi}_{l_{\theta+l},j}^{E TEL} - \bar{\Psi}_{l_{\theta+l},j}^{EL}) \bar{Q}_{,j} = \frac{1}{2} \sum_j H_{l_j} \bar{g}_j \bar{g}' P \bar{g}$  where  
 $\bar{g} = n^{-1/2} \sum_i g_i$ ,  $H = (G' \Omega^{-1} G)^{-1} G' \Omega^{-1}$  and  $P = \Omega^{-1} - \Omega^{-1} G (G' \Omega^{-1} G)^{-1} G' \Omega^{-1}$
- 5b)  $(\Psi_{l_{\theta+l},jk}^{E TEL} - \Psi_{l_{\theta+l},jk}^{EL}) \bar{\Psi}_k \bar{Q}_j = -\frac{1}{2} \sum_j H_{l_j} \bar{g}_j \bar{g}' P \bar{g} + \Xi_{1,l}$  with  $E [\Xi_l \bar{\Psi}_{l_{\theta+m}}] = o(n^{-2})$
- 5c)  $\bar{\Psi}_j (\bar{\Psi}_{l,jk}^{E TEL} - \bar{\Psi}_{l,jk}^{EL}) \bar{\Psi}_k = 0$
- 5d)  $E [(\Psi_{l_{\theta+l},jkh}^{E TEL} - \Psi_{l_{\theta+l},jkh}^{EL}) \bar{\Psi}_j \bar{\Psi}_k \bar{\Psi}_h \bar{\Psi}_{l_{\theta+m}}^{EL}] = o(n^{-2})$ .
- 6) (5a) through (5d)  $\Rightarrow \text{Covar} [\bar{R}_{l_{\theta+l}}^{E TEL} - \bar{R}_{l_{\theta+l}}^{EL}, \bar{\Psi}_{l_{\theta+m}}] = o(n^{-2})$ .
- 7) (3) and (6)  $\Rightarrow$  ETEL and EL share the same  $O(n^{-2})$  variance. ■

**Proof of Theorem 8.** Let  $g_{i,a}$  and  $g_{i,b}$  denote the subvectors of  $g_i$  that are mutually independent and let  $\lambda_a$  and  $\lambda_b$  denote the corresponding subvectors of the Lagrange multiplier. Independence holds if and only if for any measurable functions  $a(g_{i,a})$  and  $b(g_{i,b})$ ,  $E[a(g_{i,a}) b(g_{i,b})] = E[a(g_{i,a})] E[b(g_{i,b})]$  whenever these expectations are defined. The exponentially tilted empirical distribution estimates the

moment  $E[a(g_{i,a})]$  by

$$\begin{aligned}
\hat{Q}_a &= \left( n^{-1} \sum_j \exp(\hat{\lambda}' g_j) \right)^{-1} n^{-1} \sum_i a(g_{i,a}) \exp(\hat{\lambda}' g_i) \xrightarrow{p} (E[\exp(\lambda' g_j)])^{-1} \times \\
&\quad \times E[a(g_{i,a}) \exp(\lambda' g_i)] \\
&= (E[\exp(\lambda'_a g_{i,a}) \exp(\lambda'_b g_{i,b})])^{-1} E[a(g_{i,a}) \exp(\lambda'_a g_{i,a}) \exp(\lambda'_b g_{i,b})] \\
&= \frac{E[a(g_{i,a}) \exp(\lambda'_a g_{i,a})] E[\exp(\lambda'_b g_{i,b})]}{E[\exp(\lambda'_a g_{i,a})] E[\exp(\lambda'_b g_{i,b})]} = \frac{E[a(g_{i,a}) \exp(\lambda'_a g_{i,a})]}{E[\exp(\lambda'_a g_{i,a})]} \equiv Q_a.
\end{aligned}$$

and similarly for  $E[b(g_{i,b})]$ . The exponentially tilted empirical distribution estimates the moment  $E[a(g_{i,a})b(g_{i,b})]$  by

$$\begin{aligned}
\hat{Q}_{ab} &= \left( n^{-1} \sum_j \exp(\hat{\lambda}' g_j) \right)^{-1} n^{-1} \sum_i a(g_{i,a}) b(g_{i,b}) \exp(\hat{\lambda}' g_i) \\
&\xrightarrow{p} (E[\exp(\lambda'_a g_{i,a}) \exp(\lambda'_b g_{i,b})])^{-1} E[a(g_{i,a}) \exp(\lambda'_a g_{i,a}) b(g_{i,b}) \exp(\lambda'_b g_{i,b})] \\
&= \frac{E[a(g_{i,a}) \exp(\lambda'_a g_{i,a})] E[b(g_{i,b}) \exp(\lambda'_b g_{i,b})]}{E[\exp(\lambda'_a g_{i,a})] E[\exp(\lambda'_b g_{i,b})]} = Q_a Q_b \equiv Q_{ab}
\end{aligned}$$

by the independence of  $g_{i,a}$  and  $g_{i,b}$  under the true untilted distribution. Hence  $\text{plim } \hat{Q}_{ab} = \text{plim } \hat{Q}_a \text{plim } \hat{Q}_b$  as claimed. A similar result does not hold for EL because  $(1 - \lambda' g_i)^{-1} \neq (1 - \lambda'_a g_{i,a})^{-1} (1 - \lambda'_b g_{i,b})^{-1}$ , unless  $\max_{i \leq n} |\lambda' g_i| \xrightarrow{p} 0$ , which is impossible under global misspecification. ■

**Proof of Lemma 9.** From Equation (42), the first-order condition for  $\hat{\theta}$  is

$$-\tilde{G}' \tilde{\Omega}^{-1} \hat{g} - \sum_i \hat{w}_i G'_i \hat{\lambda} g'_i \tilde{\Omega}^{-1} \hat{g} + n^{-1} \sum_i G'_i \hat{\lambda} - \sum_i \hat{w}_i G'_i \hat{\lambda} = 0. \quad (52)$$

(after transposition) where  $\hat{\lambda}$  satisfies

$$\sum_i \exp(\hat{\lambda}' g_i) g_i = 0. \quad (53)$$

Equation (52) contains products of sample moments which are difficult to analyze. Our goal is thus to define auxiliary parameters that will allow us to the rewrite this first-order conditions as a linear function of sample moments.

Let us introduce the quantity  $\hat{\tau}_i = \exp(\hat{\lambda}' g_i)$  and

$$\hat{\tau} = n^{-1} \sum_i \hat{\tau}_i. \quad (54)$$

Noting that  $\hat{w}_i = n^{-1} \hat{\tau}_i / \hat{\tau}$ , Equation (52) becomes,

$$\begin{aligned} & - \left( n^{-1} \sum_i \hat{\tau}_i G'_i \right) \left( n^{-1} \sum_i \hat{\tau}_i g_i g'_i \right)^{-1} \hat{g} - n^{-1} \sum_i \hat{\tau}_i G'_i \hat{\lambda} g'_i \left( n^{-1} \sum_i \hat{\tau}_i g_i g'_i \right)^{-1} \hat{g} + \\ & + n^{-1} \sum_i G'_i \hat{\lambda} - \frac{1}{\hat{\tau}} n^{-1} \sum_i \hat{\tau}_i G'_i \hat{\lambda} = 0. \end{aligned} \quad (55)$$

Now, we introduce  $\hat{\kappa} = - \left( n^{-1} \sum_i (\hat{\tau}_i / \hat{\tau}) g_i g'_i \right)^{-1} \hat{g}$ , or equivalently,

$$\left( n^{-1} \sum_i \hat{\tau}_i g_i g'_i \right) \hat{\kappa} + \hat{\tau} n^{-1} \sum_i g_i = 0. \quad (56)$$

Substituting the  $\hat{\kappa}$  whenever it appears in Equation (55), after multiplying through by  $\hat{\tau}$ , yields:

$$n^{-1} \sum_i \hat{\tau}_i G'_i \hat{\kappa} + n^{-1} \sum_i \hat{\tau}_i G'_i \hat{\lambda} g'_i \hat{\kappa} + n^{-1} \sum_i \hat{\tau} G'_i \hat{\lambda} - n^{-1} \sum_i \hat{\tau}_i G'_i \hat{\lambda} = 0. \quad (57)$$

Equation (57) is now linear in the sample moments. Equations (54), (53), (56) and, (57) can be collected into a single vector of moment conditions  $n^{-1} \sum_i \phi(x_i, \hat{\beta}) = 0$  where  $\hat{\beta} = (\hat{\tau}, \hat{\kappa}', \hat{\lambda}', \hat{\theta}')$  and

$$\phi(x_i, \hat{\beta}) = \begin{bmatrix} \hat{\tau}_i - \hat{\tau} \\ \hat{\tau}_i g_i \\ (\hat{\tau} - \hat{\tau}_i) g_i + \hat{\tau}_i g_i g'_i \hat{\kappa} \\ \hat{\tau}_i G'_i \hat{\kappa} + \hat{\tau}_i G'_i \hat{\lambda} g'_i \hat{\kappa} - \hat{\tau}_i G'_i \hat{\lambda} + \hat{\tau} G'_i \hat{\lambda} \end{bmatrix}. \quad (58)$$

(For convenience, the third block is obtained by subtracting Equation (53) from Equation (56).) Noting that  $\frac{\partial \hat{\tau}_i}{\partial \lambda} = \hat{\tau}_i g_i$ ,  $\frac{\partial \hat{\tau}}{\partial \lambda} = 0$ , and  $\frac{\partial \hat{\tau}_i}{\partial \theta} = \hat{\tau}_i G'_i \lambda$  the first expression for  $\phi(x_i, \hat{\beta})$  in Equation (23) also follows. ■

**Proof of Theorem 10.** We first establish consistency of  $\hat{\beta}$  in three steps. (i) Show that  $\hat{\lambda}(\theta) \xrightarrow{p} \lambda^*(\theta)$  uniformly for  $\theta \in \Theta$ . (ii) Show that  $\hat{\theta} \xrightarrow{p} \theta^*$  and therefore that  $\hat{\lambda}(\hat{\theta}) \xrightarrow{p} \lambda^*(\theta^*)$ . (iii) Show that it implies  $\hat{\tau} \xrightarrow{p} \tau^*$  and  $\hat{\kappa} \xrightarrow{p} \kappa^*$ .

Step 1. By Lemma 2.4 in [46], continuity of  $\exp(\lambda'g(x_i, \theta))$  in  $\lambda$  and  $\theta$ , Assumption 3(1) and (4) imply that  $\hat{M}_\theta(\lambda) \equiv n^{-1} \sum_i \exp(\lambda'g(x_i, \theta)) \xrightarrow{p} M_\theta(\lambda) \equiv E[\exp(\lambda'g(x_i, \theta))]$  uniformly over the compact set  $\{(\lambda', \theta) : \lambda \in \Lambda(\theta), \theta \in \Theta\}$  where  $\Lambda(\theta)$  is as in Assumption 3(4). We can then show that for any  $\eta > 0$ ,  $P[\sup_{\theta \in \Theta} \|\bar{\lambda}(\theta) - \lambda^*(\theta)\| \leq \eta] \rightarrow 1$  where  $\bar{\lambda}(\theta) = \arg \min_{\lambda \in \Lambda(\theta)} \hat{M}_\theta(\lambda)$  as follows. For a given  $\eta > 0$ , select  $\varepsilon = \inf_{\theta \in \Theta} \inf_{\lambda \in \Lambda(\theta): \|\lambda - \lambda^*(\theta)\| \geq \eta} (M_\theta(\lambda) - M_\theta(\lambda^*(\theta)))$ , which is nonzero by the strict convexity of  $M_\theta(\lambda)$  in  $\lambda$  and the fact that  $\Theta$  is compact. By the definition of  $\varepsilon$ , whenever  $\sup_{\theta} (M_\theta(\bar{\lambda}(\theta)) - M_\theta(\lambda^*(\theta))) \leq \varepsilon$ , then  $\sup_{\theta \in \Theta} \|\bar{\lambda}(\theta) - \lambda^*(\theta)\| \leq \eta$ . However, using the fact that  $(\hat{M}_\theta(\bar{\lambda}(\theta)) - \hat{M}_\theta(\lambda^*(\theta))) < 0$ , we have

$$\begin{aligned} & \sup_{\theta} (M_\theta(\bar{\lambda}(\theta)) - M_\theta(\lambda^*(\theta))) \\ & \leq \sup_{\theta} (M_\theta(\bar{\lambda}(\theta)) - \hat{M}_\theta(\bar{\lambda}(\theta))) + \sup_{\theta} (\hat{M}_\theta(\bar{\lambda}(\theta)) - \hat{M}_\theta(\lambda^*(\theta))) + \\ & \quad + \sup_{\theta} (\hat{M}_\theta(\lambda^*(\theta)) - M_\theta(\lambda^*(\theta))) \\ & \leq \sup_{\theta} |M_\theta(\bar{\lambda}(\theta)) - \hat{M}_\theta(\bar{\lambda}(\theta))| + \sup_{\theta} |\hat{M}_\theta(\lambda^*(\theta)) - M_\theta(\lambda^*(\theta))| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

w.p.a. 1. Hence,  $\sup_{\theta \in \Theta} \|\bar{\lambda}(\theta) - \lambda^*(\theta)\| \leq \eta$  w.p.a. 1. In order to obtain the same conclusion for  $\hat{\lambda}(\theta)$  rather than  $\bar{\lambda}(\theta)$ , we employ an argument similar to the proof of Theorem 2.7 in [46]. Since  $\hat{M}_\theta(\lambda)$  is convex in  $\lambda$  for any  $\theta$ , if the minimum  $\bar{\lambda}(\theta)$  lies in the interior of  $\Lambda(\theta)$ , no other points in the complement of  $\Lambda(\theta)$  can achieve a lower value and thus minimizing  $\hat{M}_\theta(\lambda)$  over  $\Lambda(\theta)$  or  $\mathbb{R}^{N_g}$  yields the same answer asymptotically. This establishes that  $\sup_{\theta \in \Theta} \|\hat{\lambda}(\theta) - \lambda^*(\theta)\| \xrightarrow{p} 0$ .

Step 2.  $\ln \hat{L}(\theta) \equiv -\ln \left( n^{-1} \sum_i \exp(\hat{\lambda}'(\theta)g(x_i, \theta)) \right) + \hat{\lambda}'(\theta)\hat{g}(\theta) \xrightarrow{p} \ln L(\theta)$  uniformly for  $\theta \in \Theta$ , because (i)  $\sup_{\theta \in \Theta} \|\hat{\lambda}(\theta) - \lambda^*(\theta)\| \xrightarrow{p} 0$  (ii)  $\sup_{\theta \in \Theta} \|\hat{g}(\theta) - E[g(x_i, \theta)]\| \xrightarrow{p} 0$  since  $g(x_i, \theta)$  is continuous in  $\theta$  and  $E[\sup_{\theta \in \Theta} \|g(x_i, \theta)\|] < \infty$  by Assumption 3(4) and the inequality  $|s| \leq \exp(-s) + \exp(s)$  for any  $s \in \mathbb{R}$ . (iii)  $\exp(\lambda'g(x_i, \theta))$  is

continuous in  $\theta$  and  $E \left[ \sup_{\theta \in \Theta} \sup_{\lambda \in \Lambda(\theta)} \exp(\lambda' g(x_i, \theta)) \right] < \infty$  by Assumption 3(4) (using Lemma 2.4 in [46]). Since  $\ln L(\theta)$  is uniquely maximized at  $\theta^*$ , this implies, along with the uniform convergence of  $\ln \hat{L}(\theta)$  and its continuity, that  $\hat{\theta} \xrightarrow{p} \theta^*$ . Since  $\sup_{\theta \in \Theta} \left\| \hat{\lambda}(\theta) - \lambda^*(\theta) \right\| \xrightarrow{p} 0$  we also have that  $\hat{\lambda}(\hat{\theta}) \xrightarrow{p} \lambda^*(\theta^*)$ .

Step 3. As we have shown that  $\hat{\theta} \xrightarrow{p} \theta^*$  and  $\hat{\lambda} \xrightarrow{p} \lambda^*$  and since  $\hat{\tau}$  and  $\hat{\kappa}$  can be written as explicit continuous functions of  $\hat{\lambda}$  and  $\hat{\theta}$ , by Equations (54) and (56), it follows that  $\hat{\tau} \xrightarrow{p} E[\tau_i] \equiv \tau^*$  and  $\hat{\kappa} \xrightarrow{p} (E[\tau_i g_i(x_i, \theta^*) g_i'(x_i, \theta^*)])^{-1} (\tau^* E[g_i(x_i, \theta^*)]) \equiv \kappa^*$ , where the fact that  $E[\tau_i g_i(x_i, \theta^*) g_i'(x_i, \theta^*)]$  is invertible is implied by the assumption that  $\Gamma$  is nonsingular.

Having established that  $\hat{\beta} \xrightarrow{p} \beta^*$ , we now turn to asymptotic normality. Since Lemma (9) defines a just-identified GMM estimator, we can use Theorem 3.4 in [46], specialized to the just-identified case, if we can show that (i)  $E \left[ \sup_{\beta \in \mathcal{B}} \|\partial \phi(x_i, \beta) / \partial \beta\| \right] < \infty$  for some neighborhood  $\mathcal{B}$  of  $\beta^*$  and that (ii)  $E \left[ \phi(x_i, \beta^*) \phi'(x_i, \beta^*) \right]$  exists.

The matrix  $\partial \phi(x_i, \beta) / \partial \beta'$  consists of terms of the form  $\alpha \exp(k_\tau \lambda' g_i) g_i^{k_g} G^{k_G} S^{k_S}$  for  $0 \leq k_g + k_G + k_S \leq 3$  and  $k_\tau = 0, 1$  and where  $g$ ,  $G$ , and  $S$  respectively denote elements of  $g_i$ ,  $G_i$ , and  $S_{jl}(x_i, \theta)$  and where  $\alpha$  denotes products of elements of  $\beta$  that are necessarily bounded for  $\beta \in \mathcal{B}$ . By Assumption 3(6), we can establish (i):  $\exp(k_\tau \lambda' g_i) |g|^{k_g} |G|^{k_G} |S|^{k_S} \leq \exp(k_\tau \lambda' g_i) |b(x_i)|^{k_g + k_G + k_S} \Rightarrow E[\sup_{\beta \in \mathcal{B}} \exp(k_\tau \lambda' g_i) |g|^{k_g} |G|^{k_G} |S|^{k_S}] \leq E[\sup_{\beta \in \mathcal{B}} \exp(k_\tau \lambda' g(x_i, \theta)) (b(x_i))^{k_2}] = E[\sup_{\theta \in \mathcal{N}} \sup_{\lambda \in \Lambda(\theta)} \exp(k_\tau \lambda' g(x_i, \theta)) (b(x_i))^{k_2}] < \infty$ . The matrix  $\phi(x_i, \beta) \phi'(x_i, \beta)$  has elements of the form  $\alpha \exp(k_\tau \lambda' g_i) |g|^{k_g} |G|^{k_G}$  with  $k_\tau = 0, 1, 2$  and  $0 \leq k_g + k_G \leq 4$  and a similar reasoning implies (ii). ■

## References

- [1] AKAIKE, H. (1973). Information theory and an extension of the likelihood principle. In *Proceedings of the Second International Symposium on Information Theory* (B. N. Petrov and F. Csáki, eds.). Budapest; Akadémiai Kiado.
- [2] ANGRIST, J., CHERNOZHUKOV, V. and FERNÁNDEZ-VAL, I. (2004). Quantile regression under misspecification, with an application to the U.S. wage structure. Working Paper, MIT.
- [3] BACK, K. and BROWN, D. P. (1993). Implied probabilities in GMM estimators. *Econometrica* **61** 971–975.
- [4] BAGGERLY, K. A. (1998). Empirical likelihood as goodness-of-fit measure. *Biometrika* **85** 535–547.
- [5] BICKEL, P. J. and GHOSH, J. K. (1990). A decomposition for the likelihood ratio statistic and the bartlett correction — a bayesian argument. *Ann. Statist.* **18** 1070–1090.
- [6] BONNAL, H. and RENAULT, E. (2004). On the efficient use of the informational content of estimating equations: Implied probabilities and euclidean empirical likelihood. Working paper, Université de Montréal.
- [7] BROWN, B. W. and NEWEY, W. K. (2002). Generalized method of moments, efficient bootstrapping, and improved inference. *J. Bus. Econom. Statist.* **20** 507–517.
- [8] CHEN, X., HONG, H. and SHUM, M. (2002). Nonparametric likelihood selection tests for parametric versus moment condition models. Working Paper, Johns Hopkins University.

- [9] COCHRANE, J. H. (1996). A cross-sectional test of an investment-based asset pricing model. *J. Polit. Econ.* **104** 572–621.
- [10] CORCORAN, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika* **85** 967–972.
- [11] CRESSIE, N. and READ, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. R. Statist. Soc. B* **46** 440–464.
- [12] CSISZAR, I. (1991). Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *Ann. Statist.* **19** 2032–2066.
- [13] DAHLHAUS, R. and WEFELMEYER, W. (1996). Asymptotically optimal estimation in misspecified time series models. *Ann. Statist.* **24** 952–974.
- [14] DICICCIO, T. J., HALL, P. and ROMANO, J. P. (1991). Empirical likelihood is bartlett correctable. *Ann. Statist.* **19** 1053–1061.
- [15] DICICCIO, T. J. and ROMANO, J. P. (1990). Nonparametric confidence limits by resampling methods and least favorable families. *Int. Stat. Rev.* **58** 59–76.
- [16] FAN, J., ZHANG, C. and ZHANG, J. (2001). Generalized likelihood ratio statistics and wilks phenomenon. *Ann. Statist.* **29** 153–193.
- [17] GOLAN, A., JUDGE, G. and MILLER, D. (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley and Sons, New York.
- [18] HALL, A. R. and INOUE, A. (2003). The large sample behavior of the generalized method of moments estimator in misspecified models. *J. Econometrics* **114** 361–394.

- [19] HALL, P. and HOROWITZ, J. (1996). Bootstrap critical values for tests based on generalized method of moment estimators. *Econometrica* **64** 891–916.
- [20] HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.
- [21] HANSEN, L. P. (1982). Large sample properties of generalized method of moment estimators. *Econometrica* **50** 1029–1054.
- [22] HANSEN, L. P., HEATON, J. and YARON, A. (1996). Finite-sample properties of some alternative GMM estimators. *J. Bus. Econom. Statist.* **14** 262–280.
- [23] HONG, H., PRESTON, B. and SHUM, M. (2003). Generalized empirical likelihood-based model selection criteria for moment condition models. *Econometric Theory* **19** 923.
- [24] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101.
- [25] HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* **1** 221–233.
- [26] IMBENS, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Rev. Econom. Stud.* **64** 359–383.
- [27] IMBENS, G. W. (2002). Generalized method of moments and empirical likelihood. *J. Bus. Econom. Statist.* **20** 493–506.
- [28] IMBENS, G. W. and SPADY, R. (2002). Confidence intervals in generalized method of moments models. *J. Econometrics* **107** 87–98.

- [29] IMBENS, G. W. and SPADY, R. H. (2001). The performance of empirical likelihood and its generalizations. Working Paper, University of California, Berkeley.
- [30] IMBENS, G. W., SPADY, R. H. and JOHNSON, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica* **66** 333–357.
- [31] JING, B.-Y. and WOOD, A. T. A. (1996). Exponential empirical likelihood is not bartlett correctable. *Ann. Statist.* **24** 365–369.
- [32] KITAMURA, Y. (2000). Comparing misspecified dynamic econometric models using nonparametric likelihood. Mimeo, Department of Economics, University of Wisconsin.
- [33] KITAMURA, Y. (2001). Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica* **69** 1661–1672.
- [34] KITAMURA, Y. (2003). A likelihood-based approach to the analysis of a class of nested and non-nested models. Working Paper, University of Pennsylvania.
- [35] KITAMURA, Y. and STUTZER, M. (1997). An information-theoretic alternative to generalized method of moment estimation. *Econometrica* **65** 861–874.
- [36] KITAMURA, Y. and STUTZER, M. (2002). Connections between entropic and linear projections in asset pricing estimation. *J. Econometrics* **107** 159–174.
- [37] KOLACZYK, E. D. (1994). Empirical likelihood and generalized linear models. *Statist. Sinica* **4** 199–218.
- [38] KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, Newyork.

- [39] LAZAR, N. A. and MYKLAND, P. (1999). Empirical likelihood in the presence of nuisance parameters. *Biometrika* **86** 203–211.
- [40] LEE, S. M. S. and YOUNG, G. A. (1999). Nonparametric likelihood ratio confidence intervals. *Biometrika* **86** 107–118.
- [41] MAASOUMI, E. and PHILLIPS, P. C. B. (1982). On the behavior of inconsistent instrumental variable estimators. *J. Econometrics* **19** 183–201.
- [42] MEGHIR, C. and WEBER, G. (1996). Intertemporal nonseparability or borrowing restrictions? a disaggregate analysis using a U.S. consumption panel. *Econometrica* **64** 1151–1181.
- [43] MITTELHAMMER, R. C., JUDGE, G. G. and SCHOENBERG, R. (2001). Empirical evidence concerning the finite sample performance of el-type structural equation estimation and inference methods. Working Paper, University of California, Berkeley.
- [44] MONFORT, A. (1996). A reappraisal of misspecified econometric models. *Econometric Theory* **12** 597–619.
- [45] MYKLAND, P. A. (1995). Dual likelihood. *Ann. Statist.* **23** 396–421.
- [46] NEWEY, W. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics* (R. F. Engel and D. L. McFadden, eds.), vol. IV. Elsevier Science.
- [47] NEWEY, W. and SMITH, R. J. (2004). Higher-order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72** 219–255.

- [48] OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.
- [49] OWEN, A. B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120.
- [50] OWEN, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, New York.
- [51] PATILEA, V. (2001). Convex models, mle and misspecification. *Ann. Statist.* **29** 94–123.
- [52] PFANZAGL, J. and WEFELMEYER, W. (1978). A third-order optimum property of the maximum likelihood estimator. *J. Multivariate Anal.* **8** 1–29.
- [53] QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325.
- [54] RAMALHO, J. J. S. and SMITH, R. J. (2002). Generalized empirical likelihood non-nested tests. *J. Econometrics* **107** 99–125.
- [55] ROTHENBERG, T. J. (1986). Approximating the distribution of econometric estimators and test statistics. In *Handbook of Econometrics*, vol. II. Elsevier Science, 882–932.
- [56] SAWA, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica* **46** 1273–1291.
- [57] SCHENNACH, S. M. (2005). Accompanying document to ‘point estimation using exponentially tilted empirical likelihood’. Technical Report, University of Chicago. <http://arxiv.org/abs/math.ST/0512181>.

- [58] SCHENNACH, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika* **92** 31–46.
- [59] SCHENNACH, S. M. and SPADY, R. H. (2003). Higher-order properties of GEL/EL estimators. Working paper, University of Chicago. <http://home.uchicago.edu/~smschenn/gelel01.pdf>.
- [60] SHORE, J. and JOHNSON, R. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inform. Theory* **26** 26–37.
- [61] SMITH, R. J. (1997). Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Econ. J.* **107** 503–510.
- [62] SPADY, R. H. (1996). Nonparametric inference by quasi-likelihood methods. Working Paper, Nuffield College.
- [63] VUONG, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57** 307–333.
- [64] WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–26.

Estimator	$\gamma$	$h(w)$	$\rho(\xi)$	$\tau(\xi)$
EL	-1	$-\ln nw$	$\ln(1 - \xi)$	$(1 - \xi)^{-1}$
ET	0	$nw \ln nw$	$-\exp(\xi)$	$-\exp(\xi)$
CU	1	$(nw)^2$	$-(1 + \xi)^2/2$	$-(1 + \xi)$
ECR	$\gamma$	$\frac{(nw)^{\gamma+1}-1}{\gamma(\gamma+1)}$	$-\frac{1}{\gamma+1}(1 + \gamma\xi)^{(\gamma+1)/\gamma}$	$-(1 + \gamma\xi)^{1/\gamma}$

Table 1: The EL, ET and CU estimators as particular cases of MED and GEL estimators (adapted from Newey and Smith (2004)).

	EL	ETEL	ET
$K = 4$	0.063	0.061	0.103
$K = 10$	0.129	0.103	0.232

Table 2: The bias of the EL, ETEL and ET estimators for the Hall-Horowitz design.

Estimator	$n = 1000$			$n = 5000$		
	EL	ETEL	ET	EL	ETEL	ET
Model C	0.032	0.032	0.032	0.014	0.014	0.014
Model M	0.054	0.038	0.031	0.052	0.019	0.014

Table 3: The standard deviations of the EL, ETEL and ET estimators for Models C and M defined in the text. The number of replications is 10000 for the  $n = 1000$  sample and 2000 for the  $n = 5000$  sample.

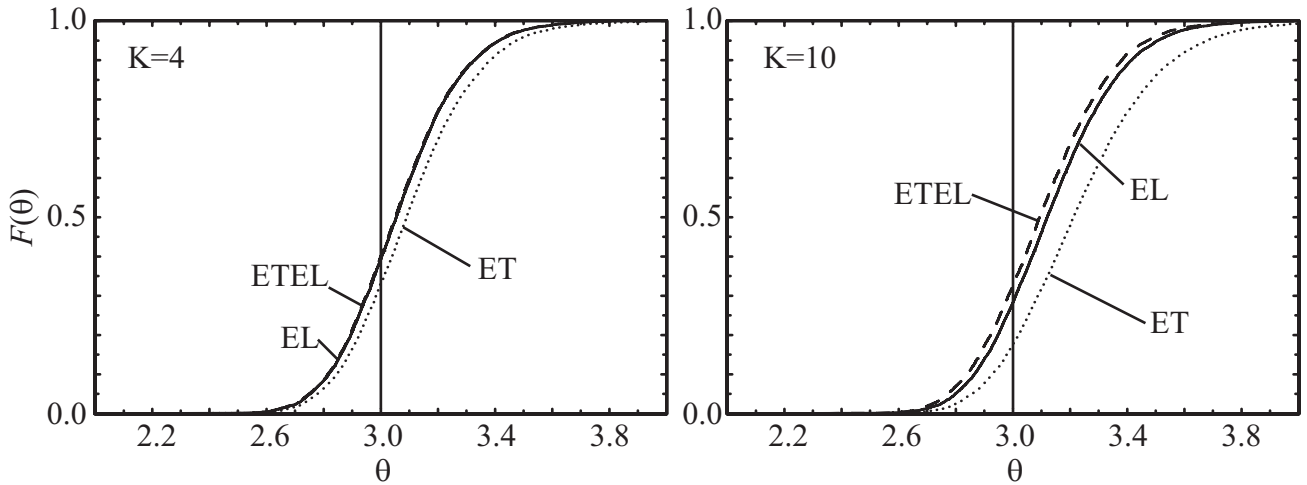


Figure 1: Cumulative distribution of the EL, ET and ETEL estimators for the Hall-Horowitz design with 4 (left) and 10 (right) moment conditions. The sample size is  $n = 200$  and 10000 replications were used to calculate this empirical cdf.

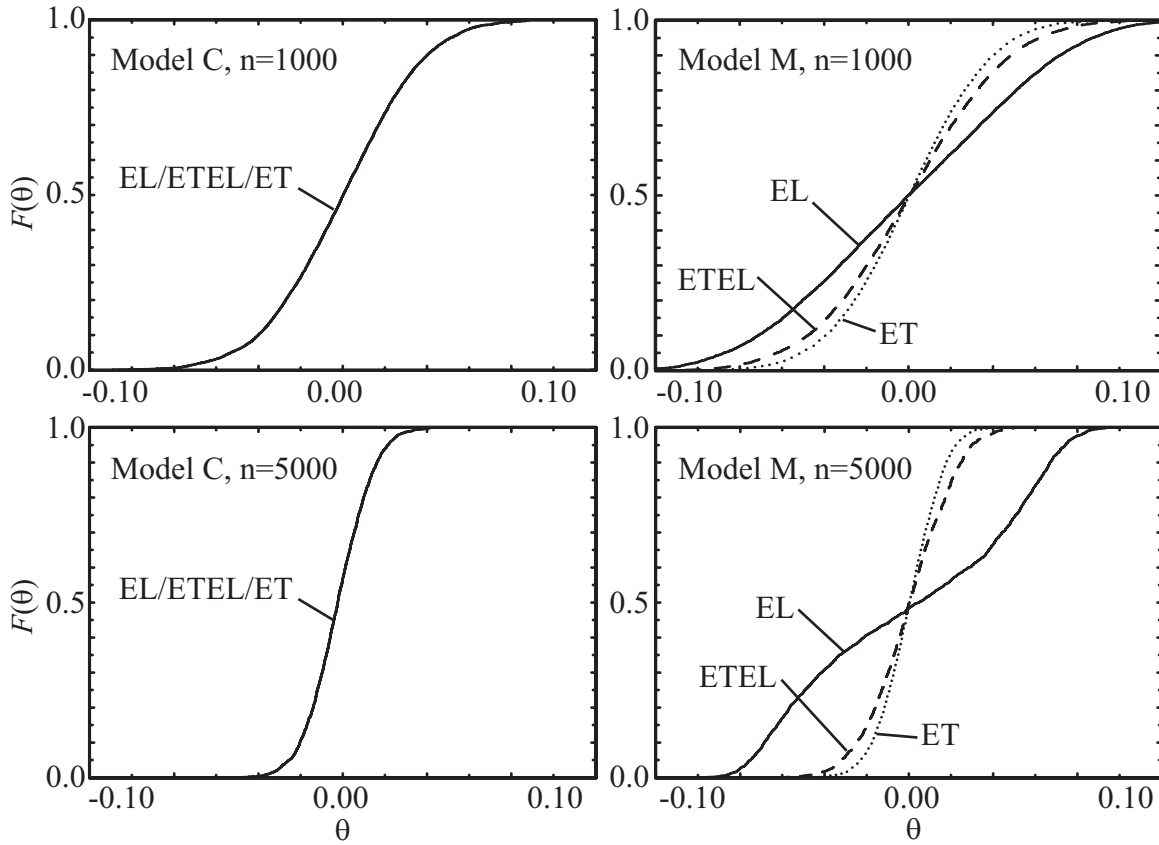


Figure 2: Cumulative distribution of the EL, ET and ETEL estimators for model C and M defined in the text. For the top portion of the figure, the sample size is  $n = 1000$  and 10000 replications were used. For the bottom portion of the figure,  $n = 5000$  with 2000 replications were used.

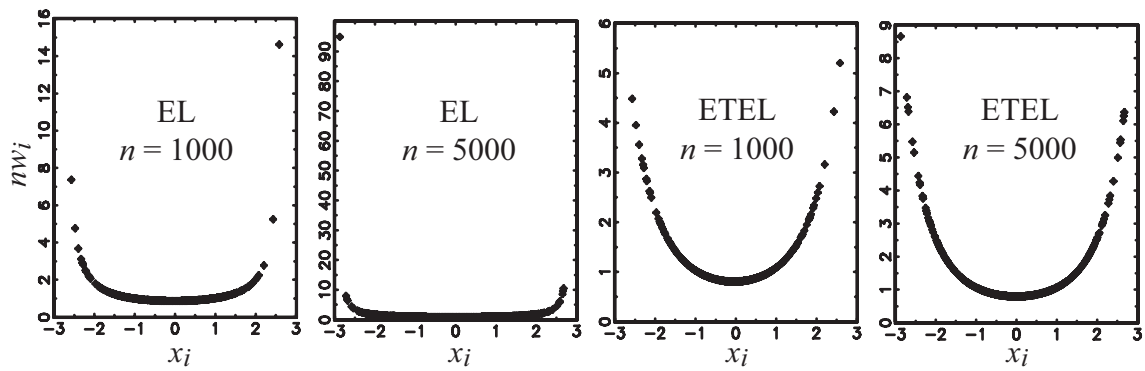


Figure 3: EL and ETEL implied probabilities in simulated samples drawn from the misspecified Model M as a function of sample size. Note the differences in the scale of the vertical axes.