ROBUST ESTIMATION WITH EXPONENTIALLY TILTED HELLINGER DISTANCE

BERTILLE ANTOINE AND PROSPER DOVONON

(November 2018)

Abstract. This paper is concerned with estimation of parameters defined by moment equalities. In this context, Kitamura, Otsu and Evdokimov (2013a) have introduced the minimum Hellinger distance (HD) estimator which is asymptotically semiparametrically efficient when the model is correctly specified and achieves optimal minimax robust properties under small deviations from the model (local misspecification). This paper evaluates the performance of inference procedures under two complementary types of misspecification, local and global. After showing that HD is not robust to global misspecification, we introduce, in the spirit of Schennach (2007), the exponentially tilted Hellinger distance (ETHD) estimator by combining the Hellinger distance and the Kullback-Leibler information criterion. Our estimator shares the same desirable asymptotic properties as HD under correct specification and local misspecification, and remains well-behaved under global misspecification. ETHD therefore appears to be the first estimator that is efficient under correct specification, and robust to both global and local misspecification.

Keywords: moment condition models; global misspecification; local misspecification; Hellinger distance; minimax robust estimation; semiparametric efficiency.

1. Introduction

It is well-recognized that economic models are simplification of reality and, as such, are intrinsically bound to be misspecified (see e.g. Maasoumi (1990), Hall and Inoue (2003), and Schennach (2007)). As a result, the choice of an inference procedure should not solely be based on its performance under correct specification, but also on its robustness to misspecification.

Two types of misspecification are outlined in the literature, so-called local and global misspecification. If the model of interest is one that describes the parameter of interest through moment restrictions, this model is globally misspecified if, under the true distribution of the data, no parameter value is compatible with the moment restrictions (see e.g. Kitamura (2000), Hall and Inoue (2003), and Schennach (2007)). This type of misspecification has been acknowledged for instance in modern asset pricing theory which advocates the use of moment condition models that depend on a pricing kernel to price financial assets. Unlike what the economic theory suggests, it is long recognized that no pricing

We would like to thank Pierre Chaussé, René Garcia, Christian Gouriéroux, Eric Renault, Susanne Schennach and Richard Smith for helpful discussions. The paper has also benefitted from the comments of an associate editor and two anonymous referees. Financial support from SSHRC (Social Sciences and Humanities Research Council) is gratefully acknowledged.

B. Antoine: Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, CANADA. *Email address:* Bertille_Antoine@sfu.ca.

P. Dovonon: Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, Quebec, CANADA. *E-mail address:* prosper.dovonon@concordia.ca (Corresponding author).

kernel can correctly price all financial securities. As a consequence, the pricing kernel used in applications is the one that is the least misspecified; see e.g. Hansen and Jagannathan (1997), Almeida and Garcia (2012), Kan, Robotti and Shanken (2013), Gospodinov, Kan and Robotti (2014), Gospodinov and Maasoumi (2017), and Antoine, Proulx and Renault (2018). Recently, misspecification has also been considered in moment based models by Ashley (2009), Nevo and Rosen (2012), Conley, Hansen, and Rossi (2012), Guggenberger (2012), and Kolesar, Chetty, Friedman, Glaeser, and Imbens (2015) who have studied inference in instrumental variable models including non-exogenous instruments; by Bravo (2014) who has studied inference in globally misspecified moment condition models including a semiparametric component; and by Cheng, Liao and Shi (2016) who have introduced an averaging GMM estimator from possibly misspecified models.

A moment condition is locally misspecified if, under the true distribution of the data, the moment condition is invalid for any finite sample size but the magnitude of violation is so small that it disappears asymptotically. Examples of local misspecification include the case where an asymptotically vanishing proportion of data sample is contaminated or exposed to measurement errors.

In this paper, we consider economic models defined by moment restrictions, and evaluate the performance of inference procedures of interest under these two complementary types of misspecification. Since the extent and nature of the misspecification are unknown in practice, it appears ideal to rely on inference procedures that are asymptotically efficient in correctly specified models, and asymptotically robust to both types of misspecification. To our knowledge, such an inference procedure is not currently available, and the main contribution of this paper is to fill this gap. An estimator robust to global misspecification remains asymptotically normal with the same rate of convergence as when the model is correctly specified. The appeal of such an estimator comes from the fact that its asymptotic distribution that is valid under both global misspecification and correct specification can be derived making inference immune to global misspecification routinely possible. Such an estimator is asymptotically centered around a *pseudo-true value* that matches the *true* parameter value if the model is correct.

By contrast, local misspecification is only noticeable in small samples (and not at the limit). Since the true distribution of the data is expected to match the one postulated by the researcher as the sample size gets large, one can define the true parameter value as the value that solves the assumed model. An efficient estimator is robust to local misspecification when its worse mean square error (computed over all possible small deviations of data distribution) remains the smallest in a certain class of estimators. Estimators that are robust to local misspecification remain consistent (for the true parameter value) so long as the true data-distribution is sufficiently close to the postulated distribution.

The study of large sample behaviour of estimators under model misspecification has registered a close attention in the econometric literature for more than three decades. Earlier work include White (1982) and Gouriéroux, Monfort and Trognon (1984) who study the maximum likelihood estimator. Hall and Inoue (2003) study the generalized method of moments (GMM) estimator under global misspecification

in a general setting extending the work of Maasoumi and Phillips (1982) and Gallant and White (1988) who focused on some GMM-type of estimators with special choice of weighting matrices. They show that, in the context of independent and identically distributed data, the two-step GMM estimator is asymptotically normal and its asymptotic distribution robust to global misspecification is provided.

More recently developed estimators for moment condition models have also been analyzed under global misspecification. We can cite the Euclidean Empirical Likelihood (EEL), also known as the continuously updated GMM, the exponential tilting (ET) and the maximum empirical likelihood (EL) estimators; all belonging to the Cressie-Read (CR) minimum power divergence class of estimators. These estimators rely on implied probabilities to re-weight the sample observations in order to guarantee that the moment condition is exactly satisfied (in sample). These estimators are defined as minimizers of some measure of discrepancy between the implied probabilities and the uniform weights $(1/n)$. Kitamura (2000) studies ET and establishes its robustness. The main advantage of EL is that, under correct specification, it has fewer sources of higher-order bias (see Newey and Smith, 2004). Schennach (2007) studies EL under global misspecification and shows that it is not robust. She identifies some singularity issues in the implied probability function of EL that are responsible for its lack of robustness; see also Smith (2007) and Broniatowski and Keziou (2012) for related conjectures. Then, observing that ET's implied probabilities do not display any such singularity, Schennach (2007) extends an approach previously considered in Corcoran (1998) and Jing and Wood (1996) to propose the exponentially tilted empirical likelihood (ETEL) estimator that combines EL's discrepancy function with ET's implied probabilities ETEL is quite appealing: it is efficient and shares the higher-order bias properties of EL in correct models, and remains as stable as ET in globally misspecified models; see also related results by Smith (2007) who combines GEL implied probabilities with EL criterion function. In addition to these estimators, a computationally friendly alternative to EL and ETEL, the so-called three-step EEL estimator, has been introduced by Antoine, Bonnal and Renault (2007) and proven to be robust by Dovonon (2016).

The concept of robust estimation to local misspecification has been formalized by Kitamura, Otsu and Evdokimov (KOE hereafter, 2013a) for parameters defined by general estimating equations in the form of a moment condition model. Building on the work of Beran (1977a,b) for fully parametric models, they equip the family of possible data distributions with the Hellinger topology and derive the asymptotic minimax bound for the mean square error of regular and Fisher consistent estimators. They also introduce the minimum Hellinger distance (HD) estimator which is shown to be asymptotically minimax robust; in addition, HD is much easier to compute than its fully parametric analogue due to Beran (1977a,b) which requires data density estimation. The behaviour of HD in globally misspecified models is unknown.

In this paper, we first explore the properties of HD in globally misspecified models and show that, similarly to EL, it does not behave well in general. HD turns out to be a member of the family of minimum power divergence estimators and the intuition for its lackluster performance follows

from the conjecture of Schennach (2007, p.641) that connects the poor performance of estimators from this family to the negative value of their indexing parameter (such as HD and EL). Actually, the only candidate from this family that retains good properties under global misspecification is ET. We then introduce the exponentially tilted Hellinger distance (ETHD) estimator that, in the spirit of Schennach's ETEL, combines ET and HD to deliver an estimator that retains the desirable properties of ET under global misspecification and those of HD under correct specification and local misspecification. Specifically, ETHD is efficient in correctly specified models and robust to both local and global misspecification. It is important to emphasize that ETHD, unlike HD, is not a saddle-point estimator and as a result its theoretical treatment calls for new proof techniques.

This paper is organized as follows. In Section 2, we briefly review the properties of HD under correct specification and local mispecification, and present a simple result that highlights its lackluster behavior under (global) misspecification. In Section 3, we introduce ETHD and derive its asymptotic properties under correct specification. Section 4 establishes that this estimator is asymptotically minimax robust to local misspecification while in Section 5, we show that ETHD is well-behaved and robust to global misspecification. The finite sample performance of this estimator is investigated in Section 6 through Monte Carlo simulations with a comparison to existing alternative estimators. Appendix A collects the graphs and tables of results of the Monte-Carlo study, while the proofs of our theoretical results are gathered in Appendices B, C and D.

2. HD under global misspecification

In this section, we introduce the minimum Hellinger distance estimator (HD) of Kitamura, Otsu and Evdokimov (2013) along with some of its properties and study its asymptotic behaviour under global misspecification. Let $\{X_i : i = 1, \dots, n\}$ be a random sample of independent and identically distributed random vectors distributed as X, with value in $\mathcal{X} \subset \mathbb{R}^d$; throughout, $E(.)$ denotes the expectation taken with respect to the true distribution of X. We assume that this sample is described by the moment restriction:

$$E\left(g(X, \theta^*)\right) = 0, \tag{1}$$

where θ^*, the parameter of interest, belongs to Θ, a compact subset of $\mathbb{R}^p$, $g(\cdot, \cdot)$ is an $\mathbb{R}^m$-valued function defined on $\mathcal{X} \times \Theta$, and $m \geq p$. Consider the Borel σ-field $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$ and let $\mathcal{M}$ be the set of all probability measures on this σ-field. Let π and ν be two elements of $\mathcal{M}$. The Hellinger distance between π and ν is given by

$$H(\pi, \nu) = \left[\frac{1}{2}\int \left(\sqrt{d\pi} - \sqrt{d\nu}\right)^2\right]^{1/2}. \tag{2}$$

If $\mathcal{X}$ is a finite or countable set, this distance takes the form

$$H(\pi, \nu) = \left[\frac{1}{2}\sum_{i \in \mathcal{X}} \left(\sqrt{\pi_i} - \sqrt{\nu_i}\right)^2\right]^{1/2}, \tag{3}$$

where π_i and ν_i are the measures of the outcome $\{i\}$ by π and ν, respectively. Throughout the paper, we let P_n denote the uniform discrete probability on $\mathcal{X}_d \equiv \{x_i : i = 1, \ldots, n\}$ where $\mathcal{X}_d$ is a realization of $\{X_i : i = 1, \ldots, n\}$.

2.1. **Definition and properties of HD.** The minimum Hellinger distance estimator $\hat{\theta}$ of θ^* is defined as

$$\hat{\theta}_{HD} \equiv \arg \inf_{\theta \in \Theta} \inf_{\pi \in \mathcal{M}_d} H^2(\pi, P_n), \qquad s.t. \qquad \sum_{i=1}^{n} \pi_i g(x_i, \theta) = 0, \tag{4}$$

where $\mathcal{M}_d$ is the set of all probability measures on $(\mathcal{X}_d, \mathscr{B}(\mathcal{X}_d))$.

By some simple algebra, one can see that HD belongs to the empirical Cressie-Read class of estimators and is associated to the power divergence function $h_{-1/2}$, where

$$h_a(\pi_i) = \frac{(n\pi_i)^{a+1} - 1}{a(a+1)}. \tag{5}$$

Recall that the empirical likelihood (EL) and the exponential tilting (ET) estimators are obtained for limit functions $h_{-1}(\pi) = -\ln(n\pi)$ and $h_0(\pi) = (n\pi)\ln(n\pi)$, respectively whereas the continuously updated estimator (CUE) is obtained for the quadratic divergence function $h_1(\pi)$.

Also, under some mild conditions and using some convex duality arguments, HD is alternatively defined as solution to the saddle-point problem (see KOE (2013a)):

$$\hat{\theta}_{HD} = \arg \min_{\theta \in \Theta} \max_{\gamma \in \mathbb{R}^m} -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \gamma' g(x_i, \theta)}, \qquad s.t. \qquad 1 + \hat{\gamma}' g(x_i, \hat{\theta}_{HD}) > 0. \tag{6}$$

Under this definition, HD fits into the generalized empirical likelihood (GEL) class of estimators introduced by Newey and Smith (2004) and is characterized by the saddle-point estimating function $\rho(v) = -1/(1+v)$ defined on the domain $\mathcal{V} = (-1, +\infty)$.

Remark 1. *The definition* (6) *explicitly*[1] *requires that,*

$$1 + \hat{\gamma}' g(x_i, \hat{\theta}_{HD}) > 0 \ \textit{for} \ (\hat{\theta}_{HD}, \hat{\gamma}) \ \textit{solving} \ (6) \ \textit{and for all} \ i = 1, \ldots, n,$$

since this condition is essential for the two definitions of the HD estimator in (4) *and* (6) *to be equivalent. This is due to the fact that the first order condition associated with the Lagrangian of the inner optimization program in* (4) *is*

$$1 + \gamma' g(x_i, \theta) = \frac{1}{\sqrt{n\pi_i}},$$

for all $i = 1, \ldots, n$ *in the direction of* π*. Hence, solutions for* π *exist only if*

$$1 + \hat{\gamma}' g(x_i, \hat{\theta}_{HD}) > 0 \, , \textit{for all} \ i = 1, \ldots, n.$$

See also the formal result and related discussions in Almeida and Garcia (2017) for all GEL estimators.

[1] KOE (2013a) do not explicitly maintain such positivity constraint.

In correctly specified models, this condition can be overlooked since the Lagrange multiplier $\hat{\gamma}$ associated to $\hat{\theta}$ obtained from (6) *converges sufficiently fast to 0 (under regularity conditions) to guarantee that $\hat{\gamma}' g(x_i, \hat{\theta})$ is uniformly negligible for n large enough. However, in possibly misspecified models, this condition may matter. This has non trivial advantage in case of model misspecification since the probability limit of* (6) *can then be interpreted as the parameter value with induced set of probability distributions*[2] *closest to the true distribution of the data under the Hellinger distance. Such an interpretation is built in the definition in* (4)*.*

If the moment restriction in (1) is correctly specified and point identified, meaning that (1) holds at only one point θ^* in the parameter space Θ, then $\hat{\theta}_{HD}$ is consistent for θ^*. In fact, as a member of the GEL class of estimators, under Assumptions 1 and 2 of Newey and Smith (2004), their Theorem 3.2 applies to HD. Letting

$$G = E\left(\frac{\partial g(X,\theta^*)}{\partial \theta'}\right), \quad \Omega = E\left(g(X,\theta^*)g(X,\theta^*)'\right) \quad \textit{and} \quad \Sigma = \left(G'\Omega^{-1}G\right)^{-1},$$

it is established that

$$\sqrt{n}(\hat{\theta}_{HD} - \theta^*) \xrightarrow{d} N(0, \Sigma). \tag{7}$$

This shows that in correctly specified models, HD is $\sqrt{n}$-consistent and asymptotically normal and efficient as it reaches the semiparametric efficiency bound. As we shall see in Section 4, KOE (2013a) show that this estimator is also minimax robust to local misspecification of the data generating process. Specifically, under some small perturbations of the data generating process, the maximum asymptotic mean square error of this estimator is smallest in the family of regular and Fisher consistent estimators (see Definition 1).

2.2. **Behavior of HD under global misspecification.** Statistical models being simplifications of reality, the data generating process may be such that the moment condition model in (1) does not actually have a solution in the parameter set Θ. This can actually be expected in settings where the model is overidentifying in the sense that more moment restrictions than unknown parameters are available i.e. $(m > p)$. This type of misspecification is referred to as global misspecification (see Hall and Inoue (2003) and Schennach (2007)). Formally, the moment condition model (1) is globally misspecified if

$$E\left(g(X,\theta)\right) \neq 0, \quad \forall \theta \in \Theta.$$

Under global misspecification, the notion of consistent estimator no longer makes much sense even though a particular estimator is expected to converge to a specific value in the parameter set which is referred to as its pseudo-true value. Of course, in correctly specified models and under mild identification conditions, pseudo-true values are the same for all consistent estimators with common limit being the solution of the model.

[2]For a given value $\theta \in \Theta$, an *induced distribution* is any distribution P satisfying $E_P(g(X,\theta)) = 0$, where $E_P(\cdot)$ stands for expectation under probability P.

In fact, asymptotic theory for estimators can be derived either assuming that the model is correctly specified or allowing for global misspecification. If the asymptotic distribution of an estimator derived allowing for global misspecification is equivalent, under correct specification, to the asymptotic distribution of that estimator derived assuming correct specification, this estimator is said to be robust to global misspecification. Such robustness is desirable because it allows for the possibility to carry out valid and reliable inference whether the model is correctly specified or not by using the misspecification-robust asymptotic distribution of the concerned estimator. Hall and Inoue (2003) show that GMM is robust to global misspecification. One can also refer to White (1982) who derives the asymptotic distribution of the maximum likelihood estimator under possible model misspecification.

The next result explores the asymptotic behaviour of HD under global mispecification. We derive for HD a result similar to that of Schennach (2007, Theorem 1) for empirical likelihood (EL), and according to which, EL is not robust to global misspecification since it is not $\sqrt{n}$-convergent in globally misspecified models.

Theorem 2.1. *(Lack of robustness of HD under global misspecification)*
Let $\{X_i : i = 1, \ldots, n\}$ be an i.i.d. sequence of random vectors distributed as X. Assume $g(x, \theta)$ to be twice continuously differentiable at all $\theta \in \Theta$ and for all x and is such that

$$\sup_{\theta \in \Theta} E\left[\|g(X, \theta)\|^2\right] < \infty .$$

If

$$\inf_{\theta \in \Theta} \|E[g(X, \theta)]\| \neq 0 \quad \textit{and} \quad \sup_{x \in \mathcal{X}} u' g(x, \theta) = \infty$$

for any $\theta \in \Theta$ and any unit vector u, then there does not exist any $\theta^* \in \Theta$ such that

$$\|\hat{\theta}_{HD} - \theta^*\| = O_P\left(\frac{1}{\sqrt{n}}\right) .$$

This result shows that HD does not converge to its potential pseudo-true value at the standard rate of $\sqrt{n}$ in general in case of global misspecification. Existence of second moments of the estimating function $g(X, \theta)$ and its unboundedness are sufficient conditions for HD not to be $\sqrt{n}$-consistent. Such conditions are fulfilled for instance if $g(X, \theta)$ is normally distributed with non degenerate variance. In the light of the standard behaviour of HD under correct specification, as shown in (7), $\sqrt{n}$-convergence under global misspecification is a necessary condition for HD to be robust to global misspecification which clearly is not always the case as shown by this result.

It is worth mentioning that the lack of robustness of HD to global misspecification is not surprising. The intuition for such a lackluster performance follows from Schennach's (2007, p.641) conjecture that connects the poor performance of estimators from the Cressie-Read family to the negative value of their indexing parameter; see also related conjectures in Smith (2007, p.250) and Broniatowski and Keziou (2012, p.2566). As recalled in (5), HD is associated with index $a = -1/2$. Actually, it is expected that power divergence estimators associated with negative Cressie-Read index have nonnegative implied

probabilities, π_i's, but are not robust to global misspecification whereas those with positive index are robust to global misspecification but have implied probabilities that can be negative. It turns out that the only Cressie-Read estimator that is well-behaved under global misspecification with nonnegative implied probabilities is the exponentially tilted (ET) estimator with index $a = 0$.

This desirable property of ET has motivated its use in two-step estimation procedures that yield estimators robust to global misspecification with interesting bias properties such as the exponentially tilted empirical likelihood estimator (ETEL) of Schennach (2007). We follow this approach and introduce in the next section the exponentially tilted Hellinger distance estimator (ETHD). We subsequently show that this new estimator has the same first-order asymptotic properties as HD under correct specification, the same minimax robustness properties as HD under local misspecification and the additional advantage of being robust to global misspecification.

3. The Exponentially Tilted Hellinger Distance estimator

The exponentially tilted Hellinger distance estimator (ETHD) that we introduce in this section borrows an idea similar to Schennach (2007) who introduces ETEL. ETHD exploits the robustness of ET's implied probabilities and is equal to the value in the parameter space that sets the Hellinger distance between these implied probabilities and the empirical distribution to the minimum. This estimator is formally introduced next. We also discuss its first-order asymptotic properties in correctly specified models.

3.1. Definition and characterization of ETHD.

The exponentially tilted Hellinger distance estimator (ETHD), $\hat{\theta}$, is defined as:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} H(\hat{\pi}(\theta), P_n), \tag{8}$$

where H is given by (3) and $\hat{\pi}(\theta) = \{\hat{\pi}_i(\theta)\}_{i=1}^n$ is the solution of

$$\min_{\{\pi_i\}_{i=1}^n} \sum_{i=1}^n \pi_i \ln(n\pi_i) \tag{9}$$

subject to

$$\sum_{i=1}^n \pi_i g(x_i, \theta) = 0 \quad \text{and} \quad \sum_{i=1}^n \pi_i = 1. \tag{10}$$

It follows from (9)-(10) that for any $\theta \in \Theta$, the implied probabilities are functions of θ and given by:

$$\hat{\pi}_i(\theta) = \frac{\exp\left(\hat{\lambda}(\theta)' g(x_i, \theta)\right)}{\sum_{j=1}^n \exp\left(\hat{\lambda}(\theta)' g(x_j, \theta)\right)}, \quad i = 1, \ldots, n \tag{11}$$

with $\hat{\lambda}(\theta)$ implicitly determined by the equation (see Kitamura (2007)):

$$\frac{1}{n}\sum_{i=1}^n g(x_i, \theta) \exp(\hat{\lambda}(\theta)' g(x_i, \theta)) = 0.$$

As a result,

$$H^2(\hat{\pi}(\theta), P_n) = 1 - \mathbf{\Delta}_{P_n}(\hat{\lambda}(\theta), \theta),$$

with

$$\mathbf{\Delta}_{P_n}(\lambda, \theta) = \frac{\frac{1}{n}\sum_{i=1}^{n} \exp\left(\lambda' g(x_i, \theta)/2\right)}{\left(\frac{1}{n}\sum_{i=1}^{n} \exp\left(\lambda' g(x_i, \theta)\right)\right)^{\frac{1}{2}}} = \frac{E_{P_n}[\exp\left(\lambda' g(X, \theta)/2\right)]}{\sqrt{E_{P_n}[\exp\left(\lambda' g(X, \theta)\right)]}},$$

with $E_{P_n}(f(X)) = \sum_{i=1}^{n} f(x_i)/n$. The next theorem gives an alternative definition of ETHD along with the first-order optimality condition that it solves.

Theorem 3.1. *Assume $g(x, \theta)$ to be continuously differentiable at all $\theta \in \Theta$ and for all x. The ETHD estimator $\hat{\theta}$ maximizes $\mathbf{\Delta}_{P_n}\left(\hat{\lambda}(\theta), \theta\right)$ and if it is an interior optimum, it solves the first-order condition:*

$$\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\hat{\pi}_i(\hat{\theta})}\right)\left(\sum_{j=1}^{n}\hat{\pi}_j(\hat{\theta})\frac{d(\hat{\lambda}(\hat{\theta})' g(x_j, \hat{\theta}))}{d\theta}\right) - \frac{1}{n}\sum_{i=1}^{n}\sqrt{\hat{\pi}_i(\hat{\theta})}\frac{d(\hat{\lambda}(\hat{\theta})' g(x_i, \hat{\theta}))}{d\theta} = 0.$$

Remark 2. *(i) The square root function being strictly concave, the Jensen's inequality ensures that $0 \leq \mathbf{\Delta}_{P_n}(\lambda, \theta) \leq 1$ for all $(\lambda, \theta) \in \mathbb{R}^m \times \Theta$ and, under very mild conditions, $\mathbf{\Delta}_{P_n}(\lambda, \theta) = 1$ only for $\lambda = 0$.*

(ii) It is worth mentioning that it appears sometimes more convenient to define $\hat{\lambda}(\theta)$ as:

$$\hat{\lambda}(\theta) = \arg\max_{\lambda \in \mathbb{R}^m} -\frac{1}{n}\sum_{i=1}^{n}\exp\left(\lambda' g(x_i, \theta)\right). \tag{12}$$

This definition is particularly useful for computing ETHD as it involves an inner and an outer loop optimization; a computation procedure similar to the nested optimization routine described by Kitamura (2007) for EL. We refer to Kitamura (2007, Section 8.1) for practical details on the implementation of such programs. We shall also rely on this definition in Section 4 as we establish the robustness of ETHD to local misspecification.

(iii) By definition, the implied probabilities $\hat{\pi}(\hat{\theta})$ yielded by ETHD are positive. This estimator also enjoys some invariance properties both to one-to-one (model) parameter transformations and to nonsingular model transformations. By the latter, we mean that if $A(\theta)$ is a nonsingular matrix, ETHD of $E(A(\theta)g(X, \theta)) = 0$ and that of $E(g(X, \theta)) = 0$ are numerically equal.

3.2. **First-order asymptotic properties of ETHD.** This section establishes consistency and asymptotic normality of ETHD. We also show that the maximum of $\mathbf{\Delta}_{P_n}(\hat{\lambda}(\theta), \theta)$ reached at ETHD can be used for model specification testing. We maintain the following regularity assumptions.

Assumption 1. *(i) $\{X_i : i = 1, \ldots, n\}$ is a sequence of i.i.d. random vectors distributed as X.*
(ii) $g(X, \theta)$ is continuous at each $\theta \in \Theta$ with probability one and Θ is compact.

(iii) $E(g(X,\theta)) = 0 \Leftrightarrow \theta = \theta^*$.
(iv) $E(\sup_{\theta\in\Theta} \|g(X,\theta)\|^\alpha) < \infty$ *for some* $\alpha > 2$.
(v) $Var(g(X,\theta))$ *is nonsingular for all* $\theta \in \Theta$ *with smallest eigenvalues* $\underline{\ell}$ *bounded away from 0.*
(vi) $E\left(\sup_{(\theta\in\Theta,\lambda\in\Lambda)} \exp(\lambda' g(X,\theta))\right) < \infty$, *where* Λ *is a compact subset of* $\mathbb{R}^m$ *containing an open neighborhood of 0.*

Assumptions 1*(i)-(iv)* are standard in the literature on inference based on moment condition models. Newey and Smith (2004) have established the consistency of the generalized empirical likelihood class of estimators under this set of assumptions. Assumption 1*(iii)* imposes strong identification of the true parameter value and thereby rules out weak identification settings studied by Stock and Wright (2000), Kleibergen (2005) and Andrews and Cheng (2012) among others. Because of the two-step nature of our estimation procedure, it is useful to maintain a dominance condition over $\Lambda \times \Theta$ and this explains our additional Assumption 1*(vi)*. Schennach (2007) has also made use of a similar assumption to establish the consistency of ETEL. Assumption 1*(v)* is stronger than the standard assumption (see Assumption 1(e) of Newey and Smith, 2004) which imposes nonsingularity of variance only at θ_0. We require this stronger version again because of the two-step nature of our problem. This assumption rules out points $\theta \in \Theta$ such that $\lambda(\theta) \equiv \arg\min_{\theta\in\Lambda} E[\exp(\lambda' g(X,\theta))] \neq 0$ and $\lambda(\theta)' g(X,\theta) = cst$, P-almost surely. At such points, $\mathbf{\Delta}(\lambda(\theta),\theta)$ would be maximum without them necessarily being the true value; with $\mathbf{\Delta}$ the population version of $\mathbf{\Delta}_{P_n}$. Actually, our main results below hold if we replace *(v)* by: *(v')*: $Var(g(X,\theta_0))$ *is nonsingular and,*

$$\forall \theta \in \Theta, \quad \left(\lambda(\theta)' g(X,\theta) \text{ is constant } P\text{-a.s. if and only if } \lambda(\theta) = 0\right).$$

However, we prefer *(v)* over *(v')* since the former is potentially easier to investigate.

It is worth mentioning that all the results in this section continue to hold if Λ is set to be a neighborhood of 0 that shrinks with increasing n, but at a rate slightly slower than $O(1/\sqrt{n})$. In this case, both *(v)* and *(vi)* can be removed and replaced by $Var(g(X,\theta_0))$ nonsingular.

Under Assumption 1, instead of (12), we shall consider the following alternative definition of $\hat{\lambda}(\theta)$:

$$\hat{\lambda}(\theta) = \arg\max_{\lambda\in\Lambda} -\frac{1}{n}\sum_{i=1}^{n} \exp\left(\lambda' g(x_i,\theta)\right). \tag{13}$$

This definition is theoretically more tractable in the proof of consistency, thanks to the compactness of Λ. For practical purposes, Λ can be taken arbitrarily large. Importantly, this definition of $\hat{\lambda}(\theta)$ does not alter the asymptotic properties of $\hat{\theta}$ so long as the interior of Λ contains 0 which is the population value of λ in correctly specified models.

Theorem 3.2. *(Consistency of the ETHD estimator)*
If Assumption 1 holds, then
(i) $\hat{\theta} \xrightarrow{P} \theta^*$; *(ii)* $\hat{\lambda}(\hat{\theta}) = O_P(n^{-1/2})$; *and* *(iii)* $\frac{1}{n}\sum_{i=1}^{n} g(x_i,\hat{\theta}) = O_P(n^{-1/2})$.

To establish asymptotic normality of ETHD, we further assume the following.

Assumption 2. *(i) $\theta^* \in int(\Theta)$; there exists a neighborhood $\mathcal{N}$ of θ^* such that $g(X,\theta)$ is twice continuously differentiable almost surely on $\mathcal{N}$ and $E\left(\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial g(X,\theta)}{\partial\theta'}\right\|\right) < \infty$.*
(ii) Rank$(G) = p$, with $G = E(\partial g(X,\theta^*)/\partial\theta')$.

Similarly to the two-step GMM procedure, the maximum of $\mathbf{\Delta}_{P_n}(\hat{\lambda}(\theta),\theta)$, reached at $\hat{\theta}$ can be used to test for the validity of the moment condition model. We consider the specification test statistics:

$$S_{1,n} = 8nH^2(\hat{\pi}(\hat{\theta}), P_n) = 8n(1 - \mathbf{\Delta}_{P_n}(\hat{\lambda}(\hat{\theta}),\hat{\theta})), \quad \text{and} \quad S_{2,n} = n\hat{\lambda}(\hat{\theta})'\hat{\Omega}\hat{\lambda}(\hat{\theta}), \tag{14}$$

with $\hat{\Omega}$ any consistent estimator of Ω. The asymptotic distributions of $S_{1,n}$ and $S_{2,n}$, along with that of ETHD are given by the following result.

Theorem 3.3. *(Asymptotic distribution of the ETHD estimator)*
Let $\hat{\lambda} = \hat{\lambda}(\hat{\theta})$. If Assumptions 1 and 2 hold, then:

(i)

$$\sqrt{n}\begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\lambda} \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \Sigma & 0 \\ 0 & \Omega^{-1/2}M\Omega^{-1/2} \end{pmatrix}\right),$$

with $\Omega = E(g(X,\theta^*)g(X,\theta^*)')$, $\Sigma = \left[G'\Omega^{-1}G\right]^{-1}$ *and* $M = I_m - \Omega^{-1/2}G\Sigma G'\Omega^{-1/2}$.

(ii) $S_{1,n} = S_{2,n} + o_P(1)$ *and both* $S_{1,n}, S_{2,n} \xrightarrow{d} \chi^2_{m-p}$.

This result shows that, under correct specification, ETHD has the same limiting distribution as the efficient two-step GMM, which also corresponds to the limiting distribution of the HD estimator as recalled in (7). Convergence and asymptotic normality properties of ETHD result from the strong identification framework imposed by Assumptions 1 and 2. These properties show that this estimator does not suffer from the so-called no moment problem of GEL estimators that occurs under weak identification as pointed out by Guggenberger (2008). The specification test statistics $S_{j,n}$ $(j = 1, 2)$ have the same asymptotic distribution as the Hansen's (1982) J-test statistic. The proof actually reveals that these test statistics are asymptotically equivalent under the conditions of the theorem.

4. ETHD under local misspecification

KOE has provided a framework to study robustness of estimators of finite dimension parameter of models defined with moment equality. Following the work of Beran (1977a,b) for parametric models, they express robustness properties in terms of local minimax loss properties. Assuming that X has the probability distribution P, an estimator of θ^* is minimax robust if, under small perturbations of data distributions around P, that estimator has the smallest worst loss as measured for instance by the estimator's mean square error. Because of the local nature of this robustness property, we shall refer to it as robustness to local misspecification to emphasize the difference with global misspecification as

introduced in section 2. It is important here to stress that robustness to global misspecification does not imply robustness to local misspecification and vice-versa. The GMM estimator is an example of estimator that is robust to global misspecification without being minimax robust to local misspecification. Also, as shown in the previous section, HD is not robust to global misspecification but is locally minimax robust.

In this section, we establish that ETHD is minimax robust to local misspecification. To this end, letting again $\mathcal{M}$ be the set of all probability measures on the Borel σ-field $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$, $\mathcal{X} \subset \mathbb{R}^d$, and $g : \mathcal{X} \times \Theta \to \mathbb{R}^m$, we introduce the functionals $\quad T_1 : \mathcal{M} \times \Theta \to \Lambda \quad$ and $\quad T : \mathcal{M} \to \Theta$ as follows (we shall subsequently discuss their well-definedness)

$$T(P) = \arg\max_{\theta \in \Theta} \frac{\int \exp\left(T_1(\theta, P)' g(X, \theta)/2\right) dP}{\left(\int \exp\left(T_1(\theta, P)' g(X, \theta)\right) dP\right)^{\frac{1}{2}}}, \tag{15}$$

and

$$T_1(\theta, P) = \arg\max_{\lambda \in \Lambda} \left(- \int \exp(\lambda' g(X, \theta)) dP \right). \tag{16}$$

ETHD is then given by $\hat{\theta} = T(P_n)$. The common approach to study minimax robustness to local misspecification consists in evaluating the magnitude of the mean square error of the estimator of interest,

$$E_Q\left(\sqrt{n}(T(P_n) - \theta^*)^2\right),$$

where θ^* is the true parameter value associated with the genuine probability distribution of the data that we denote P_*, and Q is a probability measure lying in a shrinking Hellinger-neighborhood of P_*. Specifically, Q is assumed to lie in a Hellinger ball, $B_H(P_*, r/\sqrt{n})$, centered at P_* and with radius $r/\sqrt{n}$ for some $r > 0$:

$$B_H(P_*, r/\sqrt{n}) = \left\{Q \in \mathcal{M} : H(Q, P_*) \leq r/\sqrt{n}\right\}.$$

Note that $T(P_*) = \theta^*$. Since Q stands as the hypothetical distribution of the data for a given n, $T(Q)$ would stand for the true parameter value under Q and the decomposition

$$T(P_n) - \theta_* = (T(P_n) - T(Q)) + (T(Q) - \theta_*)$$

appears convenient for the analysis of the mean square error, with $T(Q) - \theta_*$ representing the bias resulting from estimating θ_* by $T(P_n)$. However, because Q is an arbitrary element of $B_H(P_*, r/\sqrt{n})$, the functional T may not be well-defined at all Q and this is in particular due to the unboundedness of $g(x, \theta)$ for some $\theta \in \Theta$. To overcome this technical limitation, we follow KOE and resort to trimming.

Let

$$\mathcal{X}_n = \left\{ x \in \mathcal{X} : \sup_{\theta \in \Theta} \|g(x, \theta)\| \leq m_n \right\}, \quad g_n(x, \theta) = g(x, \theta) \mathbb{I}(x \in \mathcal{X}_n),$$

$$\boldsymbol{\Delta}_{n,Q}(\lambda,\theta)=\frac{\int\exp\{\lambda' g_n(X,\theta)/2\}dQ}{\left(\int\exp\{\lambda' g_n(X,\theta)\}dQ\right)^{1/2}}$$

and define:

$$\bar{T}(Q)=\arg\max_{\theta\in\Theta}\boldsymbol{\Delta}_{n,Q}(T_1(\theta,Q),\theta)\quad\text{with}\quad T_1(\theta,Q)=\arg\max_{\lambda\in\Lambda}-\int\exp\{\lambda' g_n(X,\theta)\}dQ. \tag{17}$$

If well-defined, $\bar{T}(\cdot)$ is the value of $\theta\in\Theta$ that minimizes the Hellinger distance between $P(\theta)$ and Q, where $P(\theta)$ is the distribution that minimizes the Kullback-Leibler information criterion between Q and the set of distributions P that satisfy $E_P\left(g_n(X,\theta)\right)=0$; see Lemma C.1 for a proof.

By continuity (in λ) of its objective function and compactness of Λ, the argmax set $T_1(\theta,Q)$ is nonempty for any $\theta\in\Theta$ and $Q\in\mathcal{M}$. But this set may not be a singleton in general and $\boldsymbol{\Delta}_{n,Q}(T_1(\theta,Q),\theta)$ is not guaranteed to be a proper function. Because of this, one may rather consider the following alternative definition for $\bar{T}(Q)$:

$$\bar{T}(Q)=\arg\max_{\theta\in\Theta}\max_{\lambda\in\bar{T}_1(\theta,Q)}\boldsymbol{\Delta}_{n,Q}(\lambda,\theta)\quad\text{with}\quad\bar{T}_1(\theta,Q)=\arg\max_{\lambda\in\Lambda}-\int\exp\{\lambda' g_n(X,\theta)\}dQ, \tag{18}$$

where we keep the same notation as in (17) for the estimator. The maximization over $\bar{T}(\theta,Q)$ makes it easier to prove that $\bar{T}$ is well-defined over $\mathcal{M}$ (see Lemma C.2(i)). However, as shown by the second section of the same lemma, if we further impose that Λ is convex with interior containing the origin 0, for n large enough, there exists a neighborhood of θ^* over which $\bar{T}_1(\theta,Q)$ is a singleton for any Q lying in the Hellinger ball $B_H(P_*,r/\sqrt{n})$. In fact, Lemma C.3(iv) shows that $\bar{T}(Q_n)$ converges to θ^* for any sequence Q_n in that ball. Also, for n large enough, under some mild conditions, $-E_{P_n}\left[\exp(\lambda' g_n(X,\theta))\right]$ is strictly concave in λ and therefore its maximum over the convex and compact set Λ is reached at a unique point. Hence, both $\bar{T}_1(\bar{T}(Q_n),Q_n)$ and $\bar{T}_1(\bar{T}(P_n),P_n)$ are sets containing a single element for n large enough. As a result, for the sequences of measures of interest for local misspecification studies, the inner maximization can be dropped out and the same is also true regarding estimation.

The asymptotic minimax robustness of HD has been established by KOE by comparing its asymptotic worst loss to the smallest worst loss achievable by any estimator that is asymptotically Fisher consistent and regular. We refer to their Definition 3.1 for these properties that we recall below to be self-contained.

Definition 1. ***(Fisher consistent and regular estimator)*** *Let $T_a(P_n)$ be an estimator of θ^* based on a mapping $T_a:\mathcal{M}\to\Theta$. Let $\mathcal{P}$ be the set of all probability measures P for which there exists $\theta\in\Theta$ satisfying $E_P(g(X,\theta))=0$ and let $P_{\theta,\zeta}$ be a regular parametric submodel (see Bickel,*

Klassen, Ritov, and Wellner (1993, p. 12) or Newey (1990)) of $\mathcal{P}$ *such that* $P_{\theta^*,0} = P_*$ *and such that* $P_{\theta^*+t/\sqrt{n},\zeta_n} \in B_H(P_*, r/\sqrt{n})$ *holds for* $\zeta_n = O(n^{-1/2})$ *eventually.*

(i) T_a *is asymptotically Fisher consistent if for every* $(P_{\theta^*+t/\sqrt{n},\zeta_n})_{n\in\mathbb{N}}$ *and* $t \in \mathbb{R}^p$,

$$\sqrt{n}\left(T_a(P_{\theta^*+t/\sqrt{n},\zeta_n}) - \theta^*\right) \to t.$$

(ii) T_a *is regular for* θ^* *if, for every* $(P_{\theta_n,\zeta_n})_{n\in\mathbb{N}}$ *with* $\theta_n = \theta + O(n^{-1/2})$ *and* $\zeta_n = O(n^{-1/2})$, *there exists a probability measure* M *such that:*

$$\sqrt{n}(T_a(P_n) - T_a(P_{\theta_n,\zeta_n})) \xrightarrow{d} M, \quad \textit{under} \quad P_{\theta_n,\zeta_n},$$

where the measure M *does not depend on the sequence* (θ_n, ζ_n).

Following KOE, we consider the estimation problem of the transformed scalar parameter $\tau(\theta^*)$, where τ is an arbitrary smooth function defined on Θ with value in $\mathbb{R}$. We shall focus on the one-dimensional problem and derive the bias associated to $\tau\circ\bar{T}(Q)$ and the mean square error of $\tau\circ T(P_n)$. Theorem 3.1(i) of KOE derives the asymptotic minimax lower bound of any estimator $\tau\circ T_a$ of $\tau(\theta^*)$ where T_a is a Fisher consistent and regular estimator of θ^*. They establish under some regularity conditions that for each $r > 0$,

$$\liminf_{n\to\infty} \sup_{Q\in B_H(P_*,r/\sqrt{n})} n(\tau\circ T_a(Q) - \tau(\theta^*))^2 \geq 4r^2 B^*,$$

with

$$B^* = \left(\frac{\partial\tau(\theta^*)}{\partial\theta}\right)' \Sigma \left(\frac{\partial\tau(\theta^*)}{\partial\theta}\right). \tag{19}$$

The asymptotic minimax lower bound for the square bias is then $4r^2B^*$ which is reached by the functional determining the minimum Hellinger distance (HD) estimator. Our next result establishes that the square bias of $\bar{T}(Q)$, the functional associated with ETHD, also reaches this bound. This is an essential step towards the derivation of the limit mean square error of the ETHD estimator $\tau\circ T(P_n)$. We make the following assumptions:

Assumption 3. *(i)* $\{X_i : i = 1,\ldots,n\}$ *is a sequence of i.i.d. random vectors distributed as* X.

(ii) Θ *is compact and* $\theta^* \in int(\Theta)$ *is a unique solution to* $E_{P_*}(g(X,\theta)) = 0$.

(iii) $g(x,\theta)$ *is continuous over* Θ *at each* $x \in \mathcal{X}$.

(iv) $E_{P_*}(\sup_{\theta\in\Theta}\|g(X,\theta)\|^\alpha) < \infty$ *for some* $\alpha > 2$, *and there exists a neighborhood* $\mathcal{N}$ *of* θ^* *such that* $g(x,\theta)$ *is twice continuously differentiable over* $\mathcal{N}$ *at each* $x \in \mathcal{X}$, *such that:* $\sup_{x\in\mathcal{X}_n,\theta\in\mathcal{N}} \left\|\frac{\partial g(x,\theta)}{\partial\theta'}\right\| = o(n^{1/2})$, *and there exists a measurable function* $d(X)$ *such that* $E_{P_*}(d(X)) < \infty$ *and*

$$\max\left(\sup_{\theta\in\mathcal{N}}\|g(X,\theta)\|^4,\ \sup_{\theta\in\mathcal{N}}\left\|\frac{\partial g(X,\theta)}{\partial\theta'}\right\|^2\right) \leq d(X).$$

(v) $G = E_{P_*}(\partial g(X,\theta^*)/\partial\theta')$ *has full column rank and* $Var_{P_*}(g(X,\theta))$ *is nonsingular for all* $\theta \in \Theta$ *with smallest eigenvalue* $\underline{\ell}$ *bounded away from 0.*

(vi) $\{m_n\}_{n\geq 0}$ *satisfies* $m_n \propto n^a$ *with* $1/\alpha < a < 1/2$.

(vii) Let $\mathbf{a}_n(\lambda,\theta) = \exp(\lambda' g_n(X,\theta))$, $\mathbf{a}(\lambda,\theta) = \exp(\lambda' g(X,\theta))$. $E_{P_*}(\mathbf{a}(\lambda,\theta))$ *is continuous in* (λ,θ) *over* $\Lambda\times\Theta$ *and, for any* $r>0$ *and any sequence* $Q_n \in B_H(P_*, r/\sqrt{n})$,

$E_{Q_n}(\mathbf{a}_n(\lambda,\theta))$ *converges to* $E_{P_*}(\mathbf{a}(\lambda,\theta))$, *uniformly over* $\Lambda\times\Theta$, *with* Λ *a convex and compact subset of* $\mathbb{R}^m$ *with interior containing 0.*

In addition, there exists a neighborhood $\mathcal{V}$ *of 0 such that* $E_{P_*}\left(\sup_{(\lambda,\theta)\in\mathcal{V}\times\mathcal{N}} \mathbf{a}(\lambda,\theta)\right) < \infty$ *and* $E_{Q_n}[g_n(X,\theta)\mathbf{a}_n(\lambda,\theta)]$, $E_{Q_n}[g_n(X,\theta)g_n(X,\theta)'\mathbf{a}_n(\lambda,\theta)]$, $E_{Q_n}\left[g_n(X,\theta)\left(\frac{\partial g_{n,k}(X,\theta)}{\partial\theta_l}\right)\mathbf{a}_n(\lambda,\theta)\right]$ *converge uniformly over* $\mathcal{V}\times\mathcal{N}$ *to* $E_{P_*}[g(X,\theta)\mathbf{a}(\lambda,\theta)]$, $E_{P_*}[g(X,\theta)g(X,\theta)'\mathbf{a}(\lambda,\theta)]$, $E_{P_*}\left[g(X,\theta)\left(\frac{\partial g_k(X,\theta)}{\partial\theta_l}\right)\mathbf{a}(\lambda,\theta)\right]$, *respectively, for* $k=1,\ldots,m$, $l=1,\ldots,p$.

(viii) τ *is continuously differentiable at* θ^*.

Assumptions 3*(i)-(vi)* and *(viii)* are the assumptions of KOE under which the local robustness property of HD is established. Similar to Assumption 1*(vi)*, Assumption 3*(vii)* is useful here because ETHD is determined by two separate optimization procedures as opposed to HD which is a saddle point estimator. It is not hard to establish that this assumption holds if $g(\cdot,\cdot)$ is bounded. It is also worthwhile to mention that one can do away with it if the optimization set for λ is set to Λ_n, a convex and compact neighborhood of 0 that shrinks at a rate slightly slower than $O(1/\sqrt{n})$, as discussed in Section 3.

The next result shows that $\tau\circ\bar{T}$ is Fisher consistent and that its worst square bias - when the data is distributed as Q in a suitable Hellinger-neighborhood of P_* - is equal (in the limit) to the lower bound derived by KOE.

Theorem 4.1. *Under Assumption 3, the mapping* $\bar{T}$ *is Fisher consistent and satisfies:*

$$\lim_{n\to\infty} \sup_{Q\in B_H(P_*, r/\sqrt{n})} n(\tau\circ\bar{T}(Q) - \tau(\theta^*))^2 = 4r^2 B^*, \tag{20}$$

for each $r>0$, *with* B^* *given by* (19).

The limit provided for the bias in Theorem 4.1 is useful to study the mean square error of ETHD $\hat{\theta}$. Recall that, by definition, $\hat{\theta} = T(P_n)$ as given by (15). The following result derives the asymptotic worst mean square error of $\tau\circ T(P_n)$ for the estimation of $\tau(\theta^*)$. The supremum of mean square error is taken over possible distributions Q of the data lying in the Hellinger ball centered at P_* with radius $r/\sqrt{n}$ and with respect to which the estimation function $g(X,\theta)$ has moments up to α. Let

$$\bar{B}_H^{\delta}(P_*, r/\sqrt{n}) = B_H(P_*, r/\sqrt{n}) \cap \left\{ Q\in\mathcal{M} : E_Q\left(\sup_{\theta\in\Theta}\|g(X,\theta)\|^{\alpha} \leq \delta < \infty\right)\right\},$$

with $r>0$ and $\delta>0$ and let $Q^{\otimes n}$ denote the joint distribution of n independent copies of X, with X distributed as Q. We have the following result.

Theorem 4.2. *If Assumption 3 holds, the mapping T is Fisher consistent and regular, and the ETHD estimator, $\hat{\theta} = T(P_n)$, satisfies:*

$$\lim_{b\to\infty}\lim_{\delta\to\infty}\lim_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int b\wedge n(\tau\circ T(P_n)-\tau(\theta^*))^2 dQ^{\otimes n} = (1+4r^2)B^*$$

for each $r>0$, with B^* given by (19).

Fisher consistency and regularity of the functional T ensure, from Theorem 3.2(i) of KOE, that $(1+4r^2)B^*$ is the minimum of the limit expressed in the theorem. The fact that equality holds establishes that ETHD is asymptotically minimax robust with respect to the mean square error of $\tau\circ T(P_n)$ estimating $\tau(\theta^*)$.

Following KOE, we can also consider a more general class of loss functions and explore the asymptotic risk associated to the estimation of $\bar{T}(Q)$. Let ℓ be a loss function satisfying the following assumption.

Assumption 4. *The loss function $\ell:\bar{\mathbb{R}}^p\to[0,\infty]$ is (i) symmetric subconvex (i.e., for all $z\in\mathbb{R}^p$ and $c\in\mathbb{R}$, $\ell(z)=\ell(-z)$ and $\{z\in\mathbb{R}^p:\ell(z)\leq c\}$ is convex); (ii) upper semicontinuous at infinity; and (iii) continuous on $\bar{\mathbb{R}}^p$.*

We can state the following result.

Theorem 4.3. *If Assumptions 3 and 4 hold, then the mapping T is Fisher consistent and the ETHD estimator, $\hat{\theta} = T(P_n)$, satisfies:*

$$\lim_{b\to\infty}\lim_{\delta\to\infty}\lim_{r\to\infty}\lim_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int b\wedge\ell\left(\sqrt{n}(\tau\circ T(P_n)-\tau\circ\bar{T}(Q)\right)dQ^{\otimes n} = \int \ell dN(0,B^*),$$

with B^* given by (19).

This theorem shows that, similarly to HD, ETHD is asymptotically minimax risk optimal for a general class of risk functions. Theorem 4.3 specifically shows that the supremum of expected loss under Q associated to the estimation of $\bar{T}(Q)$ by $T(P_n)$ is equal in the limit to the minimum bound established by KOE (2013a, Th. 3.3(i)) for Fisher consistent estimators.

Theorems 4.2 and 4.3 establish ETHD as an alternative to HD when it comes to minimax robustness to local misspecification. The full picture of the properties of ETHD in misspecified models is obtained in the next section where we study the large sample behaviour of this estimator under global misspecification.

5. ETHD under global misspecification

Our main motivation in proposing ETHD is to introduce an estimator that preserves most of the qualities of HD in addition to being robust to global misspecification. The simulation study in

Section 6.1 below reveals that HD is much more affected by global misspecification than ETHD and other standard estimators such as GMM, ET and ETEL. We derive in this section the asymptotic distribution of ETHD under global misspecification. Let

$$R_n(\lambda,\theta) = \begin{pmatrix} R_{\theta,\theta}(\lambda,\theta) & R_{\theta,\lambda}(\lambda,\theta) \\ R_{\lambda,\theta}(\lambda,\theta) & R_{\lambda,\lambda}(\lambda,\theta) \end{pmatrix}$$

be the $(m+p, m+p)$-matrix with components $R_{ab}(\lambda,\theta)$ $(a,b = \theta,\lambda)$ defined by Equation (D.9) in Appendix D and let

$$\boldsymbol{\Delta}_P(\lambda(\theta),\theta) = \frac{E_P\left[\exp(\lambda(\theta)'g(X,\theta)/2)\right]}{\sqrt{E_P\left[\exp(\lambda(\theta)'g(X,\theta))\right]}}, \quad \lambda(\theta) = \arg\min_{\lambda\in\Lambda} E_P\left[\exp(\lambda(\theta)'g(X,\theta))\right],$$

where P is the true probability distribution of the data X. We maintain the following set of regularity assumptions.

Assumption 5. *(Regularity conditions under global misspecification)*
(i) $\{X_i : i = 1,\ldots,n\}$ *is a sequence of i.i.d. random vectors distributed as* X.
(ii) The objective function $\boldsymbol{\Delta}_P(\theta,\lambda(\theta))$ *is maximized at a unique "pseudo-true" value* θ^* *with* $\theta^* \in int(\Theta)$ *and* Θ *compact.*
(iii) $g(x,\theta)$ *is continuous on* Θ *and twice continuously differentiable in a neighborhood* $\mathcal{N}$ *of* θ^* *for almost all* x.
(iv) $E\left(\sup_{\theta\in\Theta,\lambda\in\Lambda}\exp(\lambda' g(X,\theta))\right) < \infty$ *where* Λ *is a compact and convex subset of* $\mathbb{R}^m$ *such that* $\lambda^* \equiv \arg\max_\Lambda -E[\exp(\lambda' g(X,\theta^*))]$ *is interior to* Λ. *Furthermore,* $E\left(\|g(X,\theta^*)\|^4\right)$, $E\left\|\frac{\partial g(X,\theta^*)}{\partial\theta'}\right\|^2$ *and* $E[\exp(4\lambda^{*'} g(X,\theta^*))]$ *are all finite.*
(v) $R_n(\lambda,\theta)$ *converges in probability uniformly in a neighborhood of* (λ^*,θ^*) *with limit* $R(\lambda,\theta)$ *such that* $R \equiv R(\lambda^*,\theta^*)$ *is nonsingular.*

These assumptions are quite standard in the literature on global misspecification. Assumption 5*(ii)* imposes that the population version $\boldsymbol{\Delta}_P$ of the ETHD objective function $\boldsymbol{\Delta}_{P_n}$ is maximized at a unique point θ^* in the parameter space. From Lemma C.1 in Appendix C, the pseudo-true value θ^* is the parameter value that minimizes the Hellinger distance between the true distribution P of the data and the family of distributions consistent with the moment restrictions and closest to P in terms of Kullback-Leibler divergence. It is essential to maintain the uniqueness of the pseudo-true value to ensure the convergence of ETHD. In this paper, we do not address the case where the uniqueness condition fails, and save it for future extensions. $R_n(\lambda,\theta)$ is the first-order term appearing in the mean-value expansion of the first-order condition in θ and λ of the two optimization programs leading to ETHD, namely: $\max_\theta \boldsymbol{\Delta}_{P_n}\left(\hat{\lambda}(\theta),\theta\right)$ and (13). The non-singularity condition in Assumption 5*(v)* amounts to the first-order local identification condition in correctly specified moment condition models. We have the following result.

Theorem 5.1. *(Asymptotics under global misspecification)*
Under regularity assumption 5, we have

$$\sqrt{n}\begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\lambda} - \lambda^* \end{pmatrix} \xrightarrow{d} N\left(0, R^{-1}\Omega^* R^{-1}\right) .$$

with $\hat{\lambda} \equiv \hat{\lambda}(\hat{\theta})$ *(see Equation* (13)*), and* R *and* Ω^* *explicitly defined in the proof in Appendix D.*

This result shows that ETHD is asymptotically centered around its pseudo-true value θ^* (as defined in Assumption 5*(ii)*) and that it is $\sqrt{n}$-convergent and asymptotically normal under global misspecification. Of course, the pseudo-true value, as the probability limit of ETHD, corresponds to the true parameter value when the model is actually correctly specified and in this case, the tilting parameter value λ^* is 0; see Theorem 3.3. The above result is similar to Theorem 10 in Schennach (2007) for ETEL and to Theorem 6.1 in Smith (2007) for GEL-EL.

As discussed in Section 2.2, an estimator is said to be robust to global misspecification when its asymptotic distribution derived under global misspecification coincides with its asymptotic distribution under correct specification. In Appendix D, we also show that the above asymptotic variance corresponds to the one of Theorem 3.3 under correct specification. This means that ETHD is robust to global misspecification. ETHD is therefore the first estimator that is efficient under correct specification, and robust to both global and local misspecification.

6. Monte Carlo simulations

In this section, we report some simulation results that illustrate the finite sample properties of the estimators considered in this paper. First, we consider simulation designs that display settings of correct specification and global misspecification. These experiments confirm the lack of robustness of HD under global misspecification and also confirm that, like ETEL, ETHD is robust to global misspecification. The second set of simulations focus mainly on designs that display local misspecification, or slight perturbations - contamination - in the observed data. The results show that ETHD and HD display about the same root mean square error and underscore the local robustness properties of ETHD established in the previous section.

6.1. Experiment 1: estimation of a population mean.

We use the experimental design suggested in Schennach (2007), where we wish to estimate the mean while imposing a known variance. The moment condition model consists of two restrictions:

$$E\left(g(X_i, \theta)\right) \equiv E\left[X_i - \theta \quad (X_i - \theta)^2 - 1\right]' = 0\,,$$

where X_i is drawn from either a correctly specified model C, or a misspecified model M, with

$$\begin{array}{ll} X_i \sim N(0,1) & \text{(for Model C)} \\ X_i \sim N(0, s^2) & \text{(with } 0.72 \leq s < 1 \text{ for Model M).} \end{array}$$

The estimators that we consider for θ are: the 2-step GMM (we use the identity weighting matrix for the first step GMM estimation), HD, EL, ET, EEL (the Euclidean Empirical Likelihood, also known as the continuous updated GMM), ETEL and ETHD. Under Model C, the true parameter value is $\theta^* = 0$. Under Model M, the pseudo-true value for each estimator listed above is $\theta^* = 0$ as well. As explained by Schennach (2007), the equality of true value and pseudo-true values is useful to have a meaningful comparison of simulated variances.

Table 2 displays the simulated standard deviations of the considered estimators for sample sizes of 1,000, 5,000 and 10,000 over 10,000 replications. Under correct specification, all the estimators perform equally well, as expected since all the estimators share the same asymptotic distribution. Indeed, the sample sizes considered here are large enough for the asymptotic approximation to be quite accurate. The $\sqrt{n}$-convergence rate under correct specification of all estimators is noticeable by the fact that, as the sample size increases from 1,000 to 5,000, their respective simulated standard deviations shrink by the ratio of $\sqrt{5}$, and by the ratio of $\sqrt{2}$ when the sample size doubles from 5,000 to 10,000.

Under global misspecification, these estimators show different patterns. For $s = 0.75$, we can see that ETHD, ETEL, ET and GMM all have their standard deviations shrinking with increasing sample size whereas those of HD and EL do not shrink although HD is better among the two with smaller standard deviations.

Figure 1 shows the ratio of standard deviations for sample sizes 1,000, 5,000 and 10,000 over a grid of misspecification parameters s. As s moves farther away from 1, the ratios of standard deviations seems to depart from their reference levels - $\sqrt{5}$, $\sqrt{10}$, and $\sqrt{2}$, respectively for the three graphs in display - first for EL followed by HD. All the other estimators have their ratios significantly closer to reference with EEL looking the most stable followed by ET, GMM, ETHD and ETEL. ETHD and ETEL have similar range with the ratio of ETHD slightly closer to the reference than that of ETEL.

Figure 2 displays the cumulative distribution of ETHD, ETEL and HD for the three sample sizes and $s = 1$ and 0.75. The distributions of these estimators, as expected, are undistinguishable under correct specification while, under misspecification, the range of HD does not seem to narrow around 0 in contrast to ETHD and ETEL. The difference between the latter two seems to merely reflect the difference in their respective standard deviations. Overall, our proposed estimator ETHD performs very well both under correct specification and global misspecification.

6.2. **Experiment 2: estimation of a production function.** We now turn our attention to the estimation of a production function, a primitive component of many economic models. For example, the estimation of production functions plays a key role in the empirical analysis of issues such as the contribution of different factors to economic growth, the degree of complementarity and substitutability between inputs, estimation of economies of scale and economies of scope, evaluation of the effects of new technologies, among many others; see e.g. Aguirregabiria, 2018. There are at least three well-known issues that arise when estimating productions functions: (a) data problems, such as measurement

error in output and/or inputs; (b) specification problems, such as functional form assumptions; (c) simultaneity/endogeneity. All these issues can yield misspecification.

We consider the Cobb-Douglas production function (Cobb and Douglas, 1928) with constant returns to scale and estimate the productivity of labour function. This amounts to the following non-linear regression model of production per capita y on capital per capita k: $y = k^{\theta_0} + error$, where θ_0 is the technological parameter or the output elasticity of capital. More specifically, our data generating process is:

$$y_i = k_i^{\theta_0} + u_i \quad \text{with} \quad u_i = \rho \cdot (k_i - \mu_k)/\sigma_k + \varepsilon_i \,, \tag{21}$$

with $k_i \sim \text{IID}|t_4|$ (folded-t distribution with 4 degrees-of-freedom), $\varepsilon_i \sim \text{NID}(0,1)$, and

$$\mu_k = E(k_i) = \frac{2a_\nu}{\nu - 1}\sqrt{\frac{\nu}{\pi}} \quad \text{and} \quad \sigma_k^2 = Var(k_i) = \frac{(2\nu a_\nu)^2}{\pi(\nu-2)(\nu-1)^2}$$
$$\text{where} \qquad a_\nu = \frac{\Gamma(\nu+1)/2}{\Gamma(\nu/2)} \,.$$

Throughout, our parameter of interest is $\theta_0 = 0.3$. To estimate θ_0, we rely on the following moment restrictions:

$$E\left[Z(k_i)(y_i - k_i^\theta)\right] = 0. \tag{22}$$

where $Z(k_i) \in \mathbb{R}^m$ is the vector of m instruments. We consider $m = 2, 3, 5, 8,$ and 10 with the

$$m = 2, \quad Z(k_i) = (1, k_i^{0.5})';$$

$$m = 3, \quad Z(k_i) = (1, k_i^{0.5}, k_i^{0.75})';$$

associated vector of instruments, $\quad m = 5, \quad Z(k_i) = (1, k_i^{0.2}, k_i^{0.5}, k_i^{0.6}, k_i^{0.75})';$

$$m = 8, \quad Z(k_i) = (1, k_i^{0.2}, k_i^{0.3}, k_i^{0.4}, k_i^{0.5}, k_i^{0.6}, k_i^{0.75} k_i^{0.8})'$$

$$m = 10, \quad Z(k_i) = (1, k_i^{0.1}, k_i^{0.2}, k_i^{0.3}, k_i^{0.4}, k_i^{0.5}, k_i^{0.6}, k_i^{0.75} k_i^{0.8}, k_i^{0.9})'.$$

Finally, the parameter ρ controls the degree of endogeneity of k_i: the moment restrictions in (22) are correctly specified when $\rho = 0$, and incorrect otherwise. The instruments are chosen above as powers of k_i to guarantee that the estimating function has at least a second moment under misspecification ($\rho \neq 0$).

• **Study under global misspecification:** We set $\rho = 0.2$. We consider a sample size equal to $n = 50$, and a growing number of moment restrictions with $m = 2, 3, 5, 8,$ and 10.

The pseudo-true values for each estimator have been obtained by solving the population version of their respective optimization functionals and are displayed in Table 1; in the sequel, we rely on numerical integration[3].

[3]The probability density function of the folded t-distribution with ν degree-of-freedom ($|t_\nu|$) is

$$f_{|t_\nu|}(x) = \frac{a_\nu}{\sqrt{\pi\nu}}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} 1(x \geq 0).$$

m	GMM	MHDE	EL	ET	EEL	ETEL	ETHD
2	0.4572	0.45773	0.45175	0.45775	0.45775	0.45725	0.45725
3	0.4584	0.454	0.451625	0.4598125	0.4598125	0.459875	0.459875
5	0.4579	0.4615	0.46	0.458	0.4585	0.4585	0.4585
8	0.4859	0.457625	0.4575	0.45925	0.45775	0.45925	0.45925
10	0.4858	0.4585	0.4538	0.4578	0.4581	0.4578	0.4578

TABLE 1. Pseudo-true values of different estimators under global misspecification

Overall, the pseudo-true values (PTV hereafter) of all the estimators remain close to each other, and do not vary too much with the number of moment restrictions used for the estimation: in particular, the PTV of ET, ETEL and ETHD are identical in almost all cases and display little variation when m changes.

In Figure 3, we display the RMSE and Bias per PTV[4] for different values of m. First and foremost, many estimators have their RMSE and Bias that are very close to one another: notable exceptions include, EEL with RMSE significantly larger than the others throughout; GMM with Bias slightly smaller than the others. In addition, the RMSE remain fairly stable as the number of moment increases, while the Bias decreases slightly[5]. Finally, it is worth pointing out that our estimator ETHD's RMSE is smaller then HD's for all cases with $m > 2$, and smaller than ET's for all cases except $m = 8$.

• **Study under local misspecification with local-to-zero endogeneity:** We consider (slight) perturbations in the probability measure that generates the observations by using values of ρ that are local to 0: specifically, $\rho_{n,j} = a_j/\sqrt{n}$ with $a_j \in \{-1.0, -0.9, \ldots, 0, 0.1, 0.2, \ldots, 1.0\}$. Our sample size is set to $n = 50$ and we rely on $m = 5$ moment restrictions for the estimation of θ_0.

In Figure 4, we display the standard deviations, RMSE, Bias, as well as the probabilities[6] $Pr(|\hat{\theta} - \theta_0| > 0.1)$ as a function of $\rho_{n,j}$. Our estimator ETHD is the second best according to each criterion (behind GMM): in particular, it is worth pointing out that it clearly outperforms ET in terms of standard deviation and RMSE, while it also (though modestly) outperforms HD in terms of standard deviation and RMSE and ETEL in terms of Probas. Overall, our estimator ETHD is relatively well-behaved.

• **Study under local misspecification from partial contamination:** In this experiment, we assess how deviations from the data generating process (21) by part of the observed sample (rather than the whole sample) affect the properties of the estimators. We set $\rho = 0$ and induce deviations

[4]Since the PTV of the different estimators are in general different, we report RMSE/PTV and Bias/PTV to allow meaningful comparisons between estimators.

[5]When $m = 10$, the Bias of all estimators tend to increase slightly, but with a small sample size of $n = 50$, we are not too concerned about it.

[6]These probabilities can be interpreted as a measure of stability of each estimator around the true value, as they account for large deviations away from it.

by replacing ε_i in (21) by

$$\tilde{\varepsilon}_i = \begin{cases} \varepsilon_i & \text{with probability } 1-\pi \\ \varepsilon_i + cw_i & \text{with probability } \pi, \end{cases}$$

where $w_i = r \cdot \varepsilon_i + \sqrt{1-r^2} \cdot \xi_i$, with ξ_i independent of ε_i, and iid with distribution either χ^2_1, or $t_{1.5}$. When $\xi_i \sim \chi^2_1$, the moment condition (22) is violated by the related part of the sample. In the case where $\xi_i \sim t_{1.5}$, the moment restrictions are still satisfied across the whole sample, but the estimating function does not have second moments for the related part of the sample.

We consider $c = 0$ (no contamination) and $c = 0.5, 1.0, 2.0$; $r = 0.0$ and 0.5; and $\pi = 0.05$ (small contamination) and $\pi = 0.5$ (large contamination). Once again our sample size is set to $n = 50$ and we rely on $m = 5$ moment restrictions for the estimation.

In Table 3, we display the RMSE, Probas, Bias and Standard deviation for all the estimators and all the above-mentioned cases. Our estimator is overall very well-behaved. In particular, ETHD display standard deviations, RMSE and Probas that are smaller than ET and HD throughout: the largest differences are observed when contaminated data are generated with $c = 2$. For example, with $r = 0.5$, $\pi = 0.5$ and contaminated data generated with χ^2_1, the standard deviation of ETHD is 0.575 while ET and HD's standard deviations are respectively 1.535 and 0.730. In addition, ETHD also dominates ETEL when focusing on small contaminations with $\pi = 0.05$, though the differences remain modest throughout.

To conclude this section, our simulation results on local misspecification have some connection with the work of Lindsay (1994) that is worth highlighting. In a fully parametric framework, Lindsay (1994) has shown that minimum power divergence estimators with positive index a (see Equation (5)) entail large second-order bias in their so-called *residual adjustment function* that prevent them to show some robustness property while efficient, whereas those estimators with negative index have some robustness feature in addition to being efficient. Even though our framework in this paper is semiparametric (based on moment condition models), Lindsay's results seem to be confirmed for EEL which, with index $a = 1$, appears to be the less robust among the simulated estimators. The closeness of the RMSE performance of the other estimators is also in line with Lindsay (1994) since they all have non-positive index. Of course, our results in Section 4 and those of KOE (2013a) predict a better performance from ETHD and HD as we observed in these experiments.

7. Conclusion

In this paper, we consider moment condition models that may be suffering from two complementary types of misspecification often present in economic models, global and local misspecification.

Our first contribution is to show that the recent minimum Hellinger distance estimator (HD) proposed by KOE is not well-behaved under global misspecification. More specifically, despite desirable properties under correct specification and local misspecification, HD does not remain root-n consistent

when the model is misspecified when the functions defining the moment conditions are unbounded (even when their expectations are bounded).

Our second contribution is to propose a new estimator that is not only semiparametrically efficient under correct specification, but also robust to both types of misspecification - a desirable property since the extent and nature of the misspecification is always unknown in practice. Our estimator is obtained by combining exponential tilting (ET) and HD - so-called ETHD - and we show that it retains the advantages of both. ETHD is semiparametrically efficient under correct specification, and it remains asymptotically normal with the same rate of convergence when the model is globally misspecified. In addition, we show that it is asymptotically minimax robust to local misspecification.

Our third contribution is to document the finite sample properties of a variety of inference procedures under correct specification, as well as under local and global misspecification through a series of Monte-Carlo simulations. Overall, ETHD consistently performs very well and is competitive under most - if not all - simulation designs.

References

Aguirregabiria, V. (2018). 'Empirical Industrial Organization: Models, Methods, and Applications', Discussion paper, Department of Economics, University of Toronto, http://www.individual.utoronto.ca/vaguirre/courses/eco2901/teaching_io_toronto.html.

Almeida, C., and Garcia, R. (2012). 'Assessing Misspecified Asset Pricing Models with Empirical Likelihood Estimators', *Journal of Econometrics*, 170: 519–537.

——— (2017). 'Economic Implications of Nonlinear Pricing Kernels', *Management Science*, 63(10): 3361–3380.

Andrews, D., and Cheng, X. (2012). 'Estimation and inference with weak, semi-strong, and strong identification', *Econometrica*, 80: 2153–2211.

Antoine, B., Bonnal, H., and Renault, E. (2007). 'On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood', *Journal of Econometrics*, 138: 461–487.

Antoine, B., Proulx, K., and Renault, E. (2018). 'Smooth minimum distance for robust inference in conditional asset pricing models', *Journal of Financial Econometrics*.

Ashley, R. (2009). 'Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis', *Journal of Applied Econometrics*, 24: 325–337.

Beran, R. (1977a). 'Minimum Hellinger distance estimates for parametric models', *Annals of Statistics*, 5: 445–463.

——— (1977b). 'Robust location estimates', *Annals of Statistics*, 5: 431–444.

Bickel, P., Klassen, C., Ritov, Y., and Wellner, J. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Press, Baltimore.

Bravo, F. (2014). 'Semiparametric generalized estimating equations in misspecified models', in M. Akritas, S. Lahiri, and D. Politis (eds.), *Topics in Nonparametric Statistics. Springer Proceedings in Mathematics & Statistics*, vol. 74, pp. 43–52. Springer, New York, NY.

Broniatowski, M., and Keziou, A. (2012). 'Divergences and duality for estimation and test under moment condition models', *Journal of Statistical Planning an Inference*, 142(9): 2554–2573.

Cheng, X., Liao, Z., and Shi, R. (2016). 'An averaging GMM estimator robust to misspecification', *Working paper, PIER WP15-017*.

Cobb, C., and Douglas, P. (1928). 'A Theory of Production', *American Economic Review*, 18-1: 139–165.

Congley, T., Hansen, C., and Rossi, P. (2012). 'Plausibly exogenous', *Review of Economics and Statistics*, 94: 260–272.

Corcoran, S. (1998). 'Bartlett Adjustment of Empirical Discrepancy Statistics', *Biometrika*, 85: 965–972.

Dovonon, P. (2016). 'Large sample properties of the three-step Euclidean likelihood estimators under model misspecification', *Econometric Reviews*, 35: 465–514.

Feinberg, E. A., Kasyanov, P. O., and Voorneveld, M. (2014). 'Berges maximum theorem for noncompact image sets', *Journal of Mathematical Analysis and Applications*, 413: 1040–1046.

Feinberg, E. A., Kasyanov, P. O., and Zadoianchuk, N. V. (2013). 'Berges theorem for noncompact image sets', *Journal of Mathematical Analysis and Applications*, 397: 255–259.

Gallant, A. R., and White, H. (1988). *A unified theory of estimation and inference in nonlinear dynamic models*. Blackwell, Oxford.

Gospodinov, N., Kan, R., and Robotti, C. (2014). 'Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors', *Review of Financial Studies*, 27: 2139–2170.

Gospodinov, N., and Maasoumi, E. (2017). 'General aggregation of misspecified asset pricing models', *Working paper 2017-10*.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). 'Pseudo maximum likelihood methods: Theory', *Econometrica*, 52: 681–700.

Guggenberger, P. (2008). 'Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator', *Econometric Reviews*, 27: 526–541.

——— (2012). 'On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption', *Econometric Theory*, 28: 387–421.

Hall, A. R., and Inoue, A. (2003). 'The Large sample behaviour of the generalized method of moments estimator in misspecified models', *Journal of Econometrics*, 114: 361–394.

Hansen, L. P. (1982). 'Large sample properties of generalized method of moments estimators', *Econometrica*, 50: 1029–1054.

Hansen, L. P., and Jagannathan, R. (1997). 'Assessing specification errors in stochastic discount factor models', *Journal of Finance*, 52: 557–590.

Jing, B.-Y., and Wood, A. (1996). 'Exponential Empirical Likelihood is not Bartlett Correctable', *Annals of Statistics*, 24: 365–369.

Kan, R., Robotti, C., and Shanken, J. (2013). 'Pricing model performance and the two-pass cross-sectional regression methodology', *Journal of Finance*, 68: 2617–2649.

Kitamura, Y. (2000). 'Comparing misspecified dynamic econometric models using nonparametric likelihood', Discussion paper, University of of Wisconsin.

——— (2007). 'Empirical likelihood methods in econometrics: theory and practice', in R. Blundell, W. Newey, and T. Persson (eds.), *Advances in Economics and Econometrics: Theory and Application: Ninth World Congress of the Econometric Society*, vol. 3. Cambridge University Press, Cambridge, UK.

Kitamura, Y., Otsu, T., and Evdokimov, K. (2009). 'Robustness, infinitesimal neighborhoods, and moment restrictions', Discussion Paper 1720, Cowles Foundation for Research in Economics, Yale University.

——— (2013a). 'Robustness, infinitesimal neighborhoods, and moment restrictions', *Econometrica*, 81: 1185–1201.

——— (2013b). 'Supplement to "Robustness, infinitesimal neighborhoods, and moment restrictions"', *Econometrica Supplemental Material*, 81, http://www. econometricsociety.org/ecta/supmat/8617_proofs.pdf.

Kitamura, Y., and Stutzer, M. (1997). 'Efficiency versus robustness: The case for minimum Hellinger distance and related methods', *Econometrica*, 65: 861–874.

Kleibergen, F. (2005). 'Testing parameters in GMM without assuming that they are identified', *Econometrica*, 73: 1103–1124.

Kolesar, M., Chetty, R., Friedman, J., Glaeser, E., and Imbens, G. (2015). 'Identification and inference with many invalid instruments', *Journal of Business Economics and Statistics*, 33: 474–484.

Lindsay, B. (1994). 'Efficiency versus robustness: The case for minimum Hellinger distance and related methods', *Annals of Statistics*, 22: 1081–1114.

Maasoumi, E. (1990). 'How to live with misspecification if you must', *Journal of Econometrics*, 44: 67–86.

Maasoumi, E., and Phillips, P. C. B. (1982). 'On the behavior of inconsistent instrumental variable estimators', *Journal of Econometrics*, 19: 183–201.

Magnus, J. R., and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, Chichester, 2nd edition edn.

Markatou, M. (2007). 'Robust statistical inference: weighted likelihoods or usual m-estimation?', *Communications in Statistics - Theory and Methods*, 25: 2597–2613.

Nevo, A., and Rosen, A. (2012). 'Identification with imperfect instruments', *Review of Economics and Statistics*, 93: 659–671.

Newey, W. K. (1990). 'Semiparametric Efficiency Bounds', *Journal of Applied Econometrics*, 5: 99–135.

Newey, W. K., and McFadden, D. L. (1994). 'Large sample estimation and hypothesis testing', in R. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp. 2113–2247. Elsevier Science Publishers, Amsterdam, The Netherlands.

Newey, W. K., and Smith, R. J. (2004). 'Higher order properties of GMM and generalized empirical likelihood estimators', *Econometrica*, 72: 219–255.

Sandberg, I. W. (1981). 'Global implicit function theorems', *IEEE Transactions on Circuits and Systems*, CS-28: 145–149.

Schennach, S. (2007). 'Point estimation with exponentially tilted empirical likelihood', *Annals of Statistics*, 35: 634–672.

Smith, R. (2007). 'Weak Instruments and Empirical Likelihood: a discussion of the papers by D.W.K. Andrews and J.H. Stock and Y. Kitamura', in R. Blundell, W. Newey, and T. Persson (eds.), *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, vol. 3, chap. 8, pp. 238–260. Cambridge: Cambridge University Press.

Stock, J., and Wright, J. (2000). 'GMM with weak identification', *Econometrica*, 68: 1055–1096.

White, H. (1982). 'Maximum likelihood estimation of misspecified models', *Econometrica*, 50: 1–25.

Appendix A. Results of the Monte Carlo study

A.1. **Experiment 1: estimation of a population mean.**

Model C with s = 1.0							
	GMM	HD	EL	ET	EEL	ETEL	ETHD
Sample size $T = 1000$	0.0316	0.0316	0.0316	0.0316	0.0316	0.0316	0.0316
Sample size $T = 5000$	0.0138	0.0138	0.0138	0.0138	0.0138	0.0138	0.0138
Sample size $T = 10000$	0.0097	0.0097	0.0097	0.0097	0.0097	0.0097	0.0097
Model M with s = 0.75							
	GMM	HD	EL	ET	EEL	ETEL	ETHD
Sample size $T = 1000$	0.0488	0.0481	0.0742	0.0331	0.0270	0.0464	0.0407
Sample size $T = 5000$	0.0215	0.0375	0.0731	0.0152	0.0176	0.0257	0.0217
Sample size $T = 10000$	0.0151	0.0374	0.0743	0.0109	0.0082	0.0200	0.0166

Table 2. Experiment 1: Standard deviations of the GMM, HD, EL, ET, EEL, ETEL, ETHD estimators for models C and M (with $s = 0.75$) with 10,000 replications

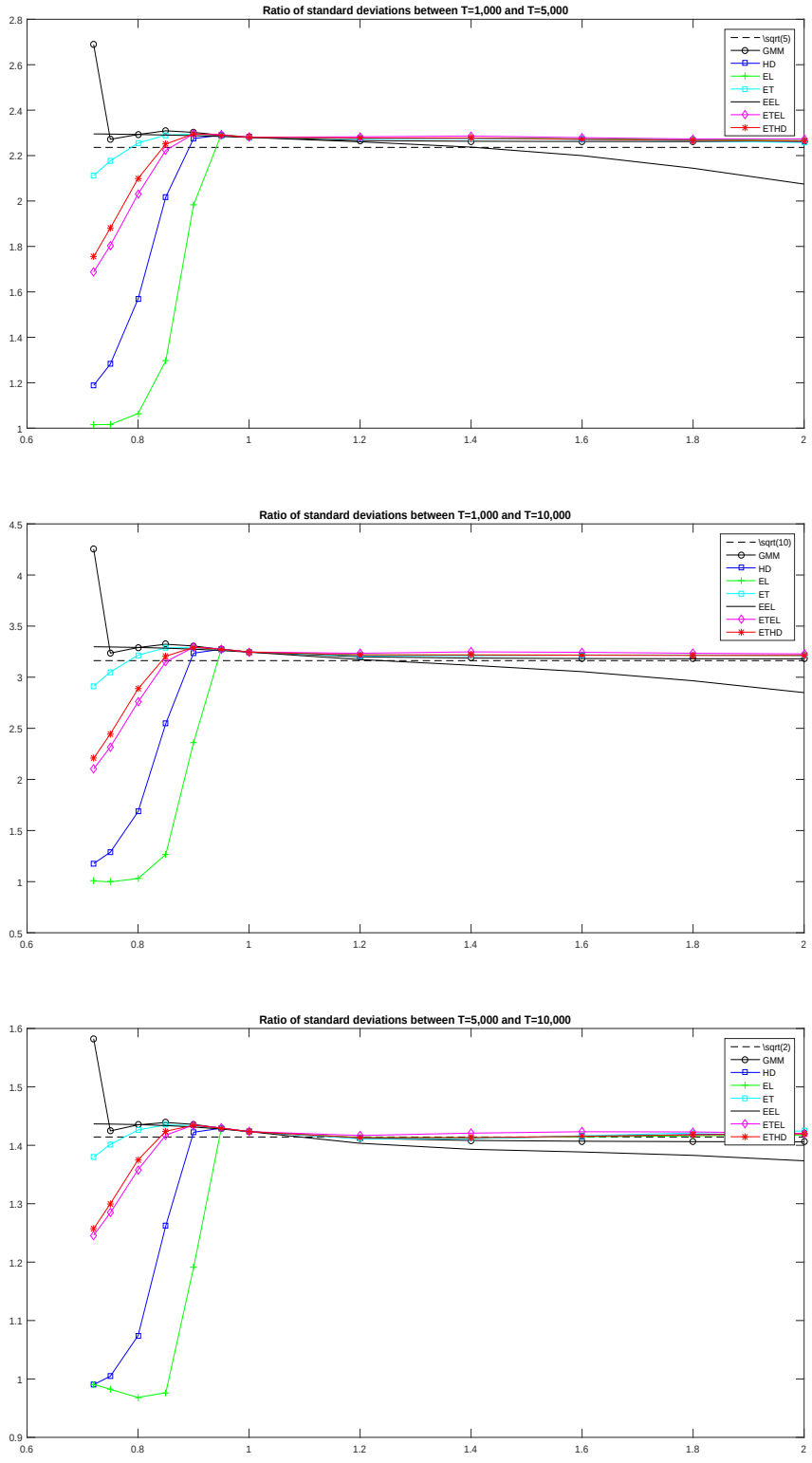

FIGURE 1. Experiment 1: Ratio of standard deviations for sample sizes (i) 1,000 and 5,000; (ii) 1,000 and 10,000; (iii) 5,000 and 10,000 over a grid of misspecification parameters s

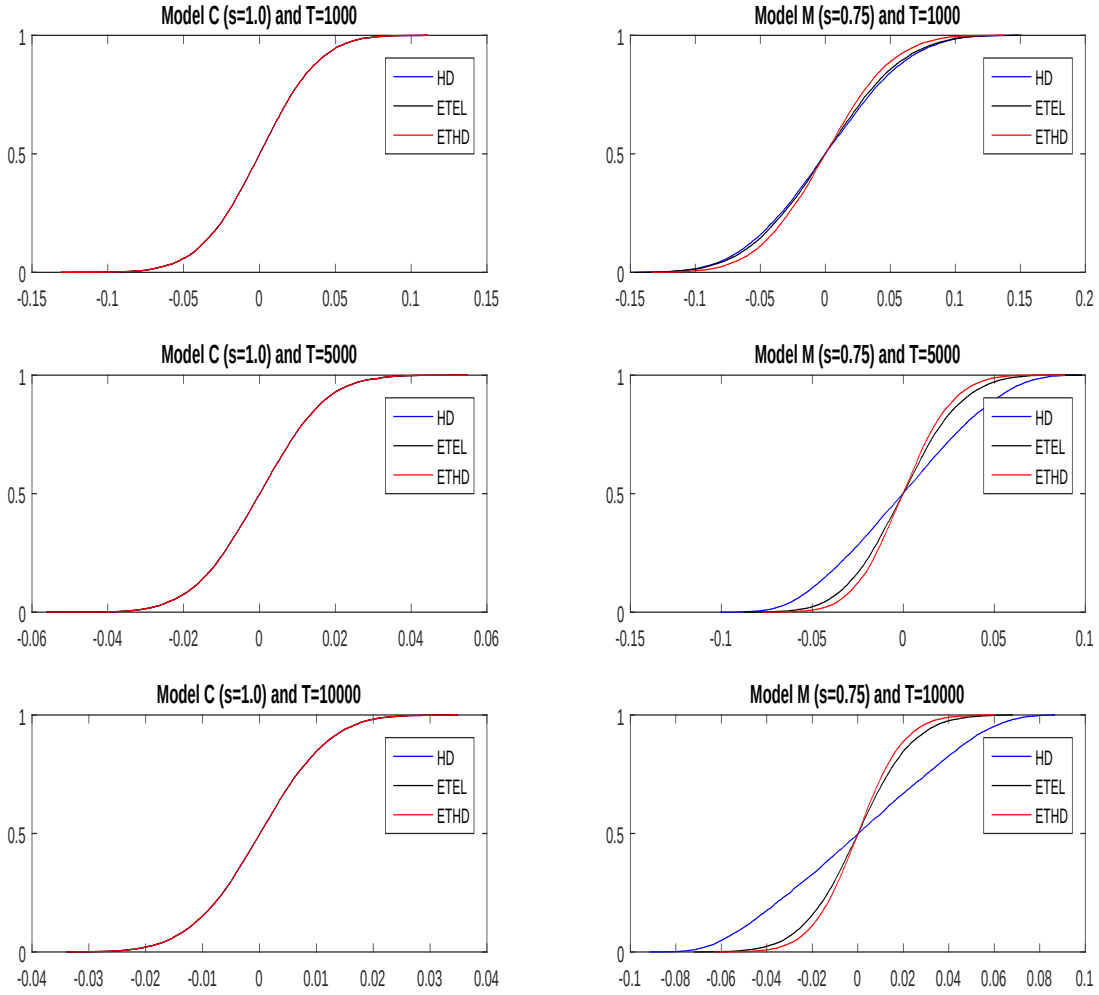

FIGURE 2. Experiment 1: Simulated cumulative distribution of HD, ETEL and ETHD under correct specification (model C with $s = 1.0$) and global misspecification (model M with $s = 0.75$)

A.2. **Experiment 2: estimation of a production function.**

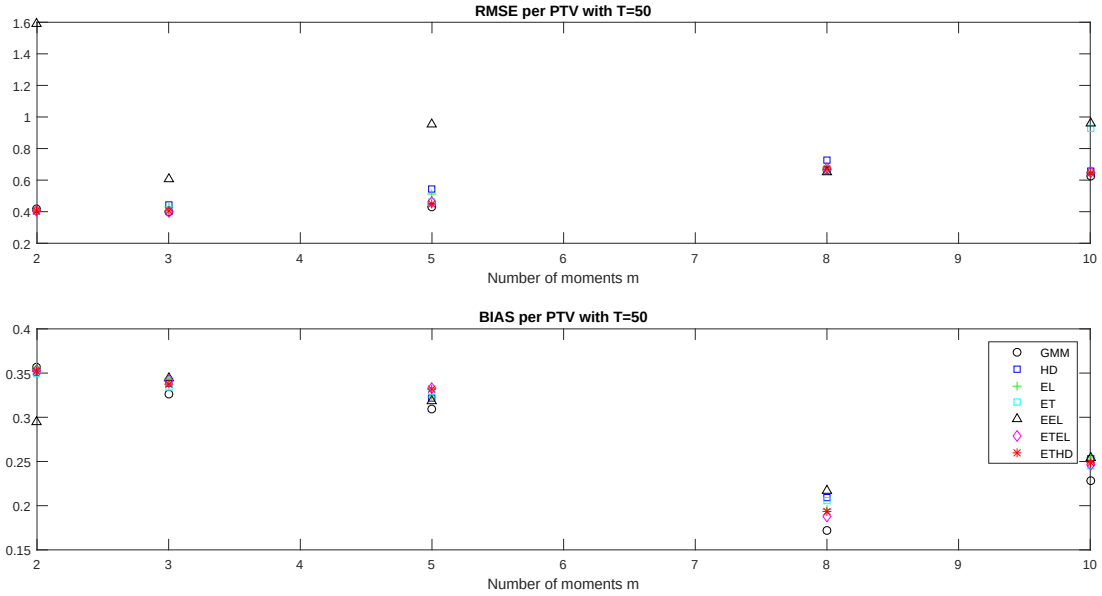

FIGURE 3. Experiment 2 under global misspecification: RMSE and Bias per pseudo-true value for $n = 50$ and different number of moments m.

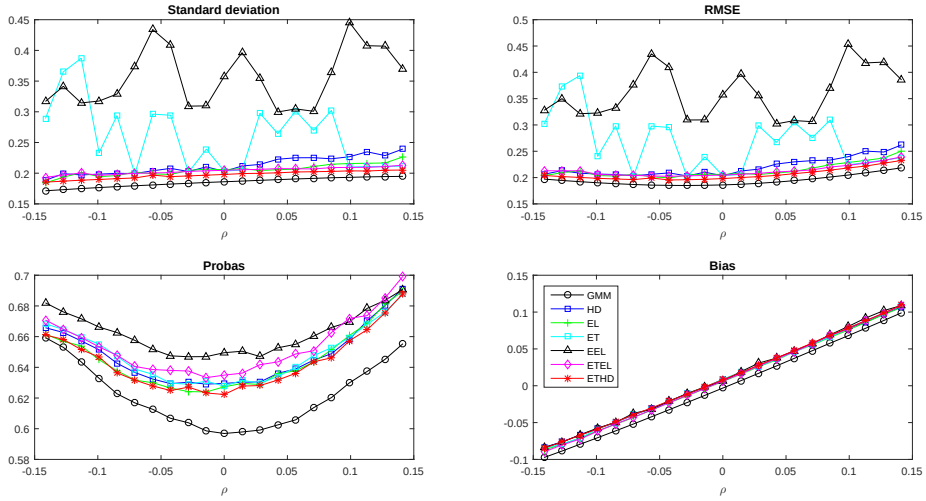

FIGURE 4. Experiment 2 under local misspecification with local-to-zero endogeneity: Standard deviations, RMSE, Bias, and Probas computed as $Pr(|\hat{\theta} - \theta_0| > 0.1)$ for $n = 50$ and $m = 5$ as a function of ρ.

				RMSE							$Pr(\|\hat{\theta}-\theta^0\|>0.1)$						
c	d	r	π	GMM	HD	EL	ET	EEL	ETEL	ETHD	GMM	HD	EL	ET	EEL	ETEL	ETHD
0	0	0	0	0.186	0.205	0.206	0.299	0.574	0.205	0.198	0.597	0.630	0.627	0.628	0.650	0.635	0.622
0.5	1	0	0.05	0.191	0.233	0.209	0.304	0.340	0.209	0.204	0.610	0.634	0.635	0.633	0.649	0.637	0.628
1.0	1	0	0.05	0.197	0.239	0.224	0.215	0.409	0.216	0.208	0.624	0.643	0.644	0.643	0.659	0.647	0.637
2.0	1	0	0.05	0.207	0.262	0.239	0.312	0.463	0.221	0.216	0.643	0.655	0.655	0.654	0.672	0.657	0.649
0.5	1	0	0.05	0.190	0.238	0.219	0.212	0.700	0.210	0.203	0.606	0.626	0.628	0.628	0.646	0.638	0.623
1.0	2	0	0.05	0.194	0.234	0.229	0.217	0.727	0.214	0.208	0.613	0.633	0.637	0.635	0.656	0.645	0.632
2.0	2	0	0.05	0.201	0.256	0.244	0.432	0.802	0.222	0.216	0.619	0.644	0.646	0.644	0.666	0.653	0.638
0.5	1	0.5	0.05	0.192	0.221	0.212	0.298	0.339	0.210	0.203	0.609	0.636	0.634	0.636	0.652	0.635	0.629
1.0	1	0.5	0.05	0.197	0.241	0.227	0.372	0.465	0.215	0.209	0.622	0.645	0.642	0.643	0.659	0.647	0.640
2.0	1	0.5	0.05	0.208	0.268	0.243	0.321	0.448	0.224	0.219	0.643	0.655	0.654	0.656	0.674	0.658	0.651
0.5	2	0.5	0.05	0.191	0.226	0.227	0.213	0.689	0.211	0.204	0.605	0.627	0.629	0.629	0.647	0.641	0.622
1.0	2	0.5	0.05	0.195	0.246	0.236	0.309	0.675	0.216	0.209	0.616	0.633	0.636	0.635	0.653	0.648	0.633
2.0	2	0.5	0.05	0.203	0.260	0.243	0.383	0.630	0.223	0.217	0.625	0.645	0.648	0.644	0.665	0.655	0.638
0.5	1	0	0.50	0.263	0.354	0.327	0.664	0.750	0.254	0.253	0.759	0.739	0.731	0.747	0.771	0.730	0.737
1.0	1	0	0.50	0.359	0.483	0.427	0.905	1.171	0.327	0.451	0.850	0.811	0.804	0.828	0.855	0.796	0.810
2.0	1	0	0.50	0.541	0.764	0.777	1.615	1.986	0.451	0.521	0.926	0.879	0.871	0.895	0.922	0.866	0.879
0.5	2	0	0.50	0.230	0.295	0.284	0.577	0.746	0.253	0.247	0.661	0.690	0.686	0.692	0.708	0.694	0.686
1.0	2	0	0.50	0.288	0.401	0.383	0.546	1.019	0.316	0.312	0.729	0.752	0.750	0.751	0.765	0.753	0.746
2.0	2	0	0.50	0.403	0.604	0.574	1.003	1.338	0.428	0.430	0.797	0.810	0.808	0.816	0.823	0.806	0.801
0.5	1	0.5	0.50	0.267	0.340	0.328	0.624	0.775	0.264	0.343	0.756	0.743	0.739	0.746	0.772	0.734	0.741
1.0	1	0.5	0.50	0.358	0.501	0.466	0.976	1.240	0.343	0.346	0.840	0.815	0.809	0.825	0.846	0.802	0.813
2.0	1	0.5	0.50	0.530	0.743	0.744	1.549	1.805	0.469	0.584	0.926	0.883	0.875	0.889	0.912	0.871	0.887
0.5	2	0.5	0.50	0.238	0.326	0.314	0.536	0.725	0.264	0.256	0.675	0.709	0.701	0.707	0.726	0.702	0.702
1.0	2	0.5	0.50	0.297	0.424	0.396	0.766	1.015	0.323	0.327	0.733	0.763	0.755	0.763	0.775	0.753	0.753
2.0	2	0.5	0.50	0.406	0.619	0.594	1.082	1.413	0.434	0.486	0.807	0.820	0.819	0.826	0.839	0.817	0.816

				Bias							Standard Deviation						
c	d	r	π	GMM	HD	EL	ET	EEL	ETEL	ETHD	GMM	HD	EL	ET	EEL	ETEL	ETHD
0	0	0	0	-0.003	0.008	0.006	0.007	0.015	0.007	0.008	0.186	0.205	0.206	0.299	0.574	0.205	0.198
0.5	1	0	0.05	-0.006	0.004	0.004	0.004	0.005	0.004	0.004	0.191	0.233	0.209	0.304	0.340	0.209	0.204
1.0	1	0	0.05	-0.009	0.002	0.001	0.004	0.005	0.002	0.003	0.196	0.239	0.224	0.215	0.409	0.216	0.208
2.0	1	0	0.05	-0.012	-0.004	-0.004	-0.003	-0.007	-0.001	-0.002	0.207	0.262	0.239	0.312	0.463	0.221	0.216
0.5	1	0	0.05	-0.004	0.005	0.004	0.008	0.020	0.007	0.006	0.190	0.238	0.219	0.212	0.700	0.210	0.203
1.0	2	0	0.05	-0.006	0.007	0.005	0.008	0.022	0.007	0.007	0.194	0.233	0.229	0.216	0.726	0.213	0.208
2.0	2	0	0.05	-0.009	0.005	0.005	0.002	0.018	0.006	0.006	0.201	0.256	0.244	0.432	0.802	0.222	0.216
0.5	1	0.5	0.05	-0.006	0.006	0.003	0.004	0.006	0.004	0.005	0.192	0.221	0.212	0.298	0.339	0.210	0.203
1.0	1	0.5	0.05	-0.008	0.002	0.001	0.000	0.003	0.004	0.003	0.197	0.241	0.227	0.372	0.465	0.215	0.209
2.0	1	0.5	0.05	-0.012	-0.002	-0.002	-0.001	0.001	0.000	0.000	0.208	0.268	0.243	0.321	0.448	0.224	0.219
0.5	2	0.5	0.05	-0.005	0.007	0.004	0.008	0.016	0.007	0.006	0.191	0.225	0.227	0.213	0.689	0.211	0.204
1.0	2	0.5	0.05	-0.006	0.006	0.005	0.006	0.020	0.007	0.006	0.195	0.246	0.236	0.309	0.675	0.216	0.209
2.0	2	0.5	0.05	-0.009	0.006	0.005	0.004	0.015	0.007	0.007	0.202	0.260	0.243	0.383	0.630	0.223	0.216
0.5	1	0	0.50	-0.063	-0.050	-0.042	-0.063	-0.079	-0.031	-0.039	0.255	0.351	0.324	0.661	0.746	0.252	0.250
1.0	1	0	0.50	-0.134	-0.102	-0.084	-0.129	-0.127	-0.063	-0.084	0.334	0.473	0.419	0.896	1.164	0.321	0.443
2.0	1	0	0.50	-0.239	-0.192	-0.148	-0.272	-0.262	-0.118	-0.145	0.486	0.739	0.763	1.592	1.969	0.435	0.501
0.5	2	0	0.50	-0.008	0.005	0.003	-0.001	0.022	0.008	0.009	0.230	0.295	0.284	0.577	0.746	0.252	0.247
1.0	2	0	0.50	-0.010	0.008	0.007	0.010	0.026	0.016	0.014	0.288	0.401	0.383	0.546	1.018	0.316	0.312
2.0	2	0	0.50	-0.002	0.015	0.006	0.005	0.016	0.030	0.030	0.403	0.604	0.574	1.003	1.338	0.427	0.429
0.5	1	0.5	0.50	-0.051	-0.037	-0.032	-0.043	-0.051	-0.022	-0.030	0.262	0.338	0.327	0.623	0.773	0.263	0.342
1.0	1	0.5	0.50	-0.105	-0.077	-0.069	-0.106	-0.137	-0.047	-0.057	0.343	0.495	0.461	0.971	1.232	0.340	0.341
2.0	1	0.5	0.50	-0.188	-0.137	-0.107	-0.212	-0.196	-0.075	-0.101	0.496	0.730	0.736	1.535	1.794	0.463	0.575
0.5	2	0.5	0.50	-0.009	0.005	0.002	0.002	0.018	0.010	0.010	0.238	0.326	0.314	0.536	0.725	0.264	0.256
1.0	2	0.5	0.50	-0.009	0.005	0.006	0.005	0.030	0.017	0.015	0.296	0.424	0.396	0.766	1.015	0.323	0.327
2.0	2	0.5	0.50	-0.001	0.009	0.013	0.001	0.015	0.034	0.028	0.406	0.619	0.594	1.082	1.413	0.433	0.485

TABLE 3. Experiment 2 under local misspecification from partial contamination: RMSE, Probas, Bias and standard deviation with $T = 50$ and 10,000 replications. The distribution of the contamination is χ^2_1 with $d = 1$, and $t_{1.5}$ with $d = 2$; the proportion π of contaminated sample is either 5%, or 50% contamination.

Appendix B. Proofs of the theoretical results in Sections 2 and 3

Proof of Theorem 2.1: Our proof closely follows the steps of the proof of Theorem 1 in Schennach (2007) written for EL. To deduce that HD is not root n consistent, we proceed as follows. We show that for any HD estimator based on random sample X with distribution $F_\infty(x)$ and unbounded support, there exists a family of other estimators $\hat{\theta}_{k,n}$ based on compactly supported $F_k(x)$ (to be precisely defined below), all having a narrower distribution than HD for each n and asymptotic variance that diverges as $k \to \infty$. Then, the same argument as the one used in Schennach (2007) allows us to conclude that HD is not root n consistent.
Let $F_k(x)$ be a sequence of distributions indexed by $k \in \mathbb{N}$, each having support C_k with $C_k \equiv \{x \in \mathcal{X} \ s.t. \ g(x,\theta) \in \mathcal{G}_k \ \forall \theta \in \Theta\}$ and $\mathcal{G}_k$ an increasing sequence of nested compact subsets of $\mathbb{R}^d$ such that $\bigcup_{k=1}^{\infty} \mathcal{G}_k = \mathbb{R}^d$. In addition, $F_k(x)$ is chosen so that the moment conditions are uniformly misspecified in the sense that,

$$\exists \overline{k} \in \mathbb{N} \ s.t. \ \inf_{k \geq \overline{k}} \inf_{\theta \in \Theta} \|E[g(x,\theta)]\| > 0$$

Let $\hat{\theta}_{k,n}$ denote the HD estimator computed with a sample of size n generated with true DGP $K_k(x)$ and let $\theta_k^* \in \Theta$ denote its corresponding pseudo-true value. From the interpretation of the HD estimator as a GEL estimator (see Newey and Smith (2004)) and KOE (2013a, p.1191), we have

$$\hat{\theta}_{k,n} = \arg\min_{\theta} \max_{\gamma} -\frac{1}{n} \sum_{i=1}^{n} \frac{2}{(1 - \gamma' g(X_i,\theta)/2)} \,.$$

The first-order condition with respect to θ and γ write, respectively:

$$-\frac{1}{n} \sum_{i=1}^{n} \sum_{i=1}^{n} \frac{\hat{G}_i' \hat{\gamma}_{k,n}}{[1 - \hat{\gamma}_{k,n}' g(X_i, \hat{\theta}_{k,n})/2]^2} = 0 \quad \text{where} \quad \hat{G}_i = \frac{\partial g(X_i, \hat{\theta}_{k,n})}{\partial \theta'}$$
$$-\frac{1}{n} \sum_{i=1}^{n} \sum_{i=1}^{n} \frac{g(X_i, \hat{\theta}_{k,n})}{[1 - \hat{\gamma}_{k,n}' g(X_i, \hat{\theta}_{k,n})/2]^2} = 0 \,.$$

The asymptotic properties of GEL-type estimators are well known:

$$\sqrt{n} \left[\begin{pmatrix} \hat{\theta}_{k,n} \\ \hat{\gamma}_{k,n} \end{pmatrix} - \begin{pmatrix} \theta_k^* \\ \gamma_k^* \end{pmatrix} \right] \xrightarrow{d} N(0, H_k^{-1} S_k H_k^{-1}) \quad \text{as } n \to \infty \text{ for } k \text{ fixed}$$

with

$$S_k = E[\phi(\theta_k^*, \gamma_k^*) \phi(\theta_k^*, \gamma_k^*)'] = \begin{pmatrix} E[\tau_i^4 G_i' \gamma_k^* \gamma_k^{*'} G_i] & E[\tau_i^4 G_i' \gamma_k^* g_i'] \\ E[\tau_i^4 g_i \gamma_k^{*'} G_i] & E[\tau_i^4 g_i g_i'] \end{pmatrix}$$
$$H_k = E\left(\frac{\partial \phi'(\theta_k^*, \gamma_k^*)}{\partial [\theta' \ \gamma']'} \right) = E \begin{pmatrix} \tau_i^3 G_i \gamma_k^* \gamma_k^{*'} G_i + \tau_i^2 \frac{\partial (G_i' \gamma_k^*)}{\partial \theta'} & \tau_i^3 G_i' \gamma_k^* g_i' + \tau_i^2 G_i' \\ \tau_i^3 g_i \gamma_k^{*'} G_i + \tau_i^2 G_i & \tau_i^3 g_i g_i' \end{pmatrix}$$

where

$$g_i = g(X_i, \theta_k^*) \,, \quad G_i = \frac{\partial g(X_i, \theta_k^*)}{\partial \theta'} \,, \quad \tau_i = \frac{1}{1 - \gamma_k^{*'} g_i/2} \,, \quad \phi(\theta, \gamma) = \begin{pmatrix} \frac{G_i' \gamma}{(1 - \gamma' g_i/2)^2} \\ \frac{g_i}{(1 - \gamma' g_i/2)^2} \end{pmatrix} .$$

From the calculations in the dual problem, we have:

$$\sqrt{\pi}_i = \frac{1}{\sqrt{n}(1 - \gamma_k^{*'} g_i/2)} > 0 \ \Rightarrow \ \frac{1}{(1 - \gamma_k^{*'} g_i/2)} > 0 \tag{B.1}$$

Since $\{g(x, \theta_k^*), x \in \mathcal{X}\}$ is unbounded in every direction, the set $\{g(x, \theta_k^*) \in C_k\}$ becomes unbounded in every direction as $k \to \infty$, Hence, the only way to have (B.1) is to have $\gamma_k^* \to 0$ as $k \to \infty$. Since $\gamma_k^* \to 0$ as $k \to \infty$, S_k and H_k can be simplified by noting that when $(H_k^{-1} S_k H_k^{-1})$ is calculated, any term containing γ_k^* will be dominated by terms not containing it. We get:

$$S_k \to \begin{pmatrix} 0 & 0 \\ 0 & E(\tau_i^4 g_i g_i') \end{pmatrix} \quad \text{and} \quad H_k^{-1} \to \begin{pmatrix} 0 & E(\tau_i^2 G_i') \\ E(\tau_i^2 G_i) & E(\tau_i^3 g_i g_i') \end{pmatrix}^{-1} \equiv \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

Define Σ_k as the (p,p) top-left submatrix of $(H_k^{-1}S_kH_k^{-1})$, that is $\Sigma_k = B_{12}E(\tau_i^4 g_i g_i')B_{21}$. Recall $\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1}$ top-right corner term is $-F^{-1}BD^{-1}$ with $F = A - BD^{-1}C$. Thus:

$$B_{12} = \left[E(\tau_i^2 G_i')\left(E(\tau_i^3 g_i g_i')\right)^{-1} E(\tau_i^2 G_i)\right]^{-1} E(\tau_i^2 G_i')\left(E(\tau_i^3 g_i g_i')\right)^{-1} = B_{21}'$$

To show that Σ_k diverges, we show the following three properties:

(i) $E(\tau_i^4 g_i g_i')$ has a divergent eigenvalue;
(ii) $\|E(\tau_i^2 G_i)\| = o\left(\left[E(\tau_i^4\|g_i g_i'\|)\right]^{1/2}\right)$;
(iii) $\|B_{12}\|\left[E(\tau_i^4\|g_i g_i'\|)\right]^{1/2}$ diverges.

(i) First, we show that $E(\tau_i^4 g_i g_i')$ has a divergent eigenvalue:

$$\begin{aligned} g_i(1-\gamma_k^{*'}g_i/2)^2 &= g_i(1-\gamma_k^{*'}g_i + (\gamma' g_i)^2/4) \\ &= g_i - g_i g_i'\gamma_k^* + g_i g_i'\gamma g_i'\gamma_k^*/4 \\ &= g_i - g_i g_i'\gamma_k^*/2(2 - g_i'\gamma_k^*/2) \\ &= g_i - g_i g_i'\gamma_k^*/2 - g_i g_i'\gamma_k^*/2(1 - g_i'\gamma_k^*/2) \end{aligned}$$

$$\begin{aligned} &\Rightarrow \quad g_i = \frac{g_i}{(1-\gamma_k^{*'}g_i/2)^2} - \frac{g_i(g_i'\gamma_k^*)/2}{(1-\gamma_k^{*'}g_i/2)^2} - \frac{g_i(g_i'\gamma_k^*)/2}{(1-\gamma_k^{*'}g_i/2)} \\ &\Rightarrow \quad E(g_i) = 0 - \left\{E\left[\frac{g_i g_i'}{(1-\gamma_k^{*'}g_i/2)^2}\right] + E\left[\frac{g_i g_i'}{(1-\gamma_k^{*'}g_i/2)}\right]\right\}\frac{\gamma_k^*}{2} \\ &\Rightarrow \quad E(g_i) \equiv -(\Omega_1+\Omega_2)\frac{\gamma_k^*}{2} \end{aligned}$$

Since $\inf_{k\geq\bar{k}} E(g(X_i,\theta_k^*)) > 0$ some $\bar{k}\in\mathbb{N}$, the only way to have $\gamma_k^*\to 0$ is if $(\Omega_1+\Omega_2)$ has a divergent eigenvalue. Let v be a unit eigenvector associated with such eigenvalue:

$$\begin{aligned} v'\Omega_1 v &= E\left(\frac{v'g_i}{(1-\gamma_k^{*'}g_i/2)^2}v'g_i\right) \leq \left[E\left(\frac{v'g_i}{(1-\gamma_k^{*'}g_i/2)^2}\right)^2\right]^{1/2} \\ v'\Omega_2 v &= E\left(\frac{v'g_i}{(1-\gamma_k^{*'}g_i/2)}v'g_i\right) \leq \left[E\left(\frac{v'g_i}{(1-\gamma_k^{*'}g_i/2)}\right)^2\right]^{1/2} = (v'\Omega_1 v)^{1/2}[E(v'g_i)^2]^{1/2} \end{aligned}$$

Hence,

$$v'\Omega v \equiv v'\Omega_1 v + v'\Omega_2 v \leq [E(v'g_i)^2]^{1/2}\left\{\left[E\left(\frac{v'g_i}{(1-\gamma_k^{*'}g_i/2)^2}\right)^2\right]^{1/2} + (v'\Omega_1 v)^{1/2}\right\}$$

Since

$$\begin{aligned} &a) \quad E(v'g_i)^2 \leq \sup_{\theta\in\Theta} E\|g(X_i,\theta)\|^2 < \infty \quad \text{by assumption,} \\ &b) \quad v'\Omega_1 v \leq \left[E\left(\frac{v'g_i}{(1-\gamma_k^{*'}g_i/2)^2}\right)^2 E(v'g_i)^2\right]^{1/2} \quad \text{diverges as shown above,} \\ &c) \quad E\left(\frac{v'g_i}{(1-\gamma_k^{*'}g_i/2)^2}\right)^2 = E[\tau_i^4(v'g_i)^2]\,, \end{aligned}$$

we conclude that $E(\tau_i^4 g_i g_i')$ has a divergent eigenvalue.

(ii) We now show that $\|E(\tau_i^2 G_i)\| = o\left(\left[E(\tau_i^4\|g_i g_i'\|)\right]^{1/2}\right)$.

$$\begin{aligned}
\tau_i^2 G_i &= \frac{1}{(1-\gamma_k^{*'} g_i/2)^2} G_i = \left[1 + \tau_i^2 \gamma_k^{*'} g_i - \tau_i^2 \left(\frac{\gamma_k^{*'} g_i}{2}\right)^2\right]^2 G_i \\
\|E(\tau_i^2 G_i)\| &= \left\| E\left[(1 + \tau_i^2 \gamma_k^{*'} g_i - \tau_i^2 \left(\frac{\gamma_k^{*'} g_i}{2}\right)^2) G_i\right]\right\| \\
&\leq E\|G_i\| + E\|\tau_i^2 \gamma_k^{*'} g_i G_i\| + E\left\|\tau_i^2 \left(\frac{\gamma_k^{*'} g_i}{2}\right)^2 G_i\right\| \\
E\tau_i^2\|\gamma_k^{*'} g_i G_i\| &= E\left(\tau_i^2 \|g_i\| \|G_i\|\right) \|\gamma\| \\
&\leq \left[E(\tau_i^4 \|g_i\|^2)\right]^{1/2} \left[\|G_i\|^2\right]^{1/2} \|\gamma_k^*\|
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Then,

$$\frac{E(\tau_i^2\|\gamma_k^{*'} g_i G_i\|)}{[E(\tau_i^4\|g_i\|^2)]^{1/2}} \to 0 \quad \Rightarrow \quad E(\tau_i^2\|\gamma_k^{*'} g_i G_i\|) = o\left(\left[E(\tau_i^4\|g_i\|^2)\right]^{1/2}\right)$$

$$E\left\|\tau_i^2 \left(\frac{\gamma_k^{*'} g_i}{2}\right)^2 G_i\right\| = E(\tau_i^2\|g_i\|^2\|G_i\|)\|\gamma_k^*\|^2 \leq \left[E(\tau_i^4\|g_i\|^2)\right]^{1/2} \left[E\|g_i\|^2\|G_i\|^2\right]^{1/2} \|\gamma_k^*\|^2$$

$$\Rightarrow \quad \|E(\tau_i^2 G_i)\| = o\left(\left[E(\tau_i^4\|g_i\|^2)\right]^{1/2}\right) = o\left(\left[E(\tau_i^4\|g_i g_i'\|)\right]^{1/2}\right) = o\left(\left[E(\tau_i^4 v' g_i g_i' v)\right]^{1/2}\right)$$

(iii) Finally, we show that $\|B_{12}\| \left[E(\tau_i^4\|g_i g_i'\|)\right]^{1/2} \to \infty$.
First, it follows from the Cauchy-Schwarz inequality that:

$$\|B_{12}\| \left[E(\tau_i^4\|g_i g_i'\|)\right]^{1/2} \geq \|B_{12} E(\tau_i^2 G_i)\| \frac{\left[E(\tau_i^4\|g_i g_i'\|)\right]^{1/2}}{\|E(\tau_i^2 G_i)\|}$$

Then, from the definition of B_{12}, we have:

$$B_{12} E(\tau_i^2 G_i) = I_p \quad \Rightarrow \quad \|B_{12} E(\tau_i^2 G_i)\| = O_p(1)$$

Finally, we showed in (ii) above that

$$\|E(\tau_i^2 G_i)\| = o\left(\left[E(\tau_i^4\|g_i g_i'\|)\right]^{1/2}\right) \quad \Rightarrow \quad \frac{\|E(\tau_i^2 G_i)\|}{\left[E(\tau_i^4\|g_i g_i'\|)\right]^{1/2}} \to 0 \quad \Rightarrow \quad \frac{\left[E(\tau_i^4\|g_i g_i'\|)\right]^{1/2}}{\|E(\tau_i^2 G_i)\|} \to \infty \quad \square$$

Proof of Theorem 3.1: To simplify the notation, we make the dependence of all quantities on $\hat{\theta}$ implicit and introduce the following notations: $\hat{\pi}_i = \hat{\pi}_i(\hat{\theta})$, $\hat{\lambda} = \hat{\lambda}(\hat{\theta})$, $g_i = g_i(x, \hat{\theta})$. In addition, $\sum_i = \sum_{i=1}^n$.

The first part follows readily from the discussion leading to the statement of the theorem. Regarding the second part, let us start with the following preliminary computation:

$$\begin{aligned}
\frac{d\hat{\pi}_i}{d\theta} &= \frac{d}{d\theta}\left[\frac{\exp(\hat{\lambda}'g_i)}{\sum_j \exp(\hat{\lambda}'g_j)}\right] \\
&= \frac{1}{\left[\sum_j \exp(\hat{\lambda}'g_j)\right]^2}\left[\frac{d(\exp(\hat{\lambda}'g_i)))}{d\theta}\sum_j \exp(\hat{\lambda}'g_j) - \exp(\hat{\lambda}'g_i)\sum_j \frac{d}{d\theta}\exp(\hat{\lambda}'g_j)\right] \\
&= \frac{1}{\left[\sum_j \exp(\hat{\lambda}'g_j)\right]^2}\left[\frac{d(\hat{\lambda}'g_i)}{d\theta}\exp(\hat{\lambda}'g_i)\sum_j \exp(\hat{\lambda}'g_j) - \exp(\hat{\lambda}'g_i)\sum_j \frac{d(\hat{\lambda}'g_j)}{d\theta}\exp(\hat{\lambda}'g_j)\right] \\
&= \frac{\exp(\hat{\lambda}'g_i)}{\sum_j \exp(\hat{\lambda}'g_j)\times\sum_k \exp(\hat{\lambda}'g_k)}\left[\frac{d(\hat{\lambda}'g_i)}{d\theta}\sum_j \exp(\hat{\lambda}'g_j) - \sum_j \frac{d(\hat{\lambda}'g_j)}{d\theta}\exp(\hat{\lambda}'g_j)\right] \\
&= \hat{\pi}_i\left[\frac{d(\hat{\lambda}'g_i)}{d\theta} - \sum_j \frac{d(\hat{\lambda}'g_j)}{d\theta}\frac{\exp(\hat{\lambda}'g_j)}{\sum_k \exp(\hat{\lambda}'g_k)}\right] \\
&= \hat{\pi}_i\left[\frac{d(\hat{\lambda}'g_i)}{d\theta} - \sum_j \hat{\pi}_j\frac{d(\hat{\lambda}'g_j)}{d\theta}\right]
\end{aligned}$$

We can now proceed from

$$H^2(\hat{\pi}, P_n) = 1 - \frac{1}{\sqrt{n}}\sum_i \sqrt{\hat{\pi}_i}$$

The differentiation with respect to θ gives:

$$\begin{aligned}
2\frac{dH^2}{d\theta} &= -\frac{1}{\sqrt{n}}\sum_i \frac{1}{\sqrt{\hat{\pi}_i}}\frac{d\hat{\pi}_i}{d\theta} \\
&= -\frac{1}{\sqrt{n}}\sum_i\left(\sqrt{\hat{\pi}_i}\frac{d(\hat{\lambda}'g_i)}{d\theta} - \sqrt{\hat{\pi}_i}\sum_j \frac{d(\hat{\lambda}'g_j)}{d\theta}\hat{\pi}_j\right) \\
&= \frac{1}{\sqrt{n}}\sum_i \sqrt{\hat{\pi}_i}\sum_j \hat{\pi}_j\frac{d(\hat{\lambda}'g_j)}{d\theta} - \frac{1}{\sqrt{n}}\sum_i \sqrt{\hat{\pi}_i}\frac{d(\hat{\lambda}'g_i)}{d\theta} \\
&= 0
\end{aligned}$$

From (12), the first-order condition for $\hat{\lambda}$ is:

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

□

Lemma B.1. *Let*

$$\mathbf{\Delta}_{P_n}(\lambda,\theta) = \frac{E_{P_n}\left[\exp(\lambda' g(X,\theta)/2)\right]}{\sqrt{E_{P_n}\left[\exp(\lambda' g(X,\theta))\right]}}, \quad \mathbf{\Delta}_{P_*}(\lambda,\theta) = \frac{E_{P_*}\left[\exp(\lambda' g(X,\theta)/2)\right]}{\sqrt{E_{P_*}\left[\exp(\lambda' g(X,\theta))\right]}}$$

and $(\hat{\lambda},\hat{\theta})$ an arbitrary sequence of $\Lambda\times\Theta$, a compact set. If (i) $\mathbf{\Delta}_{P_n}(\lambda,\theta)$ converges uniformly in probability P_* and over $\Lambda\times\Theta$ to $\mathbf{\Delta}_{P_*}(\lambda,\theta)$, with $\mathbf{\Delta}_{P_*}$ continuous in both its arguments, (ii) $Var_{P_*}(g(X,\theta))$ is non singular for all $\theta\in\Theta$ with smallest eigenvalue bounded away from 0, and (iii) $\mathbf{\Delta}_{P_n}(\hat{\lambda},\hat{\theta}) \xrightarrow{P_*} 1$, then

$$\hat{\lambda} \xrightarrow{P_*} 0.$$

Proof of Lemma B.1: By the triangle inequality, (i) and (iii) imply that $\mathbf{\Delta}_{P_*}(\hat{\lambda},\hat{\theta}) \xrightarrow{P_*} 1$. Let $\epsilon > 0$ and $N_\epsilon = \{\lambda\in\mathbb{R}^m : \|\lambda\| < \epsilon\}$ and $\bar{N}_\epsilon$ its complement. By the Jensen's inequality, since $x\mapsto\sqrt{x}$ is strictly concave, $\mathbf{\Delta}_{P_*}(\lambda,\theta)\le 1$ with equality occurring only for $\lambda' g(X,\theta)$ constant P_*-almost surely. By condition (ii), $\lambda' g(X,\theta)$

is constant P_*-almost surely if and only if $\lambda = 0$. By continuity of objective function and compactness of optimization set, there exists $(\bar{\lambda}, \bar{\theta}) \in (\bar{N}_\epsilon \cap \Lambda) \times \Theta$ such that

$$\max_{(\lambda,\theta)\in(\bar{N}_\epsilon\cap\Lambda)\times\Theta} \mathbf{\Delta}_{P_*}(\lambda,\theta) = \mathbf{\Delta}_{P_*}(\bar{\lambda},\bar{\theta}) \equiv A_\epsilon.$$

Since $\bar{\lambda} \neq 0$, $A_\epsilon < 1$. Hence, $\mathbf{\Delta}_{P_*}(\hat{\lambda},\hat{\theta}) > A_\epsilon$ with probability approaching 1 as $n \to \infty$. Therefore, $\hat{\lambda} \notin \bar{N}_\epsilon$ with probability approaching 1, that is $P_*(\|\hat{\lambda}\| < \epsilon) \to 1$ as $n \to \infty$. □

Lemma B.2. *If Assumption 1 holds and $\hat{\theta}$ is the ETHD estimator, then*

$$(a) \quad \mathbf{\Delta}_{P_n}(\hat{\lambda}(\hat{\theta}),\hat{\theta}) = 1 + O_P(n^{-1}), \quad (b) \quad \hat{\lambda}(\hat{\theta}) = O_P(n^{-1/2}), \quad (c) \quad E_{P_n}(g(X,\hat{\theta})) = O_P(n^{-1/2}).$$

Proof of Lemma B.2: We proceed in three steps. Step 1 shows that $\mathbf{\Delta}_{P_n}(\hat{\lambda}(\hat{\theta}),\hat{\theta}) = 1 + O_P(n^{-1})$. This allows, thanks to Lemma B.1 to deduce that $\hat{\lambda}(\hat{\theta}) = o_P(1)$. Step 2 derives the order of magnitude of $\hat{\lambda}(\hat{\theta})$ and Step 3 derives that of $E_{P_n}(g(X,\hat{\theta}))$.

Step 1: We first show that $\mathbf{\Delta}_{P_n}(\hat{\lambda}(\hat{\theta}),\hat{\theta}) = 1 + O_P(n^{-1})$. By definition of $\hat{\theta}$, we have:

$$\mathbf{\Delta}_{P_n}\left(\hat{\lambda}(\theta^*),\theta^*\right) \leq \mathbf{\Delta}_{P_n}\left(\hat{\lambda}(\hat{\theta}),\hat{\theta}\right) \leq 1. \tag{B.2}$$

To concludes (a), it suffices to show that $\mathbf{\Delta}_{P_n}\left(\hat{\lambda}(\theta^*),\theta^*\right) = 1 + O_P(n^{-1})$. For this, observe that by the central limit theorem, $\sqrt{n}E_{P_n}(g(X,\theta^*)) = O_P(1)$. We can therefore apply Lemma A2 of Newey and Smith (2004) to the constant sequence $\bar{\theta} = \theta^*$ and claim that $\hat{\lambda}(\theta^*) = O_P(n^{-1/2})$ and $E_{P_n}\left[\exp\left(\hat{\lambda}(\theta^*)'g(X,\theta^*)\right)\right] \geq 1 + O_P(n^{-1})$. Since $E_{P_n}[\exp(\lambda' g(X,\theta^*))]$ is minimized at $\hat{\lambda}(\theta^*)$ over Λ which contains 0, we actually have

$$1 + O_P(n^{-1}) \leq E_{P_n}\left[\exp\left(\hat{\lambda}(\theta^*)'g(X,\theta^*)\right)\right] \leq 1.$$

Thus $\varepsilon_n \equiv E_{P_n}\left[\exp\left(\hat{\lambda}(\theta^*)'g(X,\theta^*)\right)\right] - 1 = O_P(n^{-1})$.
Also, by definition of $\hat{\lambda}(\theta^*)$, $E_{P_n}\left[\exp\left(\hat{\lambda}(\theta^*)'g(X,\theta^*)\right)\right] \leq E_{P_n}\left[\exp\left(\hat{\lambda}(\theta^*)'g(X,\theta^*)/2\right)\right]$. Hence,

$$\left(E_{P_n}\left[\exp\left(\hat{\lambda}(\theta^*)'g(X,\theta^*)\right)\right]\right)^{1/2} \leq \mathbf{\Delta}_{P_n}\left(\hat{\lambda}(\theta^*),\theta^*\right) \leq 1.$$

But, $\left(E_{P_n}\left[\exp\left(\hat{\lambda}(\theta^*)'g(X,\theta^*)\right)\right]\right)^{1/2} = 1 + \frac{1}{2}\varepsilon_n + O(\varepsilon_n^2) = 1 + O_P(n^{-1})$. Thus $\mathbf{\Delta}_{P_n}\left(\hat{\lambda}(\theta^*),\theta^*\right) = 1 + O_P(n^{-1})$ and we obtain (a) using (B.2).

Step 2: Before deriving the order of magnitude in (b), we first show that $\hat{\lambda}(\hat{\theta}) \xrightarrow{P} 0$. For this, we verify the conditions of Lemma B.1. Conditions (ii) is satisfied thanks to Assumption 1(v), Condition (iii) follows from Step 1. It remains to show (i). Thanks to the dominance condition in Assumption 1(vi), Lemma 2.4 of Newey and McFadden (1994) ensures that $E_{P_n}[\exp(\lambda' g(X,\theta))]$ and $E_{P_n}[\exp(\lambda' g(X,\theta)/2)]$ converge in probability uniformly over $\Lambda \times \Theta$ to $E[\exp(\lambda' g(X,\theta))]$ and $E[\exp(\lambda' g(X,\theta)/2)]$, respectively and both limits functions are continuous in (λ,θ). To conclude (i), we show that $E[\exp(\lambda' g(X,\theta))]$ is bounded away from 0–which is enough to deduce that the ratio $\mathbf{\Delta}_{P_n}(\lambda,\theta)$ converges uniformly in probability to $\mathbf{\Delta}(\lambda,\theta)$. By convexity of $x \mapsto e^x$,

$$E[\exp(\lambda' g(X,\theta))] \geq \exp[\lambda' E_P(g(X,\theta))] \geq \exp\left[-\|\lambda\| E_P\left(\sup_{\theta\in\Theta}\|g(X,\theta)\|\right)\right] \geq \delta > 0,$$

the third and last inequalities are due to compactness of Λ and Assumption 1(iv).

Let us now establish (b). By a second order Taylor expansion of $\mathbf{\Delta}_{P_n}(\hat{\lambda}(\hat{\theta}),\hat{\theta})$ around $\lambda = 0$ with a Lagrange remainder, we have:

$$\mathbf{\Delta}_{P_n}(\hat{\lambda}(\hat{\theta}),\hat{\theta}) = \mathbf{\Delta}_{P_n}(0,\hat{\theta}) + \frac{\partial \mathbf{\Delta}_{P_n}(0,\hat{\theta})}{\partial \lambda'}\hat{\lambda}(\hat{\theta}) + \frac{1}{2}\hat{\lambda}(\hat{\theta})'\frac{\partial^2 \mathbf{\Delta}_{P_n}(\dot{\lambda},\hat{\theta})}{\partial\lambda\partial\lambda'}\hat{\lambda}(\hat{\theta}), \tag{B.3}$$

with $\dot{\lambda} \in (0,\ \hat{\lambda}(\hat{\theta}))$. We have:

$$\frac{\partial \mathbf{\Delta}_{P_n}(\lambda,\theta)}{\partial\lambda} = \frac{1}{2}\left\{\frac{E_{P_n}[g(X,\theta)\exp(\lambda' g(X,\theta)/2)]}{(E_{P_n}[\exp(\lambda' g(X,\theta))])^{1/2}} - \frac{E_{P_n}[g(X,\theta)\exp(\lambda' g(X,\theta))]\,E_{P_n}[\exp(\lambda' g(X,\theta)/2)]}{(E_{P_n}[\exp(\lambda' g(X,\theta))])^{3/2}}\right\}$$

and $\frac{\partial^2 \mathbf{\Delta}_{P_n}(\lambda,\theta)}{\partial\lambda\partial\lambda'} = \frac{1}{2}\left(\mathbf{\Delta}^{(1)}_{P_n}(\lambda,\theta) + \mathbf{\Delta}^{(2)}_{P_n}(\lambda,\theta)\right)$, with (letting $g \equiv g(X,\theta)$),

$$\mathbf{\Delta}^{(1)}_{P_n}(\lambda,\theta) = \frac{1}{2}\frac{E_{P_n}[gg'\exp(\lambda' g/2)]}{[E_{P_n}(\lambda' g)]^{1/2}} - \frac{E_{P_n}[gg'\exp(\lambda' g)]E_{P_n}[\exp(\lambda' g/2)]}{[E_{P_n}(\lambda' g)]^{3/2}}$$

$$\begin{aligned}\mathbf{\Delta}^{(2)}_{P_n}(\lambda,\theta) =& \frac{3}{2}\frac{E_{P_n}[g\exp(\lambda' g)]E_{P_n}[g'\exp(\lambda' g)]E_{P_n}[\exp(\lambda' g/2)]}{(E_{P_n}[\exp(\lambda' g)])^{5/2}} - \frac{1}{2}\frac{E_{P_n}[g\exp(\lambda' g)]E_{P_n}[g'\exp(\lambda' g/2)]}{(E_{P_n}[\exp(\lambda' g)])^{3/2}} \\ & - \frac{1}{2}\frac{E_{P_n}[g\exp(\lambda' g/2)]E_{P_n}[g'\exp(\lambda' g)]}{(E_{P_n}[\exp(\lambda' g)])^{3/2}}.\end{aligned}$$

Hence, $\frac{\partial \mathbf{\Delta}_{P_n}(0,\hat\theta)}{\partial\lambda} = 0$. We also have that:

$$\frac{\partial^2 \mathbf{\Delta}_{P_n}(\dot\lambda,\hat\theta)}{\partial\lambda\partial\lambda'} = -\frac{1}{4}Var(g(X,\hat\theta)) + o_P(1). \tag{B.4}$$

To see this, we observe that, by uniform convergence,

$$E_{P_n}\left[\exp\left(\dot\lambda' g(X,\hat\theta)\right)\right] = E\left(\exp\left(\dot\lambda' g(X,\hat\theta)\right)\right) + o_P(1).$$

By continuity of $(\lambda,\theta) \mapsto E\left(\exp\left(\lambda' g(X,\theta)\right)\right)$, the fact that $g(X,\hat\theta) = O_P(1)$ and $\dot\lambda \xrightarrow{P} 0$ implies that $E\left(\exp\left(\dot\lambda' g(X,\hat\theta)\right)\right) \to 1$ in probability as $n\to\infty$ and we have

$$E_{P_n}\left[\exp\left(\dot\lambda' g(X,\hat\theta)\right)\right] \xrightarrow{P} 1$$

as well.

We can also claim that

$$E_{P_n}\left[g(X,\hat\theta)\exp\left(\dot\lambda' g(X,\hat\theta)\right)\right] = E\left(g(X,\hat\theta)\exp\left(\dot\lambda' g(X,\hat\theta)\right)\right) + o_P(1) = E\left(g(X,\hat\theta)\right) + o_P(1).$$

To see this, let $\mathcal{N}\subset\mathbb{R}^m$ be a small neighborhood of 0. For λ near 0, we have

$$\|g(x,\theta)\exp\left(\lambda' g(x,\theta)\right)\| \le \sup_{\theta\in\Theta}\|g(x,\theta)\| \sup_{\theta\in\Theta,\lambda\in\mathcal{N}}\exp\left(\lambda' g(x,\theta)\right).$$

Applying the Hölder inequality with β: $1/\alpha + 1/\beta = 1$, have:

$$\begin{aligned} & E\left(\sup_{\theta\in\Theta}\|g(X,\theta)\|\sup_{\theta\in\Theta,\lambda\in\mathcal{N}}\exp\left(\lambda' g(X,\theta)\right)\right) \\ \le\ & \left(E\sup_{\theta\in\Theta}\|g(X,\theta)\|^\alpha\right)^{\frac{1}{\alpha}}\left(E\sup_{\theta\in\Theta,\lambda\in\mathcal{N}}\exp\left(\beta\lambda' g(X,\theta)\right)\right)^{\frac{1}{\beta}} \\ \le\ & \left(E\sup_{\theta\in\Theta}\|g(X,\theta)\|^\alpha\right)^{\frac{1}{\alpha}}\left(E\sup_{\theta\in\Theta,\lambda\in\Lambda}\exp\left(\lambda' g(X,\theta)\right)\right)^{\frac{1}{\beta}} < \infty\end{aligned}$$

This establishes the dominance condition needed for the claim to hold. We can proceed the same way to show that:

$$\begin{gathered} E_{P_n}\left[g(X,\hat\theta)g(X,\hat\theta)'\exp\left(\dot\lambda' g(X,\theta)\right)\right] = E\left(g(X,\hat\theta)g(X,\hat\theta)'\right) + o_P(1); \\ E_{P_n}\left[g(X,\hat\theta)g(X,\hat\theta)'\exp\left(\dot\lambda' g(X,\theta)/2\right)\right] = E\left(g(X,\hat\theta)g(X,\hat\theta)\right) + o_P(1); \\ E_{P_n}\left[g(X,\hat\theta)\exp\left(\dot\lambda' g(X,\theta)/2\right)\right] = E\left(g(X,\hat\theta)\right) + o_P(1); \quad \text{and} \quad E_{P_n}\left[\exp\left(\dot\lambda' g(X,\theta)/2\right)\right] = 1 + o_P(1)\end{gathered}$$

and (B.4) follows. Therefore, (B.3) can be written:

$$\mathbf{\Delta}_{P_n}(\hat\lambda(\hat\theta),\hat\theta) = 1 - \frac{1}{8}\hat\lambda(\hat\theta)'Var(g(X,\hat\theta))\hat\lambda(\hat\theta) + o_P(1)\|\hat\lambda(\hat\theta)\|^2. \tag{B.5}$$

Thus

$$\frac{1}{8}\hat\lambda(\hat\theta)'Var(g(X,\hat\theta))\hat\lambda(\hat\theta) + o_P(1)\|\hat\lambda(\hat\theta)\|^2 = O_P(n^{-1}).$$

From Assumption 1(v), this implies that:

$$\underline{\ell}\|\hat\lambda(\hat\theta)\|^2/8 + o_P(1)\|\hat\lambda(\hat\theta)\|^2 \le \frac{1}{8}\hat\lambda(\hat\theta)'Var(g(X,\hat\theta))\hat\lambda(\hat\theta) + o_P(1)\|\hat\lambda(\hat\theta)\|^2 = O_P(n^{-1})$$

with $\underline{\ell} > 0$ and we can conclude that

$$\|\hat\lambda(\hat\theta)\|^2(1 + o_P(1)) = O_P(n^{-1})$$

implying that

$$\|\hat{\lambda}(\hat{\theta})\|^2 = O_P(n^{-1})$$

or, equivalently, $\hat{\lambda}(\hat{\theta}) = O_P(n^{-1/2})$, concluding Step 2.

Step 3: Now, we show that $E_{P_n}(g(X,\hat{\theta})) = O_P(n^{-1/2})$. Let $\tilde{\lambda} = -\frac{E_{P_n}(g(X,\hat{\theta}))}{\sqrt{n}\|E_{P_n}(g(X,\hat{\theta}))\|} + \hat{\lambda}(\hat{\theta})$. By definition,

$$E_{P_n}\left[\exp\left(\hat{\lambda}(\hat{\theta})'g(X,\hat{\theta})\right)\right] \leq E_{P_n}\left[\exp\left(\tilde{\lambda}'g(X,\hat{\theta})\right)\right].$$

Second order Taylor expansions of each side around 0 with a Lagrange remainder gives:

$$E_{P_n}\left[\exp\left(\hat{\lambda}(\hat{\theta})'g(X,\hat{\theta})\right)\right] = 1 + \hat{\lambda}(\hat{\theta})'E_{P_n}\left(g(X,\hat{\theta})\right) + \frac{1}{2}\hat{\lambda}(\hat{\theta})'E_{P_n}\left[g(X,\hat{\theta})g(X,\hat{\theta})'\exp\left(\dot{\lambda}'g(X,\hat{\theta})\right)\right]\hat{\lambda}(\hat{\theta})$$

and

$$\begin{aligned} E_{P_n}\left[\exp\left(\tilde{\lambda}'g(X,\hat{\theta})\right)\right] &= 1 + \hat{\lambda}(\hat{\theta})'E_{P_n}\left(g(X,\hat{\theta})\right) - n^{-1/2}\left\|E_{P_n}\left(g(X,\hat{\theta})\right)\right\| \\ &\quad + \tfrac{1}{2}\tilde{\lambda}'E_{P_n}\left[g(X,\hat{\theta})g(X,\hat{\theta})'\exp\left(\ddot{\lambda}'g(X,\hat{\theta})\right)\right]\tilde{\lambda}, \end{aligned}$$

with $\dot{\lambda} \in (0, \hat{\lambda}(\hat{\theta}))$ and $\ddot{\lambda} \in (0, \tilde{\lambda})$. Since $\hat{\lambda}(\hat{\theta})$ and $\tilde{\lambda}$ are both $O_P(n^{-1/2})$, so are $\dot{\lambda}$ and $\ddot{\lambda}$ and, as a result, the quadratic terms in both expansions are of order $O_P(n^{-1})$. Thus:

$$1 + \hat{\lambda}(\hat{\theta})'E_{P_n}\left(g(X,\hat{\theta})\right) + O_P(n^{-1}) \leq 1 + \hat{\lambda}(\hat{\theta})'E_{P_n}\left(g(X,\hat{\theta})\right) - n^{-1/2}\left\|E_{P_n}\left(g(X,\hat{\theta})\right)\right\| + O_P(n^{-1})$$

and we can conclude that: $E_{P_n}\left(g(X,\hat{\theta})\right) = O_P(n^{-1/2})$. □

Proof of Theorem 3.2: Proofs of (ii) and (iii) follow from Lemma B.2. We show (i). We have

$$E_{P_n}\left(g(X,\hat{\theta})\right) = E(g(X,\hat{\theta})) + \left(E_{P_n}\left(g(X,\hat{\theta})\right) - E(g(X,\hat{\theta}))\right).$$

By uniform convergence in probability of $E_{P_n}(g(X,\theta))$ towards $E(g(X,\theta))$ over Θ, we have:

$$E_{P_n}\left(g(X,\hat{\theta})\right) = E(g(X,\hat{\theta})) + o_P(1).$$

From (iii), we can deduce that $E(g(X,\hat{\theta})) \xrightarrow{P} 0$ as $n \to \infty$. Since $E(g(X,\theta)) = 0$ is solved only at θ^*, the fact that $\theta \to E(g(X,\theta))$ is continuous and Θ compact, a similar argument that in Newey and McFadden (1994) allows us to conclude that $\hat{\theta} \xrightarrow{P} \theta^*$. □

Proof of Theorem 3.3: (i) We essentially rely on mean-value expansions of the first order optimality conditions for $\hat{\theta}$ and $\hat{\lambda}$. Since $\hat{\theta}$ converges in probability to θ^* which is an interior point, with probability approaching 1, $\hat{\theta}$ is an interior solution and solves the first order condition:

$$\left.\frac{d\boldsymbol{\Delta}_{P_n}(\hat{\lambda}(\theta),\theta)}{d\theta}\right|_{\theta=\hat{\theta}} = \frac{N_1(\hat{\lambda}(\hat{\theta}),\hat{\theta})}{D_1(\hat{\lambda}(\hat{\theta}),\hat{\theta})} - \frac{N_2(\hat{\lambda}(\hat{\theta}),\hat{\theta})}{D_2(\hat{\lambda}(\hat{\theta}),\hat{\theta})} = 0, \tag{B.6}$$

with

$$N_1(\lambda,\theta) = \frac{1}{2}E_{P_n}\left[\left(\frac{d\hat{\lambda}'}{d\theta}(\hat{\theta})g(X,\theta) + \frac{\partial g(X,\hat{\theta})'}{\partial\theta}\lambda\right)\exp\left(\lambda'g(X,\hat{\theta})/2\right)\right],$$

$$N_2(\lambda,\theta) = \frac{1}{2}E_{P_n}\left[\left(\frac{d\hat{\lambda}'}{d\theta}(\hat{\theta})g(X,\theta) + \frac{\partial g(X,\hat{\theta})'}{\partial\theta}\lambda\right)\exp\left(\lambda'g(X,\hat{\theta})\right)\right] \times D_0\left(\frac{\hat{\lambda}}{2},\hat{\theta}\right),$$

$$D_1(\lambda,\theta) = D_0(\lambda,\theta)^{1/2}, \quad D_2(\lambda,\theta) = D_0(\lambda,\theta)^{3/2}, \quad D_0(\lambda,\theta) = E_{P_n}\left[\exp(\lambda'g(X,\theta))\right].$$

Also, the fact that $\hat{\lambda}(\hat{\theta})$ converges in probability to 0 makes it an interior solution so that it solves in λ the first-order condition:

$$E_{P_n}\left[g(X,\hat{\theta})\exp\left(\lambda'g(X,\hat{\theta})\right)\right] = 0. \tag{B.7}$$

Note that, by a similar arguments to those in the proof of Lemma A1 of Newey and Smith (2004), $\max_{1\leq i\leq n}\sup_{\theta\in\Theta}\|g(X_i,\theta)\| = O_P(n^{1/\alpha})$ and for any sequence $\bar{\lambda}$ such that $\bar{\lambda} = O_P(n^{-1/2})$, we have:

$$\|\bar{\lambda}\|\|g(X_i,\theta)\| = o_P(1) \quad \text{and} \quad |\bar{\lambda}'g(X_i,\theta)| = o_P(1),$$

uniformly over $\theta\in\Theta$ and $i = 1,\ldots,n$. We will use these orders of magnitude routinely in the following lines. They allow us to claim that

$$D_0(\hat{\lambda}(\hat{\theta})/2,\hat{\theta}) = 1 + o_P(1), \quad D_1(\hat{\lambda}(\hat{\theta}),\hat{\theta}) = 1 + o_P(1), \quad \text{and} \quad D_2(\hat{\lambda}(\hat{\theta}),\hat{\theta}) = 1 + o_P(1).$$

We will consider the left hand sides of (B.6) and (B.7) and carry out their mean-value expansions around $(0,\theta^*)$. Regarding (B.6), we have:

$$N_2(0,\theta^*) = N_1(0,\theta^*) + o_P(n^{-1/2}), \quad \text{with} \quad N_1(0,\theta^*) = \frac{1}{2}\frac{d\hat{\lambda}(\hat{\theta})'}{d\theta}E_{P_n}\left(g(X,\theta^*)\right)$$

so that the mean-value expansion of (B.6) is:

$$o_P(n^{-1/2}) = \frac{\partial}{\partial\theta'}\left(\frac{N_1(\lambda,\theta)}{D_1(\hat{\lambda},\hat{\theta})} - \frac{N_2(\lambda,\theta)}{D_2(\hat{\lambda},\hat{\theta})}\right)\Bigg|_{(\dot{\lambda},\dot{\theta})}(\hat{\theta}-\theta^*) + \frac{\partial}{\partial\lambda'}\left(\frac{N_1(\lambda,\theta)}{D_1(\hat{\lambda},\hat{\theta})} - \frac{N_2(\lambda,\theta)}{D_2(\hat{\lambda},\hat{\theta})}\right)\Bigg|_{(\dot{\lambda},\dot{\theta})}\hat{\lambda}, \tag{B.8}$$

where $\hat{\lambda}\equiv\hat{\lambda}(\hat{\theta})$, $\dot{\lambda}\in(0,\hat{\lambda})$ and $\dot{\theta}\in(\theta^*,\hat{\theta})$ and both may vary from row to row. We have:

$$\begin{aligned}
\tfrac{\partial N_1(\dot{\lambda},\dot{\theta})}{\partial\theta'} &= \tfrac{1}{2}\tfrac{d\hat{\lambda}'}{d\theta}(\hat{\theta})E_{P_n}\left(\tfrac{\partial g(X,\dot{\theta})}{\partial\theta'}\exp(\dot{\lambda}'g(X,\hat{\theta})/2)\right) = \tfrac{1}{2}\tfrac{d\hat{\lambda}'}{d\theta}(\hat{\theta})E_{P_n}\left(\tfrac{\partial g(X,\dot{\theta})}{\partial\theta'}\right) + o_P(1).\\
\tfrac{\partial N_2(\dot{\lambda},\dot{\theta})}{\partial\theta'} &= \tfrac{1}{2}D_0(\hat{\lambda}/2,\hat{\theta})\tfrac{d\hat{\lambda}'}{d\theta}(\hat{\theta})E_{P_n}\left(\tfrac{\partial g(X,\dot{\theta})}{\partial\theta'}\exp(\dot{\lambda}'g(X,\hat{\theta}))\right) = \tfrac{1}{2}\tfrac{d\hat{\lambda}'}{d\theta}(\hat{\theta})E_{P_n}\left(\tfrac{\partial g(X,\dot{\theta})}{\partial\theta'}\right) + o_P(1).\\
\tfrac{\partial N_1(\dot{\lambda},\dot{\theta})}{\partial\lambda'} &= \tfrac{1}{2}E_{P_n}\left\{\left[\tfrac{\partial g(X,\hat{\theta})'}{\partial\theta} + \tfrac{1}{2}\left[\tfrac{d\hat{\lambda}'}{d\theta}(\hat{\theta})g(X,\dot{\theta})g(X,\hat{\theta})' + \tfrac{\partial g(X,\hat{\theta})'}{\partial\theta}\dot{\lambda}g(X,\hat{\theta})'\right]\right]\exp(\dot{\lambda}'g(X,\hat{\theta})/2)\right\}\\
&= \tfrac{1}{2}E_{P_n}\left(\tfrac{\partial g(X,\hat{\theta})'}{\partial\theta}\right) + \tfrac{1}{4}\tfrac{d\hat{\lambda}'}{d\theta}(\hat{\theta})E_{P_n}\left(g(X,\dot{\theta})g(X,\hat{\theta})'\right) + o_P(1)\\
\tfrac{\partial N_2(\dot{\lambda},\dot{\theta})}{\partial\lambda'} &= \tfrac{1}{2}D_0(\hat{\lambda}/2,\hat{\theta})E_{P_n}\left\{\left(\tfrac{\partial g(X,\hat{\theta})'}{\partial\theta} + \left[\tfrac{d\hat{\lambda}'}{d\theta}(\hat{\theta})g(X,\dot{\theta})g(X,\hat{\theta})' + \tfrac{\partial g(X,\hat{\theta})'}{\partial\theta}\dot{\lambda}g(X,\hat{\theta})'\right]\right)\exp(\dot{\lambda}'g(X,\hat{\theta}))\right\}\\
&= \tfrac{1}{2}E_{P_n}\left(\tfrac{\partial g(X,\hat{\theta})'}{\partial\theta}\right) + \tfrac{1}{2}\tfrac{d\hat{\lambda}'}{d\theta}(\hat{\theta})E_{P_n}\left(g(X,\dot{\theta})g(X,\hat{\theta})'\right) + o_P(1)
\end{aligned}$$

As a result,

$$\begin{aligned}
\tfrac{\partial}{\partial\theta'}\left(\tfrac{N_1(\lambda,\theta)}{D_1(\hat{\lambda},\hat{\theta})} - \tfrac{N_2(\lambda,\theta)}{D_2(\hat{\lambda},\hat{\theta})}\right)\Big|_{(\dot{\lambda},\dot{\theta})} &= o_P(1)\\
\tfrac{\partial}{\partial\lambda'}\left(\tfrac{N_1(\lambda,\theta)}{D_1(\hat{\lambda},\hat{\theta})} - \tfrac{N_2(\lambda,\theta)}{D_2(\hat{\lambda},\hat{\theta})}\right)\Big|_{(\dot{\lambda},\dot{\theta})} &= -\tfrac{1}{4}\tfrac{d\hat{\lambda}(\hat{\theta})'}{d\theta}E_{P_n}\left(g(X,\dot{\theta})g(X,\hat{\theta})'\right) + o_P(1).
\end{aligned}$$

Note that, the first order condition determining $\hat{\lambda}$, that is (B.7) with θ replacing $\hat{\theta}$, defines the implicit function $\hat{\lambda}(\theta)$ with derivative given by

$$\begin{aligned}
\tfrac{d\hat{\lambda}(\theta)}{d\theta'} &= -\left(E_{P_n}\left[g(X,\theta)g(X,\theta)'\exp\left(\hat{\lambda}(\theta)'g(X,\theta)\right)\right]\right)^{-1}\\
&\quad\times E_{P_n}\left[\left(\tfrac{\partial g(X,\theta)}{\partial\theta'} + g(X,\theta)\hat{\lambda}(\theta)'\tfrac{\partial g(X,\theta)}{\partial\theta'}\right)\exp\left(\hat{\lambda}(\theta)'g(X,\theta)\right)\right].
\end{aligned}$$

By similar arguments as above, we have:

$$\frac{d\hat{\lambda}(\hat{\theta})}{d\theta'} = -\left(E_{P_n}\left[g(X,\hat{\theta})g(X,\hat{\theta})'\right]\right)^{-1}E_{P_n}\left(\frac{\partial g(X,\hat{\theta})}{\partial\theta'}\right) + o_P(1) = -\Omega^{-1}G + o_P(1).$$

Thus,

$$\frac{\partial}{\partial\lambda'}\left(\frac{N_1(\lambda,\theta)}{D_1(\lambda,\theta)} - \frac{N_2(\lambda,\theta)}{D_2(\lambda,\theta)}\right)\Bigg|_{(\dot{\lambda},\dot{\theta})} = \frac{1}{4}G' + o_P(1).$$

Also, since $E_{P_n}(g(X,\hat{\theta}) = O_P(n^{-1/2})$, a standard mean-value expansion ensures that $\hat{\theta}-\theta^* = O_P(n^{-1/2})$ and (B.8) amounts to:

$$\sqrt{n}G'\hat{\lambda} = o_P(1). \tag{B.9}$$

The expansion of (B.7) around $(0, \theta^*)$ yields:

$$\begin{aligned} 0 &= E_{P_n}(g(X,\theta^*)) + E_{P_n}\left[\left(\tfrac{\partial g(X,\dot\theta)}{\partial \theta'} + g(X,\dot\theta)\dot\lambda' \tfrac{\partial g(X,\dot\theta)}{\partial \theta'}\right)\exp\left(\dot\lambda' g(X,\dot\theta)\right)\right](\hat\theta - \theta^*) \\ &\quad + E_{P_n}\left[g(X,\dot\theta)g(X,\dot\theta)'\exp\left(\dot\lambda' g(X,\dot\theta)\right)\right]\hat\lambda, \end{aligned}$$

with $(\dot\lambda, \dot\theta) \in (0, \hat\lambda(\hat\theta)) \times (\theta^*, \hat\theta)$ and may differ from row to row. By similar arguments to those previously made, this expression reduces to:

$$G\sqrt{n}(\hat\theta - \theta^*) + \Omega\sqrt{n}\hat\lambda = -\sqrt{n}E_{P_n}(g(X,\theta^*)) + o_P(1). \tag{B.10}$$

Together, (B.9) and (B.10) yield:

$$\begin{pmatrix} \Omega & G \\ G' & 0 \end{pmatrix}\sqrt{n}\begin{pmatrix} \hat\lambda \\ \hat\theta - \theta^* \end{pmatrix} = \begin{pmatrix} -\sqrt{n}E_{P_n}(g(X,\theta^*)) \\ 0 \end{pmatrix} + o_P(1) \tag{B.11}$$

By the standard partitioned inverse matrix formula (see Magnus and Neudecker (1999, p.11)), we have

$$\begin{pmatrix} \Omega & G \\ G' & 0 \end{pmatrix}^{-1} = \begin{pmatrix} \Omega^{-1/2}M\Omega^{-1/2} & \Omega^{-1}G\Sigma \\ \Sigma G'\Omega^{-1} & -\Sigma \end{pmatrix}. \tag{B.12}$$

Hence,

$$\sqrt{n}\begin{pmatrix} \hat\lambda \\ \hat\theta - \theta^* \end{pmatrix} = -\begin{pmatrix} \Omega^{-1/2}M\Omega^{-1/2} \\ \Sigma G'\Omega^{-1} \end{pmatrix}\frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i,\theta^*) + o_P(1)$$

and the statement (i) of the theorem follows easily.

To establish (ii), we use the fact that

$$\sqrt{n}\hat\lambda = -\Omega^{-1/2}M\Omega^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i,\theta^*) + o_P(1)$$

and Equation (B.5). This equation implies that

$$8n\left(1 - \mathbf{\Delta}_{P_n}(\hat\lambda,\hat\theta)\right) = n\hat\lambda'\Omega\hat\lambda + o_P(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i,\theta^*)'\Omega^{-1/2}M\Omega^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i,\theta^*) + o_P(1)$$

and the result follows since $\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \Omega^{-1/2}g(X_i,\theta^*)\right)' M \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \Omega^{-1/2}g(X_i,\theta^*)\right)$ is asymptotically distributed as a χ^2_{m-p}. □

Appendix C. Local misspecification

This section contains the proofs of the main results that appear in Section 4 as well as some useful auxiliary lemmas.

C.1. **Proofs of the main theorems.** This section provides proofs to the main results in Section 4 of the main text.

Proof of Theorem 4.1: The proof follows similar lines as those of Theorem 3.1(ii) in KOE (2013a). To establish Fisher consistency, let $P_{\theta,\zeta}$ be a regular sub-model such that for $t \in \mathbb{R}^p$, $P_{\theta_n,\zeta_n} \in B_H(P_*, r/\sqrt{n})$ for n large enough, with $\theta_n = \theta^* + t/\sqrt{n}$ and $\zeta_n = O(n^{-1/2})$. We further assume that $E_{P_{\theta_n,\zeta_n}}[\sup_{\theta\in\Theta}\|g(X,\theta)\|^\alpha] \leq \delta < \infty$ for some $\delta > 0$. (Note that the particular sub-model used by KOE to derive the lower bound in their Theorem 3.1(i) satisfies this condition.) We have to show that

$$\sqrt{n}(\bar T(P_{\theta_n,\zeta_n}) - \theta^*) \to t,$$

as $n \to \infty$. From Lemma C.5,

$$\sqrt{n}(\bar T(P_{\theta_n,\zeta_n}) - \theta^*) = -\Sigma G'\Omega^{-1}\sqrt{n}E_{P_{\theta_n,\zeta_n}}[g_n(X,\theta^*)] + o(1).$$

By a mean-value expansion, we have:

$$\sqrt{n}E_{P_{\theta_n,\zeta_n}}[g_n(X,\theta^*)] = \sqrt{n}E_{P_{\theta_n,\zeta_n}}[g_n(X,\theta_n)] - E_{P_{\theta_n,\zeta_n}}\left[\frac{\partial g_n(X,\dot\theta)}{\partial\theta'}\right]t,$$

with $\dot{\theta} \in (\theta^*, \theta_n)$ and may vary from row to row. Noting that $E_{P_{\theta_n,\zeta_n}}[g(X,\theta_n)] = 0$, $E_{P_{\theta_n,\zeta_n}}[g_n(X,\theta_n)] = E_{P_{\theta_n,\zeta_n}}[g(X,\theta_n)\mathbb{I}\{X \notin \mathcal{X}_n\}] = o(n^{-1/2})$ (we refer to Equation A.16 of KOE (2013b) for the proof). Also, thanks to Assumption 3(iv), and by the continuity of the map $\theta \mapsto E_{P_*}\left[\frac{\partial g(X,\theta)}{\partial \theta'}\right]$ in a neighborhood of θ^*, we can claim that $E_{P_{\theta_n,\zeta_n}}\left[\frac{\partial g_n(X,\dot{\theta})}{\partial \theta'}\right]$ converges to G as $n \to \infty$. This establishes that $\bar{T}$ is asymptotically Fisher consistent in the claimed family of sub-models and this is enough to apply Theorem 3.1(i) of KOE (2013a), and deduce that

$$\liminf_{n\to\infty} L_n \geq 4r^2 B^*, \tag{C.1}$$

where $L_n = \sup_{Q \in B_H(P_*, r/\sqrt{n})} n\left(\tau \circ \bar{T}(Q_n) - \tau(\theta^*)\right)^2$.

Now, let $F = \frac{\partial \tau(\theta_0)}{\partial \theta'}\Sigma G' \Omega^{-1}$ and $Q_n \in B_H(P_*, r/\sqrt{n})$. By Lemma C.3(iv), $\bar{T}(Q_n) \to \theta^*$ as $n \to \infty$ and using Lemma C.5, a Taylor expansion of $\tau(\bar{T}(Q_n))$ around θ^* ensures that:

$$\sqrt{n}\left(\tau \circ \bar{T}_{Q_n} - \tau(\theta^*)\right) = -\sqrt{n}F \int g_n(X,\theta^*) dQ_n + o(1).$$

From Lemma A.4 of KOE (2013b), we have $E_{P_*}(g_n(X,\theta^*)) = o(n^{-1/2})$. Thus,

$$-\sqrt{n}F \int g_n(X,\theta^*) dQ_n + o(1) = -\sqrt{n}F \int g_n(X,\theta^*)(dQ_n - dP_*) + o(1)$$

$$= \quad -\sqrt{n}F \int g_n(X,\theta^*)\left(dQ_n^{1/2} - dP_*^{1/2}\right) dQ_n^{1/2} - \sqrt{n}F \int g_n(X,\theta^*)\left(dQ_n^{1/2} - dP_*^{1/2}\right) dP_*^{1/2} + o(1).$$

By the triangle inequality, we have

$$n(\left(\tau \circ \bar{T}(Q_n) - \tau(\theta^*)\right)^2 \leq n(A_1 + A_2 + 2A_3) + o(1),$$

with

$$A_1 = \left| F \int g_n(x,\theta^*)\left(dQ_n^{1/2} - dP_*^{1/2}\right) dQ_n^{1/2} \right|^2, \quad A_1 = \left| F \int g_n(x,\theta^*)\left(dQ_n^{1/2} - dP_*^{1/2}\right) dP_*^{1/2} \right|^2$$

and $A_3 = \sqrt{A_1 \cdot A_2}$. By the Cauchy-Schwarz inequality and then by Lemma A.5(i) of KOE (2013b), we have:

$$A_1 \leq \left| F\left(\int g_n(X,\theta^*) g_n(X,\theta^*)' dQ_n\right) F' \right| \cdot \int \left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 \leq B^* \frac{r^2}{n} + o(n^{-1}).$$

By the same way, we have $A_2 \leq B^* \frac{r^2}{n} + o(n^{-1})$ and we can deduce that $A_3 \leq B^* \frac{r^2}{n} + o(n^{-1})$. Therefore,

$$n\left(\tau \circ \bar{T}(Q_n) - \tau(\theta^*)\right)^2 \leq 4r^2 B^* + o(1), \tag{C.2}$$

Besides, from Lemma C.2, $\bar{T}$ is well-defined on $B_H(P_*, r/\sqrt{n})$ and takes value in the compact set Θ. By continuity of τ, there exists $C > 0$ such that

$$L_n = \sup_{Q \in B_H(P_*, r/\sqrt{n})} n\left(\tau \circ \bar{T}(Q) - \tau(\theta^*)\right)^2 \leq C \cdot n < \infty.$$

Then, by definition of sup, there exists a sequence $\bar{Q}_n$ in $B_H(P_*, r/\sqrt{n})$ such that

$$L_n \leq n\left(\tau \circ \bar{T}(\bar{Q}_n) - \tau(\theta^*)\right)^2 + \frac{1}{2^n}.$$

Thus

$$\limsup_{n\to\infty} L_n \leq \limsup_{n\to\infty} n\left(\tau \circ \bar{T}(\bar{Q}_n) - \tau(\theta^*)\right)^2$$

and using (C.2), we deduce that $\limsup_{n\to\infty} L_n \leq 4r^2 B^*$. This establishes (20) recalling (C.1). □

Proof of Theorem 4.2: We proceed in two steps. First, we show that $\bar{T}$ is regular. Then, applying Theorem 3.2(i) of KOE (2013a), we can claim that, for each $r > 0$,

$$\lim_{b\to\infty} \lim_{\delta\to\infty} \liminf_{n\to\infty} \sup_{Q \in \bar{B}_H^{\delta}(P_*, r/\sqrt{n})} \int b \wedge n(\tau \circ T(P_n) - \tau(\theta^*))^2 dQ^{\otimes n} \geq (1 + 4r^2)B^*. \tag{C.3}$$

In a second step, we establish that limit superior is less or equal to $(1 + 4r^2)B^*$.

Consider again the sub-model P_{θ_n,ζ_n} as introduced in the proof of Theorem 4.1. We show that:

$$\sqrt{n}\left(T(P_n) - T(P_{\theta_n,\zeta_n})\right) \xrightarrow{d} N(0,\Sigma),$$

under P_{θ_n,ζ_n}. We have

$$\sqrt{n}\left(T(P_n) - T(P_{\theta_n,\zeta_n})\right) = \sqrt{n}\left[\left(T(P_n) - \bar{T}(P_n)\right) + \left(\bar{T}(P_n) - \bar{T}(P_{\theta_n,\zeta_n})\right) + \left(\bar{T}(P_{\theta_n,\zeta_n}) - T(P_{\theta_n,\zeta_n})\right)\right].$$

Note that from Lemma C.7, $\sqrt{n}\left(\bar{T}(P_n) - \bar{T}(P_{\theta_n,\zeta_n})\right)$ converges in distribution to $N(0,\Sigma)$ under P_{θ_n,ζ_n}. Hence, only need to show that

$$(a) \quad \sqrt{n}(\bar{T}(P_{\theta_n,\zeta_n}) - T(P_{\theta_n,\zeta_n}) = o(1) \quad \text{and} \quad (b) \quad \sqrt{n}(T(P_n) - \bar{T}(P_n)) = o_P(1) \quad \text{under} \quad P_{\theta_n,\zeta_n}.$$

To show (a), it is not hard to see, for n large enough, that $T(P_{\theta_n,\zeta_n}) = \theta_n$. Hence,

$$\sqrt{n}(\bar{T}(P_{\theta_n,\zeta_n}) - T(P_{\theta_n,\zeta_n})) = \sqrt{n}(\bar{T}(P_{\theta_n,\zeta_n}) - \theta^*) - \sqrt{n}(\theta_n - \theta^*) = \sqrt{n}(\bar{T}(P_{\theta_n,\zeta_n}) - \theta^*) - t = o(1)$$

by Fisher consistency of $\bar{T}$. To show (b), by similar reasoning as in the proof of Lemma C.7(i), it suffices to show that $\sqrt{n}(T(P_n) - \bar{T}(P_n)) = o_{P_*}(1)$. We first observe that

$$\sqrt{n}(T(P_n) - \bar{T}(P_n)) = \sqrt{n}(T(P_n) - \theta^*) - \sqrt{n}(\bar{T}(P_n) - \theta^*) = O_{P_*}(1) + O_{P_*}(1) = O_{P_*}(1),$$

where the orders of magnitude follow from Theorem 3.3 and Lemma C.7(i).
Let $\epsilon > 0$ and consider $P_*\left(\|\sqrt{n}(T(P_n) - \bar{T}(P_n))\| > \epsilon\right)$ that we show converges to 0 as $n \to \infty$. For this, let $\nu > 0$. By uniform tightness of $\sqrt{n}(T(P_n) - \bar{T}(P_n))$, there exists $\eta > \epsilon$ such that

$$\sup_n P_*\left(\|\sqrt{n}(T(P_n) - \bar{T}(P_n))\| > \eta\right) < \nu/2$$

and we have:

$$\begin{array}{rl} & P_*\left(\|\sqrt{n}(T(P_n) - \bar{T}(P_n))\| \geq \epsilon\right) \\ = & P_*\left(\epsilon \leq \|\sqrt{n}(T(P_n) - \bar{T}(P_n))\| \leq \eta\right) + P_*\left(\|\sqrt{n}(T(P_n) - \bar{T}(P_n))\| > \eta\right) \\ \leq & P_*\left(\epsilon \leq \|\sqrt{n}(T(P_n) - \bar{T}(P_n))\| \leq \eta\right) + \frac{\nu}{2}. \end{array}$$

Note that, for all X, $\epsilon\mathbb{I}\{\epsilon \leq \|X\| \leq \eta\} \leq \|X\| \wedge \eta$. Thus,

$$P_*\left(\epsilon \leq \|\sqrt{n}(T(P_n) - \bar{T}(P_n))\| \leq \eta\right) \leq \frac{1}{\epsilon^2} E_{P_*}\left(\|\sqrt{n}(T(P_n) - \bar{T}(P_n))\|^2 \wedge \eta^2\right).$$

But we know that if $(X_1,\ldots,X_n) \in \mathcal{X}_n^{\otimes n}$, (with the notation ($A^{\otimes n} = A \times \cdots \times A$, n-fold), $\bar{T}(P_n) = T(P_n)$. So,

$$\begin{array}{rl} & E_{P_*}\left(\|\sqrt{n}(T(P_n) - \bar{T}(P_n))\|^2 \wedge \eta^2\right) \\ = & \displaystyle\int_{(X_1,\ldots,X_n)\notin\mathcal{X}_n^{\otimes n}} \|\sqrt{n}(T(P_n) - \bar{T}(P_n))\|^2 \wedge \eta^2 dP_*^{\otimes n} \leq \eta^2 E_{P_*}(\mathbb{I}\{(X_1,\ldots,X_n) \notin \mathcal{X}_n^{\otimes n}\}) \\ \leq & \eta^2 \displaystyle\sum_{i=1}^n E_{P_*}(\mathbb{I}\{X_i \notin \mathcal{X}_n\}) = \eta^2 n P_*\left(\sup_{\theta\in\Theta} \|g(X,\theta)\| > m_n\right) \\ \leq & \eta^2 n m_n^{-\alpha} E_{P_*}\left(\sup_{\theta\in\Theta} \|g(X,\theta)\|^\alpha\right). \end{array}$$

Since $nm_n^{-\alpha} = n^{1-a\alpha} \to 0$ as $n \to \infty$, we claim that for n large enough,

$$P_*\left(\epsilon \leq \|\sqrt{n}(T(P_n) - \bar{T}(P_n))\| \leq \eta\right) \leq \frac{\nu}{2}$$

and we conclude that $\sqrt{n}(T(P_n) - \bar{T}(P_n)) = o_{P_*}(1)$ and (C.3) holds.

We now show that

$$\lim_{b\to\infty}\lim_{\delta\to\infty}\limsup_{n\to\infty} \sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})} \int b \wedge n(\tau\circ T(P_n) - \tau(\theta^*))^2 dQ^{\otimes n} \leq (1+4r^2)B^*. \tag{C.4}$$

We follow similar lines as in the proof of Theorem 3.2(ii) of KOE (2013a). Using the fact that, for all $b,c,d \geq 0$, $b \wedge (c+d) \leq b \wedge c + b \wedge d$, we have:

$$\begin{array}{rl} & \limsup_{n\to\infty} \sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})} \displaystyle\int b \wedge n\left(\tau\circ T(P_n) - \tau(\theta^*)\right)^2 dQ^{\otimes n} \\ = & \limsup_{n\to\infty} \sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})} \displaystyle\int b \wedge n\left((\tau\circ T(P_n) - \tau\circ\bar{T}(P_n)) + (\bar{T}(P_n) - \tau(\theta^*))\right)^2 dQ^{\otimes n} \\ \leq & A_1 + 2A_2 + A_3, \end{array}$$

with

$$\begin{aligned}
A_1 &= \limsup_{n\to\infty} \sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})} \int b\wedge n\left(\tau\circ T(P_n)-\tau\circ\bar{T}(P_n)\right)^2 dQ^{\otimes n},\\
A_2 &= \limsup_{n\to\infty} \sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})} \int b\wedge n\left|\tau\circ T(P_n)-\tau\circ\bar{T}(P_n)\right|\left|\tau\circ\bar{T}(P_n)-\tau(\theta^*)\right| dQ^{\otimes n},\\
A_3 &= \limsup_{n\to\infty} \sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})} \int b\wedge n\left(\tau\circ\bar{T}(P_n)-\tau(\theta^*)\right)^2 dQ^{\otimes n}.
\end{aligned}$$

We show that $A_1 = A_2 = 0$. As previously mentioned, $T(P_n) = \bar{T}(P_n)$ if $(X_1,\ldots,X_n)\in\mathcal{X}_n^{\otimes n}$. Thus,

$$\begin{aligned}
A_1 &\le b\times\limsup_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int_{(x_1,\ldots,x_n)\notin\mathcal{X}_n^{\otimes n}} dQ^{\otimes n} \le b\times\limsup_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\sum_{i=1}^n Q(X_i\notin\mathcal{X}_n)\\
&= b\times\limsup_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})} nQ\left(\sup_{\theta\in\Theta}\|g(X,\theta)\|\ge m_n\right) \qquad \text{(C.5)}\\
&\le b\times\limsup_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})} nm_n^{-\alpha}E_Q\left(\sup_{\theta\in\Theta}\|g(X,\theta)\|^\alpha\right)\le b\times\delta\times\limsup_{n\to\infty} nm_n^{-\alpha}=0.
\end{aligned}$$

$A_2 = 0$ is shown similarly. Consider A_3. Note that $\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int b\wedge n\left(\tau\circ\bar{T}(P_n)-\tau(\theta^*)\right)^2 dQ^{\otimes n}\le b<\infty$. Therefore, there exists $\bar{Q}_n\in\bar{B}_H^\delta(P_*,r/\sqrt{n})$ such that

$$\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int b\wedge n\left(\tau\circ\bar{T}(P_n)-\tau(\theta^*)\right)^2 dQ^{\otimes n}\le\int b\wedge n\left(\tau\circ\bar{T}(P_n)-\tau(\theta^*)\right)^2 d\bar{Q}^{\otimes n}+\frac{1}{2^n}.$$

Therefore,

$$A_3\le\limsup_{n\to\infty}\int b\wedge n\left(\tau\circ\bar{T}(P_n)-\tau(\theta^*)\right)^2 d\bar{Q}_n^{\otimes n}.$$

Note that, thanks to Lemma C.7, $\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau\circ\bar{T}(\bar{Q}_n))$ converges in distribution towards $N(0,B^*)$ under $\bar{Q}_n$. Let $\int b\wedge n\left(\tau\circ\bar{T}(P_n)-\tau(\theta^*)\right)^2 d\bar{Q}_n^{\otimes n}$ be a subsequence of this sequence that converge to the lim sup (we keep n to denote the subsequence for simplicity). This has a further subsequence along which $\sqrt{n}(\tau\circ\bar{T}(\bar{Q}_n)-\tau(\theta^*))$ converges towards its lim sup, say $\tilde{t}$. Thanks to Theorem 4.1, $\tilde{t}$ is finite. Hence, along this final subsequence,

$$\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau(\theta^*))=\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau\circ\bar{T}(\bar{Q}_n))+\sqrt{n}(\tau\circ\bar{T}(\bar{Q}_n)-\tau(\theta^*))$$

converges in distribution towards $N(\tilde{t},B^*)$ under $\bar{Q}_n$. Let $Z\sim N(0,B^*)$. We can claim that:

$$A_3\le\int b\wedge(Z+\tilde{t})^2 dN(0,B^*)\le B^*+\tilde{t}^2\le B^*+\limsup_{n\to\infty} n(\tau\circ\bar{T}(\bar{Q}_n)-\tau(\theta^*))^2\le B^*+4r^2B^*,$$

where the lim sup is taking over the initial sequence and the last inequality follows from Theorem 4.1. This establishes (C.4) which, along with (C.3) concludes the proof. □

Proof of Theorem 4.3: The Fisher consistency of $\bar{T}$ in the family of sub-models P_{θ_n,ζ_n} satisfying $E_{P_{\theta_n,\zeta_n}}[\sup_{\theta\in\Theta}\|g(X,\theta)\|^\alpha]\le\delta<\infty$ for some $\delta>0$ is established by Theorem 4.1and is sufficient to apply Theorem 3.3(i) of KOE (2013a) with $S_n=T(P_n)$. Thus, we have:

$$\lim_{b\to\infty}\lim_{\delta\to\infty}\lim_{r\to\infty}\liminf_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int b\wedge\ell(\sqrt{n}(\tau\circ T(P_n)-\tau\circ\bar{T}(Q)))dQ^{\otimes n}\ge\int\ell dN(0,B^*). \qquad \text{(C.6)}$$

To claim the expected result, it suffices to show that for all $b,r,\delta>0$,

$$\limsup_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int b\wedge\ell(\sqrt{n}(\tau\circ T(P_n)-\tau\circ\bar{T}(Q)))dQ^{\otimes n}\le\int\ell dN(0,B^*). \qquad \text{(C.7)}$$

We have:

$$\begin{aligned}
&\limsup_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int b\wedge\ell(\sqrt{n}(\tau\circ T(P_n)-\tau\circ\bar{T}(Q)))dQ^{\otimes n}\\
\le\ &\limsup_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int_{(X_1,\ldots,X_n)\notin\mathcal{X}_n^{\otimes n}} b\wedge\ell(\sqrt{n}(\tau\circ T(P_n)-\tau\circ\bar{T}(Q)))dQ^{\otimes n}\\
&+\limsup_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int_{(X_1,\ldots,X_n)\in\mathcal{X}_n^{\otimes n}} b\wedge\ell(\sqrt{n}(\tau\circ T(P_n)-\tau\circ\bar{T}(Q)))dQ^{\otimes n}
\end{aligned}$$

Using similar argument to that in (C.5), the first term is zero. Regarding the second term, we have:

$$\begin{aligned} & \sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})} \int_{(X_1,\ldots,X_n)\in\mathcal{X}_n^{\otimes n}} b\wedge\ell(\sqrt{n}(\tau\circ T(P_n)-\tau\circ\bar{T}(Q)))dQ^{\otimes n} \\ \leq & \sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})} \int b\wedge\ell(\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau\circ\bar{T}(Q)))dQ^{\otimes n}. \end{aligned}$$

Since $0\leq b\wedge\ell(\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau\circ\bar{T}(Q)))\leq b<\infty$, so is the supremum over $Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})$. Thus, similarly to the proof of Theorem 4.2, there exists $\bar{Q}_n\in\bar{B}_H^\delta(P_*,r/\sqrt{n})$ such that

$$\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int b\wedge\ell(\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau\circ\bar{T}(Q)))dQ^{\otimes n}\leq\int b\wedge\ell(\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau\circ\bar{T}(\bar{Q}_n)))d\bar{Q}_n^{\otimes n}+\frac{1}{2^n}.$$

As a result,

$$\begin{aligned} & \limsup_{n\to\infty}\sup_{Q\in\bar{B}_H^\delta(P_*,r/\sqrt{n})}\int b\wedge\ell(\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau\circ\bar{T}(Q)))dQ^{\otimes n} \\ \leq & \limsup_{n\to\infty}\int b\wedge\ell(\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau\circ\bar{T}(\bar{Q}_n)))d\bar{Q}_n^{\otimes n}. \end{aligned}$$

By Lemma C.4(ii), $\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau\circ\bar{T}(\bar{Q}_n))$ converges in distribution under $\bar{Q}_n$ to $N(0,B^*)$. Thus,

$$\limsup_{n\to\infty}\int b\wedge\ell(\sqrt{n}(\tau\circ\bar{T}(P_n)-\tau\circ\bar{T}(\bar{Q}_n)))d\bar{Q}_n^{\otimes n}=\int b\wedge\ell dN(0,B^*).$$

This establishes (C.7) which, along with (C.6) concludes the proof. □

C.2. **Auxiliary lemmas and proofs.**

Lemma C.1. *Let $Q\in\mathcal{M}$, $\mathscr{P}_\theta=\{P\in\mathcal{M}:E_P(g(X,\theta))=0\}$ with $\theta\in\Theta$ and $P(\theta)$ solution to $\min_{P\in\mathscr{P}_\theta}E_P\left[\log\left(\frac{dP}{dQ}\right)\right]$. We have*

$$\arg\min_{\theta\in\Theta}H(P(\theta),Q)=\arg\max_{\theta\in\Theta}\frac{E_Q[\exp(\lambda(\theta)'g(X,\theta)/2)]}{(E_Q[\exp(\lambda(\theta)'g(X,\theta))])^{1/2}},$$

with $\lambda(\theta)=\arg\min_{\lambda\in\Lambda}E_Q[\exp(\lambda(\theta)'g(X,\theta))]$.

Proof of Lemma C.1: From Kitamura and Stutzer (1997), the solution $P(\theta)$ to $\min_{P\in\mathscr{P}_\theta}E_P\left[\log\left(\frac{dP}{dQ}\right)\right]$ has the Gibbs canonical density with respect to Q given by:

$$\frac{dP(\theta)}{dQ}=\frac{\exp(\lambda(\theta)'g(X,\theta))}{E_Q[\exp(\lambda(\theta)'g(X,\theta))]}.$$

We can conclude using the fact that:

$$H(P(\theta),Q)^2=1-\int dP(\theta)^{1/2}dQ^{1/2}=1-E_Q\left[\left(\frac{dP(\theta)}{dQ}\right)^{1/2}\right].$$

□

Lemma C.2. *If Assumption 3 holds, then:*

(i) *For all $Q\in\mathcal{M}$ and $n\in\mathbb{N}$, $\bar{T}(Q)$ as given by (18) is well-defined.*

(ii) *There exists a neighborhood $\mathcal{V}_{\theta^*}$ of θ^* such that for any $r>0$, n large enough and any sequence $Q_n\in B_H(P_*,r/\sqrt{n})$, $\lambda_n:\theta\mapsto\bar{T}_1(\theta,Q_n)$ is a well-defined and continuous function on $\mathcal{V}_{\theta^*}$. Furthermore, λ_n is continuously differentiable on $int(\mathcal{V}_{\theta^*})$ and, for any $\theta\in int(\mathcal{V}_{\theta^*})$,*

$$\frac{\partial\lambda_n(\theta)}{\partial\theta'}=-A_n(\theta)^{-1}B_n(\theta),$$

where, letting $a_n(\theta)=\exp(\lambda_n(\theta)'g_n(X,\theta))$,

$$A_n(\theta)=E_{Q_n}\left[g_n(X,\theta)g_n(X,\theta)'a_n(\theta)\right],\quad B_n(\theta)=E_{Q_n}\left[(I_m+g_n(X,\theta)\lambda_n(\theta)')\frac{\partial g_n(X,\theta)}{\partial\theta'}a_n(\theta)\right].$$

In addition, for any sequence $(\theta_n)_n$ converging to θ_* as $n \to \infty$, we have

$$\begin{aligned}\frac{\partial \lambda_n(\theta_n)}{\partial \theta'} &= -\left(E_{Q_n}[g_n(X,\theta^*)g_n(X,\theta^*)']\right)^{-1} E_{Q_n}\left[\frac{\partial g_n(X,\theta^*)}{\partial \theta'}\right] + o(1) \\ &= -\left(E_{P_*}[g(X,\theta^*)g(X,\theta^*)']\right)^{-1} E_{P_*}\left[\frac{\partial g(X,\theta^*)}{\partial \theta'}\right] + o(1).\end{aligned} \tag{C.8}$$

Proof of Lemma C.2: (i) Let $Q \in \mathcal{M}$. The map $f_n : (\lambda,\theta) \mapsto E_Q[\exp(\lambda' g_n(X,\theta))]$ is continuous in both its arguments. Since Λ is compact, the Berge's maximum theorem (see Feinberg, Kasyanov and Zadoianchuk, 2013 and Feinberg, Kasyanov and Voorneveld, 2014) guarantees that $\theta \mapsto \bar{T}_1(\theta, Q) = \arg\min_{\lambda\in\Lambda} f_n(\lambda,\theta)$ is upper semi-continuous and compact-valued. Also, since $(\lambda,\theta) \mapsto \mathbf{\Delta}_{n,Q}(\lambda,\theta)$ is continuous in both arguments, $v(\theta) = \max_{\lambda\in\bar{T}_1(\theta,Q)} \mathbf{\Delta}_{n,Q}(\lambda,\theta)$ is upper semi-continuous on Θ. By the Weierstrass theorem, $v(\theta)$ takes a maximum value on Θ and $\bar{T}(Q)$ is therefore well-defined.

(ii) Let

$$\mathcal{V}_{\theta^*} = \{\theta \in \Theta : \|\theta - \theta^*\| \le \epsilon\delta/(2K+1)\}, \quad \text{and} \quad \Lambda_\epsilon = \{\lambda \in \mathbb{R}^m : \|\lambda\| \le 2\epsilon\},$$

with $K = E_{P_*}(\sup_{\theta\in\mathcal{N}} \|\partial g(X,\theta)/\partial\theta'\|)$, $\epsilon > 0$ sufficiently small so that $\mathcal{V}_{\theta^*} \subset \bar{\mathcal{N}} \subset \mathcal{N}$ and $\Lambda_\epsilon \subset \bar{\mathcal{V}} \subset \mathcal{V}$, $\delta > 0$ to be defined later and $\bar{\mathcal{N}}$ and $\bar{\mathcal{V}}$ compact neighborhoods of θ^* and 0, respectively.

Let $\theta \in \mathcal{V}_{\theta^*}$. We first show that $f_n : \lambda \mapsto E_{Q_n}[\exp(\lambda' g_n(X,\theta))]$ is strictly convex on the convex set Λ_ϵ for each $\theta \in \mathcal{V}_{\theta^*}$. Therefore, $\arg\min_{\lambda\in\Lambda_\epsilon} f_n(\lambda)$ is unique, $\lambda_{n,\epsilon}(\theta)$. For this, we observe that the conditions in Assumption 3 ensure that f_n is twice differentiable with

$$\frac{\partial^2 f_n(\lambda)}{\partial\lambda\partial\lambda'} = E_{Q_n}\left[g_n(X,\theta)g_n(X,\theta)'\exp(\lambda' g_n(X,\theta))\right].$$

Under Assumption 3(vii),

$$\frac{\partial^2 f_n(\lambda)}{\partial\lambda\partial\lambda'} = E_{P_*}\left[g(X,\theta)g(X,\theta)'\exp(\lambda' g(X,\theta))\right] + o(1)$$

where the neglected term is uniform over $\mathcal{V}_{\theta^*} \times \Lambda_\epsilon$. Note that $E_{P_*}\left[g(X,\theta)g(X,\theta)'\exp(\lambda' g(X,\theta))\right]$ is singular if and only if $E_{P_*}\left[g(X,\theta)g(X,\theta)'\right]$ is so. By Assumption 3(v), this latter is nonsingular over Θ. Thus, for all $(\lambda,\theta) \in \bar{\mathcal{N}} \times \bar{\mathcal{V}}$, the determinant of $E_{P_*}\left[g(X,\theta)g(X,\theta)'\exp(\lambda' g(X,\theta))\right]$ is strictly positive. By continuity of eigenvalues function and compactness of $\bar{\mathcal{N}} \times \bar{\mathcal{V}}$, the smallest eigenvalue of this matrix is bounded from below by 2δ for some $\delta > 0$. Therefore, for n large enough, the smallest eigenvalue of $\frac{\partial^2 f_n(\lambda)}{\partial\lambda\partial\lambda'}$ is bounded from below by δ.

Next, we show that $\lambda_{n,\epsilon}(\theta)$ is interior to Λ_ϵ. In this case, since f_n is convex on Λ, $\lambda_{n,\epsilon}(\theta)$ is also unique global minimum, hence equal to $\lambda_n(\theta)$ which is therefore well-defined on $\mathcal{V}_{\theta^*}$ and Berge's maximum theorem ensures that this function is continuous.

By the definition of minimum and a second order mean value expansion of f_n at $\lambda_{n,\epsilon}(\theta)$ around 0, we obtain

$$\frac{1}{2}\lambda_{n,\epsilon}(\theta)' E_{Q_n}\left[g_n(X,\theta)g_n(X,\theta)'\exp(\dot{\lambda}' g_n(X,\theta))\right]\lambda_{n,\epsilon}(\theta) \le -E_{Q_n}[g_n(X,\theta)']\lambda_{n,\epsilon}(\theta),$$

with $\dot{\lambda} \in (0, \lambda_{n,\epsilon}(\theta))$. From the previous lines, $E_{Q_n}\left[g_n(X,\theta)g_n(X,\theta)'\exp(\dot{\lambda}' g_n(X,\theta))\right]$ has its smallest eigenvalue bounded away from 0 by δ for n large enough. So, this inequality implies that,

$$\delta\|\lambda_{n,\epsilon}(\theta)\|^2 \le \|E_{Q_n}(g_n(X,\theta))\|\|\lambda_{n,\epsilon}(\theta)\|.$$

Hence,

$$\delta\|\lambda_{n,\epsilon}(\theta)\| \le \sup_{\theta\in\Theta} \|E_{Q_n}(g_n(X,\theta)) - E_{P_*}(g(X,\theta))\| + \|E_{P_*}(g(X,\theta))\| \equiv (1) + (2).$$

From the proof of Lemma A.1(ii) of KOE (2013b), (1) converges to 0 as n grows and hence, is less than $\epsilon/2$ for n large enough. By a mean value expansion and with $\dot{\theta} \in (\theta^*, \theta)$ that may vary with rows, we have

$$\|E_{P_*}(g(X,\theta))\| = \left\|E_{P_*}(g(X,\theta^*)) + E_{P_*}\left(\frac{\partial g(X,\dot{\theta})}{\partial\theta'}\right)(\theta - \theta^*)\right\| \le E_{P_*}\left(\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial g(X,\dot{\theta})}{\partial\theta'}\right\|\right)\|\theta - \theta^*\|$$

and $(2) \le K\delta\epsilon/(2K+1)$. Also, for n large enough, $(1) \le \delta\epsilon/2$. Thus, $\|\lambda_{n,\epsilon}(\theta)\| \le \epsilon < 2\epsilon$, showing that $\lambda_{n,\epsilon}(\theta)$ is interior to Λ_ϵ.

We now establish the differentiability of $\theta \mapsto \lambda_n(\theta)$ by relying on a global implicit function theorem. Since $\lambda_n(\theta)$ is interior minimum, it solves the first order condition

$$F_n(\lambda, \theta) \equiv E_{Q_n}[g_n(X,\theta)\exp(\lambda' g_n(X,\theta))] = 0. \tag{C.9}$$

Note that $(\lambda,\theta) \mapsto F_n(\lambda,\theta)$ is continuously differentiable in both its arguments on $\Lambda_\epsilon \times \mathcal{V}_{\theta^*}$ and all the other conditions of the global implicit function theorem of Sandberg (1981, Corollary 1) are fulfilled (in particular, for every $\theta \in \mathcal{V}_{\theta^*}$, (C.9) has a unique solution in Λ_ϵ and second derivatives in the direction of λ are nonsingular) and we can conclude that the implicit function $\lambda_n(\theta)$ determined by (C.9) is continuously differentiable on $int(\mathcal{V}_{\theta^*})$ with derivative's expression given by the lemma.

Let us now consider $(\theta_n)_n$, a sequence of elements of Θ converging to θ^* as $n \to \infty$. For n large enough, θ belongs to $int(\mathcal{V}_{\theta^*})$ and by a mean value expansion,

$$\lambda_n(\theta_n) - \lambda_n(\theta^*) = \frac{\partial \lambda_n(\dot\theta)}{\partial \theta'}(\theta_n - \theta^*),$$

with $\dot\theta \in (\theta_n, \theta^*)$ and may differ by row. It is not hard to see that $\left\| \frac{\partial \lambda_n(\dot\theta)}{\partial \theta'} \right\|$ is bounded. From Lemma C.4, for n large enough, $\lambda_n(\theta^*) = \bar T_1(\theta^*, Q_n) \to 0$ as $n \to \infty$. Thus, $\lambda_n(\theta_n) \to 0$, as $n \to \infty$. Also

$$A_n(\theta_n) = E_{P_*}\left[g(X,\theta_n)g(X,\theta_n)'\exp(\lambda_n(\theta_n)'g(X,\theta_n))\right] + o(1)$$

and by the Lebesgue dominated convergence theorem, $A_n(\theta_n) = E_{P_*}[g(X,\theta^*)g(X,\theta^*)'] + o(1)$. Although a bit more tedious, one obtains along similar lines that $B_n(\theta_n) = E_{P_*}\left(\frac{\partial g(X,\theta^*)}{\partial \theta'}\right) + o(1)$. □

Lemma C.3. *If Assumption 3 holds, then, for each $r > 0$ and any sequence $Q_n \in B_H(P_*, r/\sqrt{n})$,*

(i) $\mathbf{\Delta}_{n,Q_n}(\bar T_1(\bar T_{Q_n}, Q_n), \bar T_{Q_n}) = 1 + O(n^{-1})$,
(ii) $\bar T_1(\bar T_{Q_n}, Q_n) = O(n^{-1/2})$,
(iii) $E_{Q_n}(g_n(X, \bar T_{Q_n})) = O(n^{-1/2})$,
(iv) $\bar T_{Q_n} \to \theta^*$ *as* $n \to \infty$.

Proof of Lemma C.3: (i) By the definition of $\bar T_{Q_n}$, and concavity of $x \mapsto \sqrt{x}$, we have

$$\mathbf{\Delta}_{n,Q_n}(\bar T_1(\theta^*, Q_n), \theta^*) \le \mathbf{\Delta}_{n,Q_n}(\bar T_1(\bar T_{Q_n}, Q_n), \bar T_{Q_n}) \le 1$$

and by Lemma C.4, we deduce that $\mathbf{\Delta}_{n,Q_n}(\bar T_1(\bar T_{Q_n}, Q_n), \bar T_{Q_n}) = 1 + O(n^{-1})$.

(ii) Since $E_{P_*}[\exp(\lambda' g(X,\theta))]$ is continuous on $\Lambda \times \Theta$, it has a minimum, hence $E_{P_*}[\exp(\lambda' g(X,\theta))]$ is bounded away from 0 on $\Lambda \times \Theta$. This is enough, using Assumption 3(vii), to claim that the ratio $\mathbf{\Delta}_{n,Q_n}(\lambda,\theta)$ converges to $\mathbf{\Delta}_{P_*}(\lambda,\theta)$ uniformly over $\Lambda \times \Theta$. Also, thanks to (i), the conditions of Lemma B.1 are satisfied and we can claim that $\hat\lambda_n \equiv \bar T_1(\bar T_{Q_n}, Q_n) \to 0$ as $n \to \infty$.

By a second-order Taylor expansion of $\lambda \mapsto \mathbf{\Delta}_{n,Q_n}(\lambda, \bar T_{Q_n})$ at $\hat\lambda_n$ around 0, we have:

$$\mathbf{\Delta}_{n,Q_n}(\hat\lambda_n, \hat\theta) = \mathbf{\Delta}_{n,Q_n}(0,\hat\theta) + \frac{\partial \mathbf{\Delta}_{n,Q_n}(0,\hat\theta)}{\partial \lambda'}\hat\lambda_n + \frac{1}{2}\hat\lambda_n' \frac{\partial^2 \mathbf{\Delta}_{n,Q_n}(\dot\lambda,\hat\theta)}{\partial\lambda\partial\lambda'}\hat\lambda_n, \tag{C.10}$$

with $\hat\theta \equiv \bar T_{Q_n}$ and $\dot\lambda \in (0, \hat\lambda_n)$. The first and second partial derivatives of $\mathbf{\Delta}_{n,Q_n}(\lambda,\theta)$ are given in the proof of Lemma C.4. Let us admit for now that:

$$N_{1,n}(\dot\lambda,\hat\theta) = E_{P_*}(g(X,\hat\theta)) + o(1),\ \ N_{2,n}(\dot\lambda,\hat\theta) = E_{P_*}(g(X,\hat\theta)g(X,\hat\theta)') + o(1),\ \ D_n(\dot\lambda,\hat\theta) = 1 + o(1), \tag{C.11}$$

for all sequence $\dot\lambda \to 0$. Then,

$$\frac{\partial^2 \mathbf{\Delta}_{n,Q_n}(\dot\lambda,\hat\theta)}{\partial\lambda\partial\lambda'} = -\frac{1}{4}Var_{P_*}(g(X,\hat\theta)) + o(1).$$

Hence, (C.10) becomes

$$-\frac{1}{8}\hat\lambda_n' Var_{P_*}(g(X,\hat\theta))\hat\lambda_n + o(\|\hat\lambda_n\|^2) + 1 = 1 + O(n^{-1}).$$

Or, equivalently,

$$\hat\lambda_n' Var_{P_*}(g(X,\hat\theta))\hat\lambda_n + o(\|\hat\lambda_n\|^2) = O(n^{-1}).$$

Thanks to Assumption 3(v), this implies that $\underline{\ell}\|\hat\lambda_n\|^2 + o(|\hat\lambda_n\|^2) = O(n^{-1})$, and in particular that $\hat\lambda_n = O(n^{-1/2})$.

To complete the proof, we establish (C.11). Note that

$$D_n(\dot{\lambda},\hat{\theta}) = \left(E_{Q_n}[\exp(\dot{\lambda}'g(X,\hat{\theta}))] - E_{P_*}[\exp(\dot{\lambda}'g(X,\hat{\theta}))]\right) + E_{P_*}[\exp(\dot{\lambda}'g(X,\hat{\theta}))].$$

The term in the brackets converges to 0 and, by the dominance condition in Assumption 3(vii), lim and E_{P_*} can interchange and the fact that $g(X,\hat{\theta}) = O_{P_*}(1)$ implies that $\lim_n E_{P_*}[\exp(\dot{\lambda}'g(X,\hat{\theta}))] = 1$.

Similarly,

$$N_{2,n}(\dot{\lambda},\hat{\theta}) = E_{P_*}[g(X,\hat{\theta})g(X,\hat{\theta})'] + E_{P_*}\left[g(X,\hat{\theta})g(X,\hat{\theta})'\left(\exp(\dot{\lambda}'g(X,\hat{\theta})) - 1\right)\right] + o(1).$$

We have $\left\|g(X,\hat{\theta})g(X,\hat{\theta})'\left(\exp(\dot{\lambda}'g(X,\hat{\theta})) - 1\right)\right\| \leq Z$, with

$Z = \sup_{\theta\in\mathcal{N}}\|g(X,\theta)\|^2\left(\sup_{(\lambda,\theta)\in\mathbf{v}\times\mathcal{N}}\exp(\lambda'g(X,\theta)) + 1\right)$, where $\mathbf{v}$ is a small neighborhood of 0 contained in $\mathcal{V}$. By the Hölder inequality,

$$E_{P_*}(Z) \leq \left(E_{P_*}\sup_{\theta\in\mathcal{N}}\|g(X,\theta)\|^{\alpha}\right)^{2/\alpha}\left(E_{P_*}\left[\sup_{(\lambda,\theta)\in\mathbf{v}\times\mathcal{N}}\exp(\lambda'g(X,\theta)) + 1\right]^{\alpha/(\alpha-2)}\right)^{1-2/\alpha}.$$

By the c_r-inequality,

$$\begin{aligned}E_{P_*}\left(\sup_{(\lambda,\theta)\in\mathbf{v}\times\mathcal{N}}\exp(\lambda'g(X,\theta)) + 1\right)^{\alpha/(\alpha-2)} &\leq 2^{2/(\alpha-2)}E_{P_*}\left(\sup_{(\lambda,\theta)\in\mathbf{v}\times\mathcal{N}}\exp\left(\frac{\alpha}{\alpha-2}\lambda'g(X,\theta)\right) + 1\right)\\ &\leq 2^{2/(\alpha-2)}E_{P_*}\left(\sup_{(\lambda,\theta)\in\mathcal{V}\times\mathcal{N}}\exp(\lambda'g(X,\theta)) + 1\right),\end{aligned}$$

showing that $E_{P_*}(Z) < \infty$. Therefore, we pass lim through E_{P_*} and claim the result. Conclusion for $N_{1,n}(\dot{\lambda},\hat{\theta})$ is reached similarly. This completes (ii).

(iii) This is obtained along the same lines as Step 3 in the proof of Lemma B.2 with g_n, Q_n, $\hat{\lambda}_n$ and $\bar{T}_{Q_n}$ replacing g, P_n, $\hat{\lambda}(\hat{\theta})$ and $\hat{\theta}$, respectively.

(iv) Along the same lines as KOE's (2013b) proof of their Lemma A.1(ii), we can show that:

$$\sup_{\theta\in\Theta}|E_{Q_n}(g_n(X,\theta)) - E_{P_*}(g(X,\theta))| \to 0,$$

as $n\to\infty$. Also, from (iii) of the lemma, we have $E_{Q_n}(g_n(X,\bar{T}_{Q_n})) = O(n^{-1/2})$. Thus

$$\left|E_{P_*}(g(X,\bar{T}_{Q_n}))\right| \leq \left|E_{P_*}(g(X,\bar{T}_{Q_n})) - E_{Q_n}(g_n(X,\bar{T}_{Q_n}))\right| + \left|E_{Q_n}(g_n(X,\bar{T}_{Q_n}))\right|,$$

implies that $E_{P_*}(g(X,\bar{T}_{Q_n})) \to 0$ as $n\to\infty$. Since $\theta \mapsto E_{P_*}(g(X,\theta))$ is continuous and Θ is compact, the identification condition in Assumption 3(ii) allows us to conclude that $\bar{T}_{Q_n} \to \theta^*$ as $n\to\infty$. □

Lemma C.4. *If Assumption 3 holds, then, for each $r > 0$ and any sequence $Q_n \in B_H(P_*, r/\sqrt{n})$,*

(i) $\bar{T}_1(\theta^*, Q_n) = O(n^{-1/2})$,
(ii) $\boldsymbol{\Delta}_{n,Q_n}(\bar{T}_1(\theta^*,Q_n),\theta^*) = 1 + O(n^{-1})$.

Proof of Lemma C.4: (i) The map $f_n : \lambda \mapsto -E_{Q_n}[\exp(\lambda'g_n(X,\theta^*)]$ is continuous on Λ, so it has at least one maximum $\bar{T}_1(\theta^*,Q_n)$. Let $\Lambda_n = \{\lambda\in\mathbb{R}^m : \|\lambda\| \leq c/m_n^{1+\zeta}\}$ with $c > 0$ and $0 < \zeta < -1 + 1/2a$ so that $\sqrt{n}/m_n^{1+\zeta} \to \infty$ as $n\to\infty$. Let $\tilde{T}_1(\theta^*,Q_n) = \arg\max_{\lambda\in\Lambda_n} f_n(\lambda)$. Under Assumption 3, f_n is twice differentiable and

$$\frac{\partial^2 f_n}{\partial\lambda\partial\lambda'}(\lambda) = -E_{Q_n}\left(g_n(X,\theta^*)g_n(X,\theta^*)'\exp(\lambda'g_n(X,\theta^*))\right).$$

From Lemma C.6, $\frac{\partial^2 f_n}{\partial\lambda\partial\lambda'}(\lambda) = -E_{P_*}\left(g(X,\theta^*)g(X,\theta^*)'\right) + o(1)$ as $n\to\infty$. Therefore, f_n is strictly concave on Λ_n and thus has a unique maximum for n large enough. Let $\tilde{\lambda}_n = \tilde{T}_1(\theta^*,Q_n)$. By a second-order mean-value expansion of $f_n(\tilde{\lambda}_n)$ around 0, we have:

$$f_n(\tilde{\lambda}_n) = -1 - E_{Q_n}[g_n(X,\theta^*)']\tilde{\lambda}_n - \frac{1}{2}\tilde{\lambda}_n'E_{Q_n}\left(g_n(X,\theta^*)g_n(X,\theta^*)'\exp(\dot{\lambda}'g_n(X,\theta^*))\right)\tilde{\lambda}_n,$$

with $\dot{\lambda} \in (0, \tilde{\lambda}_n)$. By definition, $f_n(\tilde{\lambda}_n) \geq -1$, hence:

$$\frac{1}{2}\tilde{\lambda}_n' E_{Q_n}\left(g_n(X,\theta^*)g_n(X,\theta^*)'\exp(\dot{\lambda}' g_n(X,\theta^*))\right)\tilde{\lambda}_n \leq E_{Q_n}[g_n(X,\theta^*)']\tilde{\lambda}_n.$$

Using once again Lemma C.6 and the fact that $Var_{P_*}(g(X,\theta^*))$ is nonsingular, we can write

$$C\|\tilde{\lambda}_n\|^2 + o(\|\tilde{\lambda}_n\|^2) \leq \|\tilde{\lambda}_n\|\|E_{Q_n}[g_n(X,\theta^*)]\|,$$

for some $C > 0$. Along similar lines as in the proof of Lemma A.4(i) of KOE (2013b), we can readily show that $E_{Q_n}[g_n(X,\theta^*)] = O(n^{-1/2})$. Thus, $\tilde{\lambda}_n = O(n^{-1/2})$. As a result, we can claim that $\tilde{\lambda}_n$ is an interior maximum of $f_n(\lambda)$ over Λ_n and so, is the global maximum over Λ. Thus $\bar{T}_1(\theta^*, Q_n) = \tilde{\lambda}_n = O(n^{-1/2})$. This establishes (i).

(ii) $\lambda \mapsto \mathbf{\Delta}_{n,Q_n}(\lambda,\theta^*)$ is also differentiable up to order 2 with second-order mean-value expansion at $\bar{T}_1(\theta^*, Q_n) = \tilde{\lambda}_n$ around 0 given by

$$\mathbf{\Delta}_{n,Q_n}(\tilde{\lambda}_n,\theta^*) = 1 + \frac{\partial \mathbf{\Delta}_{n,Q_n}}{\partial \lambda'}(0,\theta^*)\tilde{\lambda}_n + \frac{1}{2}\tilde{\lambda}_n' \frac{\partial^2 \mathbf{\Delta}_{n,Q_n}}{\partial\lambda\partial\lambda'}(\dot{\lambda},\theta^*)\tilde{\lambda}_n,$$

with $\dot{\lambda} \in (0, \tilde{\lambda}_n)$. Note that

$$\frac{\partial \mathbf{\Delta}_{n,Q_n}}{\partial\lambda}(\lambda,\theta) = \frac{1}{2}\left(\frac{N_{1,n}(\lambda/2,\theta)}{D_n(\lambda,\theta)^{1/2}} - \frac{N_{1,n}(\lambda,\theta)D_n(\lambda/2,\theta)}{D_n(\lambda,\theta)^{3/2}}\right)$$

and

$$\begin{aligned}\frac{\partial^2 \mathbf{\Delta}_{n,Q_n}}{\partial\lambda\partial\lambda'}(\lambda,\theta) =& \frac{1}{4}\frac{N_{2,n}(\lambda/2,\theta)}{D_n(\lambda,\theta)^{1/2}} - \frac{1}{2}\frac{N_{2,n}(\lambda,\theta)D_n(\lambda/2,\theta)}{D_n(\lambda,\theta)^{3/2}} - \frac{1}{4}\frac{N_{1,n}(\lambda/2,\theta)N_{1,n}(\lambda,\theta)'}{D_n(\lambda,\theta)^{3/2}} \\ &- \frac{1}{4}\frac{N_{1,n}(\lambda,\theta)N_{1,n}(\lambda/2,\theta)'}{D_n(\lambda,\theta)^{3/2}} + \frac{3}{4}\frac{N_{1,n}(\lambda,\theta)N_{1,n}(\lambda,\theta)'D_n(\lambda/2,\theta)}{D_n(\lambda,\theta)^{5/2}},\end{aligned}$$

with $N_{1,n}(\lambda,\theta) = E_{Q_n}[g_n(X,\theta)\exp(\lambda' g_n(X,\theta))]$, $N_{2,n}(\lambda,\theta) = E_{Q_n}[g_n(X,\theta)g_n(X,\theta)'\exp(\lambda' g_n(X,\theta))]$, and $D_n(\lambda,\theta) = E_{Q_n}[\exp(\lambda' g_n(X,\theta))]$.

Clearly, $\frac{\partial \mathbf{\Delta}_{n,Q_n}}{\partial\lambda}(0,\theta) = 0$ and, from Lemma C.6,
$N_{1,n}(\dot{\lambda},\theta^*) = E_{P_*}(g(X,\theta^*)) + o(1)$, $N_{2,n}(\dot{\lambda},\theta^*) = E_{P_*}(g(X,\theta^*)g(X,\theta^*)') + o(1)$ and $D_n(\dot{\lambda},\theta^*) = 1 + o(1)$, same as $N_{1,n}(\dot{\lambda}/2,\theta^*)$, $N_{2,n}(\dot{\lambda}/2,\theta^*)$ and $D_{1,n}(\dot{\lambda}/2,\theta^*)$, respectively. Hence,

$$\frac{\partial^2 \mathbf{\Delta}_{n,Q_n}}{\partial\lambda\partial\lambda'}(\dot{\lambda},\theta^*) = -\frac{1}{4}Var_{P_*}(g(X,\theta^*)) + o(1).$$

as a result, using (i), we can claim that (ii) holds. □

Lemma C.5. *If Assumption 3 holds, then: for each $r > 0$ and any sequence $Q_n \in B_H(P_*, r/\sqrt{n})$,*

$$\sqrt{n}(\bar{T}_{Q_n} - \theta^*) = -\Sigma G'\Omega^{-1}\sqrt{n}E_{Q_n}\left(g_n(X,\theta^*)\right) + o(1). \tag{C.12}$$

Proof of Lemma C.5: Let $\hat{\theta}_n \equiv \bar{T}_{Q_n}$, $\lambda_n(\theta) \equiv \bar{T}_1(\theta, Q_n)$ and $\hat{\lambda}_n = \lambda_n(\hat{\theta}_n)$. Since Since $\hat{\theta}_n \to \theta^*$, Lemma C.2(ii) ensures that $\theta \mapsto \lambda_n(\theta)$ is differentiable at $\hat{\theta}_n$ for n large enough. Also, $\theta \mapsto E_{Q_n}[\exp(\lambda_n(\theta)' g_n(X,\theta))]$ and $\theta \mapsto E_{Q_n}[\exp(\lambda_n(\theta)' g_n(X,\theta)/2)]$ are both differentiable at $\hat{\theta}_n$. As an interior optimum, $\hat{\theta}_n$ satisfies the first order optimality condition

$$\frac{d}{d\theta}\mathbf{\Delta}_{n,Q_n}(\lambda_n(\theta),\theta)\Big|_{\theta=\hat{\theta}_n} = 0,$$

that is

$$\frac{N_{1n}(\hat{\lambda}_n,\hat{\theta}_n)}{D_{1n}(\hat{\lambda}_n,\hat{\theta}_n)} - \frac{N_{2n}(\hat{\lambda}_n,\hat{\theta}_n)}{D_{2n}(\hat{\lambda}_n,\hat{\theta}_n)} = 0, \tag{C.13}$$

with $N_{jn}(\lambda,\theta)$, $D_{jn}(\lambda,\theta)$ $(j = 1, 2)$ defined similarly to $N_j(\lambda,\theta)$, $D_j(\lambda,\theta)$ in (B.6) with $\hat{\lambda}(\theta)$, P_n and g replaced by $\lambda_n(\theta)$, Q_n and g_n, respectively.

Also, since $\hat{\lambda}_n$ converges to 0, it is also an interior solution for n large enough and therefore solves the first order optimality condition

$$E_{Q_n}\left[g_n(X,\hat{\theta}_n)\exp\left(\hat{\lambda}_n' g_n(X,\hat{\theta}_n)\right)\right] = 0. \tag{C.14}$$

We proceed to a mean value expansion of (C.13) and (C.14) around $(0,\theta^*)$.

Note that $N_{2n}(0,\theta^*) = N_{1n}(0,\theta^*)+o(1/\sqrt{n}) = \frac{1}{2}\frac{d\lambda_n'(\hat\theta_n)}{d\theta}E_{Q_n}(g_n(X,\theta^*))+o(1/\sqrt{n})$ and $D_{1n}(\hat\lambda_n,\hat\theta_n) = 1+o(1)$ and $D_{2n}(\hat\lambda_n,\hat\theta_n) = 1+o(1)$. Hence, a mean value expansion of (C.13) around $(0,\theta^*)$ yields

$$o(n^{-1/2}) = \frac{\partial}{\partial\theta'}\left(\frac{N_{1n}(\lambda,\theta)}{D_{1n}(\hat\lambda_n,\hat\theta_n)} - \frac{N_{2n}(\lambda,\theta)}{D_{2n}(\hat\lambda_n,\hat\theta_n)}\right)\Bigg|_{(\dot\lambda,\dot\theta)}(\hat\theta_n-\theta^*) + \frac{\partial}{\partial\lambda'}\left(\frac{N_{1n}(\lambda,\theta)}{D_{1n}(\hat\lambda_n,\hat\theta_n)} - \frac{N_{2n}(\lambda,\theta)}{D_{2n}(\hat\lambda_n,\hat\theta_n)}\right)\Bigg|_{(\dot\lambda,\dot\theta)}\hat\lambda_n, \tag{C.15}$$

with $(\dot\lambda,\dot\theta)\in(0,\hat\lambda_n)\times(\theta^*,\hat\theta_n)$ and may differ from row to row. The expressions of

$$\frac{\partial N_{jn}}{\partial\theta'}, \quad \frac{\partial N_{jn}}{\partial\lambda'},$$

$j=1,2$ are analogue to the expressions of the partial derivatives of N_j as given following (B.8) with, again, $\hat\lambda(\theta)$, g and P_n replaced by $\lambda_n(\theta)$, g_n and Q_n, respectively. Also, for n large enough, since $\hat\lambda_n = O(n^{-1/2})$, it belongs to Λ_n as defined in Lemma C.6 for some $0<\zeta<-1+1/2a$ and thanks to the same lemma, by using the fact that $E_{Q_n}(\partial g_n(X,\theta)/\partial\theta') = E_{P_*}(\partial g(X,\theta)/\partial\theta')+o(1)$ and $E_{Q_n}(g_n(X,\theta)g_n(X,\theta)') = E_{P_*}(g(X,\theta)g(X,\theta)')+o(1)$ for all θ in some neighborhood of θ^*, we have, for $j=1,2$:

$$\frac{\partial N_{jn}}{\partial\theta'}(\dot\lambda,\dot\theta) = \frac{1}{2}\frac{d\lambda_n(\hat\theta_n)'}{d\theta}E_{Q_n}\left(\frac{\partial g_n}{\partial\theta'}(X,\dot\theta)\right)+o(1),$$

$$\frac{\partial N_{1n}}{\partial\lambda'}(\dot\lambda,\dot\theta) = \frac{1}{2}E_{Q_n}\left(\frac{\partial g_n'}{\partial\theta}(X,\hat\theta_n)\right)+\frac{1}{4}\frac{d\lambda_n(\hat\theta_n)'}{d\theta}E_{Q_n}\left(g_n(X,\dot\theta)g_n(X,\hat\theta)'\right)+o(1),$$

and

$$\frac{\partial N_{2n}}{\partial\lambda'}(\dot\lambda,\dot\theta) = \frac{1}{2}E_{Q_n}\left(\frac{\partial g_n'}{\partial\theta}(X,\hat\theta_n)\right)+\frac{1}{2}\frac{d\lambda_n(\hat\theta_n)'}{d\theta}E_{Q_n}\left(g_n(X,\dot\theta)g_n(X,\hat\theta_n)'\right)+o(1).$$

As a result,

$$\frac{\partial}{\partial\theta'}\left(\frac{N_{1n}(\lambda,\theta)}{D_{1n}(\hat\lambda_n,\hat\theta_n)} - \frac{N_{2n}(\lambda,\theta)}{D_{2n}(\hat\lambda_n,\hat\theta_n)}\right)\Bigg|_{(\dot\lambda,\dot\theta)} = o(1)$$

and

$$\frac{\partial}{\partial\lambda'}\left(\frac{N_{1n}(\lambda,\theta)}{D_{1n}(\hat\lambda_n,\hat\theta_n)} - \frac{N_{2n}(\lambda,\theta)}{D_{2n}(\hat\lambda_n,\hat\theta_n)}\right)\Bigg|_{(\dot\lambda,\dot\theta)} = -\frac{1}{4}\frac{d\lambda_n(\hat\theta_n)'}{d\theta}E_{Q_n}\left(g_n(X,\dot\theta)g_n(X,\hat\theta_n)'\right)+o(1).$$

Also, from Lemma C.2(ii),

$$\frac{d\lambda_n(\hat\theta_n)}{d\theta'} = -\left(E_{Q_n}\left(g_n(X,\hat\theta_n)g_n(X,\hat\theta_n)'\right)\right)^{-1}E_{Q_n}\left(\frac{\partial g_n(X,\hat\theta_n)}{\partial\theta'}\right)+o(1).$$

The expansion in (C.15) becomes:

$$G'\sqrt{n}\hat\lambda_n = o(\|\sqrt{n}\hat\lambda_n\|)+o(\sqrt{n}\|\hat\theta_n-\theta^*\|). \tag{C.16}$$

A mean value expansion of (C.14) around $(0,\theta^*)$ yields:

$$\begin{aligned} 0 &= E_{Q_n}\left(g_n(X,\theta^*)\right)+E_{Q_n}\left[\left(I_m+g_n(X,\dot\theta)\dot\lambda'\right)\tfrac{\partial g_n}{\partial\theta'}(X,\dot\theta)\exp\left(\dot\lambda' g_n(X,\dot\theta)\right)\right](\hat\theta_n-\theta^*) \\ &\quad +E_{Q_n}\left[g_n(X,\dot\theta)g_n(X,\dot\theta)'\exp\left(\dot\lambda' g_n(X,\dot\theta)\right)\right]\hat\lambda_n, \end{aligned}$$

with $(\dot\lambda,\dot\theta)\in(0,\hat\lambda_n)\times(\theta^*,\hat\theta_n)$ and may differ from row to row. By similar arguments as above, we get:

$$G\sqrt{n}(\hat\theta_n-\theta^*)+\Omega\sqrt{n}\hat\lambda_n = -\sqrt{n}E_{Q_n}(g_n(X,\theta^*))+o(\|\sqrt{n}\hat\lambda_n\|)+o(\|\sqrt{n}(\hat\theta_n-\theta^*)\|). \tag{C.17}$$

Using (C.16) and (C.17) and solving for $(\hat\theta_n-\theta^*,\hat\lambda_n)$, we get

$$\sqrt{n}(\hat\theta_n-\theta^*)+o(\|\sqrt{n}(\hat\theta_n-\theta^*)\|) = -\sqrt{n}\Sigma G'\Omega^{-1}E_{Q_n}(g_n(X,\theta^*))+o(\|\sqrt{n}\hat\lambda_n\|)$$

which is sufficient to deduce the result. □

Lemma C.6. *Let $h(x,\theta)$ be a function measurable on $\mathcal{X}$ for each $\theta \in \Theta$ and taking value in $\mathbb{R}^\ell$. Let $\mathcal{X}_n = \{x \in \mathcal{X} : \sup_{\theta\in\Theta} \|g(x,\theta)\| \leq m_n\}$ with (m_n) a sequence of scalars satisfying $m_n \to 0$ as $n \to \infty$ and define $h_n(x,\theta) = h(x,\theta)\mathbb{I}(x \in \mathcal{X}_n)$. For some c, $\zeta > 0$, let $\Lambda_n = \left\{\lambda \in \mathbb{R}^m : \|\lambda\| \leq c/m_n^{1+\zeta}\right\}$ and let $\mathcal{N}$ be a subset of Θ. Let $r > 0$. If,*

$\sup_{\theta\in\mathcal{N},x\in\mathcal{X}_n} \|h(x,\theta)\| = o(n)$, $E_{P_*}\left(\sup_{\theta\in\mathcal{N}} \|h(X,\theta)\|^2\right) < \infty$, *and* $E_{P_*}\left(\sup_{\theta\in\mathcal{N}} \|g(X,\theta)\|\right) < \infty$, *then, uniformly over $Q_n \in B_H(P_*, r/\sqrt{n})$,*

$$\sup_{\lambda\in\Lambda_n,\theta\in\mathcal{N}} \|E_{Q_n}[h_n(X,\theta)\exp(\lambda' g_n(X,\theta))] - E_{P_*}(h(X,\theta))\| = o(1)$$

and

$$\sup_{\lambda\in\Lambda_n,\theta\in\mathcal{N}} \|E_{Q_n}[\exp(\lambda' g_n(X,\theta))] - 1\| = o(1).$$

Proof of Lemma C.6: We have:

$$\begin{array}{rl} & \|E_{Q_n}[h_n(X,\theta)\exp(\lambda' g_n(X,\theta))] - E_{P_*}(h(X,\theta))\| \\ \leq & \|E_{Q_n}[h_n(X,\theta)\exp(\lambda' g_n(X,\theta))] - E_{P_*}(h_n(X,\theta))\| + \|E_{P_*}(h_n(X,\theta)) - E_{P_*}(h(X,\theta))\| \equiv (1)+(2). \end{array}$$

Also, $(1) \leq (1.1) + (1.2)$ with

$$\begin{array}{rcl} (1.1) & = & \|E_{Q_n}[h_n(X,\theta)\exp(\lambda' g_n(X,\theta))] - E_{P_*}[h_n(X,\theta)\exp(\lambda' g_n(X,\theta))]\|, \\ (1.2) & = & \|E_{P_*}[h_n(X,\theta)(\exp(\lambda' g_n(X,\theta)) - 1)]\|. \end{array}$$

We next show that (1.1), (1.2) and (2) are all $o(1)$ uniformly on λ and θ.

$$\begin{array}{rcl} (1.1) & = & \left\| \int h_n(x,\theta)\exp(\lambda' g_n(x,\theta))(dQ_n - dP_*) \right\| \\ & = & \left\| \int h_n(x,\theta)\exp(\lambda' g_n(x,\theta)) \left\{ \left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 + 2dP_*^{1/2}\left(dQ_n^{1/2} - dP_*^{1/2}\right) \right\} \right\| \\ & \leq & \int \|h_n(x,\theta)\| \exp(\lambda' g_n(x,\theta)) \left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 \\ & & +2\left(\int \|h_n(x,\theta)\|^2 \exp(2\lambda' g_n(x,\theta)) dP_* \right)^{1/2} \left(\int \left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 \right)^{1/2} \end{array}$$

(the inequality is obtained using the triangle and the Cauchy-Schwarz inequalities.) By definition, for any $\lambda \in \Lambda_n$, $x \in \mathcal{X}$ and $\theta \in \Theta$,

$$|\lambda' g_n(x,\theta)| \leq \|\lambda\| \|g_n(x,\theta)\| \leq \frac{c}{m_n^\zeta} \to 0, \quad as \quad n \to \infty.$$

Thus, $\sup_{x\in\mathcal{X},\lambda\in\Lambda_n,\theta\in\Theta} \exp(\lambda' g_n(x,\theta)) \leq C$, a positive constant independent of n. As a result,

$$\begin{array}{rcl} (1.1) & \leq & C \sup_{x\in\mathcal{X}_n,\theta\in\mathcal{N}} \|h(x,\theta)\| \int \left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 \\ & & +2C\left(E_{P_*}\left(\sup_{\theta\in\mathcal{N}} \|h(x,\theta)\|^2 \right) \right)^{1/2} \left(\int \left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 \right)^{1/2} \\ & \leq & o(n)\frac{r^2}{n} + O(1)\frac{r}{\sqrt{n}} \to 0, \quad as \quad n \to \infty. \end{array}$$

By the Cauchy-Schwarz inequality,

$$\begin{array}{rcl} (1.2) & \leq & \left(E_{P_*}\|h(X,\theta)\|^2\right)^{1/2} \left(E_{P_*}(\exp(\lambda' g(X,\theta) - 1)^2\right)^{1/2} \\ & \leq & \left(E_{P_*} \sup_{\theta\in\mathcal{N}} \|h(X,\theta)\|^2\right)^{1/2} \left(E_{P_*} \sup_{\theta\in\mathcal{N},\lambda\in\Lambda_n} (\exp(\lambda' g(X,\theta) - 1)^2\right)^{1/2}. \end{array}$$

From the previous lines, the second term in the right hand side goes to 0 as $n \to \infty$ and we deduce that $(1.2) = o(1)$. Finally,

$$\begin{array}{rcl} (2) & = & \left\| \int_{x\notin\mathcal{X}_n} h(X,\theta) dP_* \right\| \leq E_{P_*}\left(\|h(X,\theta)\| \mathbb{I}_{\|g(X,\theta)\|\geq m_n} \right) \\ & \leq & \left(E_{P_*}\left(\|h(X,\theta)\|^2\right)\right)^{1/2} \left(P_*\left(\|g(X,\theta)\| \geq m_n\right)\right)^{1/2} \\ & \leq & \left(E_{P_*}\left(\sup_{\theta\in\mathcal{N}} \|h(X,\theta)\|^2\right)\right)^{1/2} \left(\frac{1}{m_n} E_{P_*}\left(\sup_{\theta\in\mathcal{N}} \|g(X,\theta)\|\right) \right)^{1/2} = O(m_n^{-1/2}) = o(1). \end{array}$$

This completes the first conclusion.

$$|E_{Q_n}[\exp(\lambda' g_n(X,\theta))] - 1| \leq \left| \int \exp(\lambda' g_n(X,\theta))(dQ_n - dP_*) \right| + E_{P_*}[|\exp(\lambda' g_n(X,\theta) - 1|].$$

From the preceding lines, it is not hard to see that $\sup_{(\lambda,\theta)\in\Lambda_n\times\Theta} E_{P_*}[|\exp(\lambda' g_n(X,\theta) - 1|] \to 0$. Also,

$$\left| \int \exp(\lambda' g_n(x,\theta)(dQ_n - dP_*) \right| \leq C \int \left(dQ_n^{1/2} - dP_*^{1/2} \right)^2 + 2C \left(\int \left(dQ_n^{1/2} - dP_*^{1/2} \right)^2 \right)^{1/2}$$
$$\leq C \cdot \frac{r^2}{n} + 2C \cdot \frac{r}{\sqrt{n}} \to 0, \text{ as } n \to \infty. \quad \square$$

Lemma C.7. *Let $r > 0$ and Q_n be a sequence contained in $B_H(P_*, r/\sqrt{n})$. If Assumption 3 holds, then we have:*

$$(i) \qquad \sqrt{n}(\bar{T}(P_n) - \theta^*) = -\Sigma G' \Omega^{-1} \sqrt{n} E_{P_n}[g_n(X,\theta^*)] + o_P(1) \quad \text{under } Q_n$$

$$(ii) \quad \sqrt{n}(\bar{T}(P_n) - \bar{T}(Q_n)) \xrightarrow{d} N(0,\Sigma), \quad \text{under } Q_n$$

Proof of Lemma C.7: (i) The proof of Theorem 3.3 leading to (B.12) is also valid with $\hat{\theta}$ replaced by $\bar{T}(P_n)$ and g replaced by g_n and we have:

$$\sqrt{n}(\bar{T}(P_n) - \theta^*) = -\Sigma G' \Omega^{-1} \sqrt{n} E_{P_n}[g_n(X,\theta^*)] + o_P(1),$$

where the $o_P(1)$ term is so with respect to P_*. Using the fact that $Q_n \in B_H(P_*, r/\sqrt{n})$, it is not hard to see that Q_n and P_* are contiguous probability measures in the sense that for any measurable sequence of events A_n, $(P_*(A_n) \to 0) \Leftrightarrow (Q_n(A_n) \to 0)$. Thus the $o_P(1)$ term has the same magnitude under Q_n and this establishes (i).

(ii) Using Lemma C.5 and (i), we can write:

$$\begin{aligned} \sqrt{n}(\bar{T}(P_n) - \bar{T}(Q_n)) &= \sqrt{n}(\bar{T}(P_n) - \theta^*) - \sqrt{n}(\bar{T}(Q_n) - \theta^*) \\ &= -\sqrt{n}\Sigma G' \Omega^{-1} \sqrt{n} \left(E_{P_n}[g_n(X,\theta^*)] - (E_{Q_n}[g_n(X,\theta^*)]) + o_P(1). \right. \end{aligned}$$

Relying on the central limit theorem for triangular arrays as in the proof of KOE's Lemma A.8, we can claim that

$$\sqrt{n} \left(E_{P_n}[g_n(X,\theta^*)] - (E_{Q_n}[g_n(X,\theta^*)]) \right. \xrightarrow{d} N(0,\Omega),$$

under Q_n and (ii) follows as a result. $\square$

Lemma C.8. *Let $r > 0$ and Q_n be a sequence contained in $B_H(P_*, r/\sqrt{n})$. If Assumption 3 holds, then the following statements hold under Q_n:*

(i) $\bar{T}_1(\theta^*, P_n) = O_P(n^{-1/2})$,
(ii) $E_{P_n}\left(g_n(X, \bar{T}_{P_n})\right) = O_P(n^{-1/2})$, $E_{P_n}\left(g_n(X, \bar{T}_{P_n}) g_n(X, \bar{T}_{P_n})'\right) = \Omega + O_P(n^{-1/2})$, *and* $E_{P_n}\left(\frac{\partial g_n}{\partial \theta'}(X, \bar{T}_{P_n})\right) = G + o_P(1)$,
(iii) $\bar{T}_1(\bar{T}_{P_n}, P_n) = O_P(n^{-1/2})$.

Proof of Lemma C.8: (i) Do as in the proof of Lemma C.4(i) with Q_n replaced by P_n. Then, obtain that $\bar{T}_1(\theta^*, P_n) = O_P(n^{-1/2})$ under P_*. Thanks to the mutual contiguity property of Q_n and P_* exposed in the proof of Lemma C.7, we can claim (i).

(ii) and (iii) The first equation in (ii) and (iii) are obtained along the same lines as the proof of Lemma C.4(iii) and C.4(ii), respectively whereas the other two equations in (ii) are obtained by a first order mean value expansion around θ^* and using Lemma C.7(i). $\square$

Appendix D. Global misspecification

Proof of Theorem 5.1: The proof is split into three parts: in (i), we show the convergence of $\hat{\theta}$ and $\hat{\lambda}(\hat{\theta})$; in (ii), we derive the asymptotic distribution of the estimators and discuss the estimation of the (robust) variance-covariance matrix; in (iii), we show that the asymptotic variance in Theorem 5.1 corresponds to the one in Theorem 3.3 under correct specification.

(i) First, we show the consistency of $\hat{\lambda}$ and $\hat{\theta}$. We follow the proof in three steps of Theorem 10 in Schennach (2007): (a) we show that $\hat{\lambda}(\theta) \xrightarrow{P} \lambda^*(\theta)$ uniformly for $\theta \in \Theta$ and that $\lambda^*(\cdot)$ is continuous at θ^*; (b) we show that $\hat{\theta} \xrightarrow{P} \theta^*$; (c) It follows that $\hat{\lambda}(\hat{\theta}) \xrightarrow{P} \lambda^*(\theta^*)$.
(a) Let $\lambda^*(\theta)$ denote the argument of the minimum over Λ of $\lambda \mapsto E[\exp(\lambda' g(X,\theta))]$ which is unique by strict convexity of $E[\exp(\lambda' g(X,\theta))]$ over the convex set Λ. The Berge's maximum theorem guarantees that $\lambda^*(\cdot)$ is continuous. Since $\exp(\lambda' g(x,\theta))$ is continuous in λ and θ, thanks to Assumption 5(v), we have:

$$\hat{M}_\theta(\lambda) \equiv \frac{1}{n}\sum_{i=1}^{n} \exp(\lambda' g(x_i,\theta) \xrightarrow{P} M_\theta(\lambda) \equiv E(\exp(\lambda' g(X,\theta))),$$

uniformly over the compact set $\Lambda \times \Theta$.
Recall $\hat{\lambda}(\theta) \equiv \arg\min_{\lambda\in\Lambda} \hat{M}_\theta(\lambda)$. We now show that for any $\eta > 0$,

$$P\left(\sup_{\theta\in\Theta} \|\hat{\lambda}(\theta) - \lambda^*(\theta)\| \leq \eta\right) \to 1, \quad as \quad n \to \infty.$$

For a given $\eta > 0$, define ϵ as follows:

$$\epsilon = \inf_{\theta\in\Theta} \inf_{\lambda\in\Lambda:\|\lambda-\lambda^*(\theta)\|\geq\eta} (M_\theta(\lambda) - M_\theta(\lambda^*(\theta)))$$

By strict convexity of $M_\theta(\lambda)$ in λ and compactness of Θ, we have $\epsilon > 0$. In addition, by definition of ϵ,

$$\text{if} \quad \sup_{\theta\in\Theta}(M_\theta(\hat{\lambda}(\theta)) - M_\theta(\lambda^*(\theta))) \leq \epsilon \quad \text{then} \quad \sup_{\theta\in\Theta} \|\hat{\lambda}(\theta) - \lambda^*(\theta)\| \leq \eta.$$

Since $\hat{M}_\theta(\hat{\lambda}(\theta)) - \hat{M}_\theta(\lambda^*(\theta)) < 0$, we have:

$$\begin{aligned}
\sup_{\theta\in\Theta}(M_\theta(\hat{\lambda}(\theta)) - M_\theta(\lambda^*(\theta))) &\leq \sup_{\theta\in\Theta}(M_\theta(\hat{\lambda}(\theta)) - \hat{M}_\theta(\hat{\lambda}(\theta))) + \sup_{\theta\in\Theta}(\hat{M}_\theta(\hat{\lambda}(\theta)) - \hat{M}_\theta(\lambda^*(\theta))) \\
&\quad + \sup_{\theta\in\Theta}(\hat{M}_\theta(\lambda^*(\theta)) - M_\theta(\lambda^*(\theta))) \\
&\leq \sup_{\theta\in\Theta}|M_\theta(\hat{\lambda}(\theta)) - \hat{M}_\theta(\hat{\lambda}(\theta))| + \sup_{\theta\in\Theta}|\hat{M}_\theta(\lambda^*(\theta)) - M_\theta(\lambda^*(\theta))| \\
&\leq \epsilon/2 + \epsilon/2
\end{aligned}$$

Hence, we conclude that

$$\sup_{\theta\in\Theta} \|\hat{\lambda}(\theta) - \lambda^*(\theta)\| \leq \eta,$$

with probability approaching one.

(b) To prove the consistency of $\hat{\theta}$, we will make use of the consistency of $\hat{\lambda}$. We have a uniform convergence of the objective function $\boldsymbol{\Delta}_{P_n}(\lambda(\theta),\theta)$ over (Λ,Θ) which implies that:

$$\begin{aligned}
&\forall\epsilon > 0 \ \lim_n P\left(|\boldsymbol{\Delta}_{P_n}(\lambda(\hat{\theta}),\hat{\theta}) - \Delta(\lambda(\hat{\theta}),\hat{\theta})| < \epsilon/3\right) = 1 \\
\Rightarrow \quad &\forall\epsilon > 0 \ \lim_n P\left(\boldsymbol{\Delta}_{P_n}(\lambda(\hat{\theta}),\hat{\theta}) < \Delta(\lambda(\hat{\theta}),\hat{\theta}) + \epsilon/3\right) = 1 \qquad \text{(D.1)}
\end{aligned}$$

Similarly, we can show that

$$\forall\epsilon > 0 \ \lim_n P\left(\Delta(\lambda(\theta^*),\theta^*) < \boldsymbol{\Delta}_{P_n}(\lambda(\theta^*),\theta^*) + \epsilon/3\right) = 1 \qquad \text{(D.2)}$$

By definition of $\hat{\theta}$, we have:

$$\forall\epsilon > 0 \ \lim_n P\left(\boldsymbol{\Delta}_{P_n}(\lambda(\theta^*),\theta^*) < \boldsymbol{\Delta}_{P_n}(\lambda(\hat{\theta}),\hat{\theta}) + \epsilon/3\right) = 1 \qquad \text{(D.3)}$$

From equations (D.1) and (D.3), we get:

$$\forall\epsilon > 0 \ \lim_n P\left(\boldsymbol{\Delta}_{P_n}(\lambda(\theta^*),\theta^*) < \Delta(\lambda(\hat{\theta}),\hat{\theta}) + 2\epsilon/3\right) = 1 \qquad \text{(D.4)}$$

We can now use equation (D.2) to deduce:

$$\forall\epsilon > 0 \ \lim_n P\left(\Delta(\lambda(\theta^*),\theta^*) < \Delta(\lambda(\hat{\theta}),\hat{\theta}) + \epsilon\right) = 1 \qquad \text{(D.5)}$$

We now use the identification assumption and the definition of $\hat{\theta}$ to deduce that, for every neighborhood $\mathcal{N}^*$ of θ^*, there exists a constant $\eta > 0$ such that

$$\sup_{\theta \in \Theta \setminus \mathcal{N}^*} \Delta(\lambda(\theta), \theta) + \eta < \Delta(\lambda(\theta^*), \theta^*) .$$

Then, we have

$$\hat{\theta} \in \Theta \setminus \mathcal{N}^* \Rightarrow \Delta(\lambda(\hat{\theta}), \hat{\theta}) + \eta \leq \sup_{\theta \in \Theta \setminus \mathcal{N}^*} \Delta(\lambda(\theta), \theta) + \eta < \Delta(\lambda(\theta^*), \theta^*).$$

Thus,

$$P\left(\hat{\theta} \in \Theta \setminus \mathcal{N}^*\right) \leq P\left(\Delta(\lambda(\hat{\theta}), \hat{\theta}) + \eta \leq \Delta(\lambda(\theta^*), \theta^*)\right) \to 0 \quad as \quad n \to \infty,$$

where the convergence to 0 follows directly from equation (D.5) above.

(ii) To derive the asymptotic distribution of ETHD estimator under global misspecification, we write a mean-value expansion of the first-order condition around (θ^*, λ^*). Recall that (θ^*, λ^*) is assumed to be in the interior of the parameter space (see Assumption 5). Hence, $\hat{\theta}$ solves the first order condition:

$$\frac{d\boldsymbol{\Delta}_{P_n}(\hat{\lambda}(\theta), \theta)}{d\theta} = \frac{N_1(\hat{\lambda}(\theta), \theta)}{D_1(\hat{\lambda}(\theta), \theta)} - \frac{N_2(\hat{\lambda}(\theta), \theta)}{D_2(\hat{\lambda}(\theta), \theta)} = 0, \tag{D.6}$$

$$\text{with} \quad N_1(\lambda, \theta) = \frac{1}{2} E_{P_n}\left[\left(\frac{d\hat{\lambda}'}{d\theta}(\theta) g(X, \theta) + \frac{\partial g(X, \theta)'}{\partial \theta}\lambda\right) \exp\left(\lambda' g(X, \theta)/2\right)\right],$$

$$N_2(\lambda, \theta) = \frac{1}{2} E_{P_n}\left[\left(\frac{d\hat{\lambda}'}{d\theta}(\theta) g(X, \theta) + \frac{\partial g(X, \theta)'}{\partial \theta}\lambda\right) \exp\left(\lambda' g(X, \theta)\right)\right] \times D_0\left(\frac{\lambda}{2}, \theta\right),$$

$$D_1(\lambda, \theta) = D_0(\lambda, \theta)^{1/2}, \quad D_2(\lambda, \theta) = D_0(\lambda, \theta)^{3/2}, \quad D_0(\lambda, \theta) = E_{P_n}\left[\exp(\lambda' g(X, \theta))\right].$$

Similarly, $\hat{\lambda}(\hat{\theta})$ solves the first-order condition:

$$E_{P_n}\left[g(X, \hat{\theta}) \exp\left(\lambda' g(X, \hat{\theta})\right)\right] = 0. \tag{D.7}$$

We consider the left hand sides of (D.6) and (D.7) and carry out their mean-value expansions around (λ^*, θ^*).

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} N_1(\lambda^*, \theta^*)/D_1(\lambda^*, \theta^*) - N_2(\lambda^*, \theta^*)/D_2(\lambda^*, \theta^*) \\ E_{P_n}\left[g(X, \theta^*) \exp(\lambda^{*\prime} g(X, \theta^*))\right] \end{pmatrix} + \overline{R}_n \begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\lambda} - \lambda^* \end{pmatrix} \tag{D.8}$$

where, with $\overline{\theta} \in (\theta^*, \hat{\theta})$ and $\overline{\lambda} \in (\lambda^*, \hat{\lambda})$ and both may vary from row to row,

$$\overline{R}_n = \begin{pmatrix} R_{\theta,\theta}(\overline{\lambda}, \overline{\theta}) & R_{\theta,\lambda}(\overline{\lambda}, \overline{\theta}) \\ R_{\lambda,\theta}(\overline{\lambda}, \overline{\theta}) & R_{\lambda,\lambda}(\overline{\lambda}, \overline{\theta}) \end{pmatrix},$$

$$\text{with} \quad R_{\theta,\theta}(\lambda, \theta) = \frac{\partial}{\partial \theta'}\left(\frac{N_1(\lambda, \theta)}{D_1(\lambda, \theta)} - \frac{N_2(\lambda, \theta)}{D_2(\lambda, \theta)}\right) \tag{D.9}$$

$$R_{\theta,\lambda}(\lambda, \theta) = \frac{\partial}{\partial \lambda'}\left(\frac{N_1(\lambda, \theta)}{D_1(\lambda, \theta)} - \frac{N_2(\lambda, \theta)}{D_2(\lambda, \theta)}\right) \tag{D.10}$$

$$R_{\lambda,\theta}(\lambda, \theta) = E_{P_n}\left[\left(\frac{\partial g'(X, \theta)}{\partial \theta} + \frac{\partial g(X, \theta)'}{\partial \theta}\lambda g(X, \theta)'\right)' \exp(\lambda' g(X, \theta))\right] \tag{D.11}$$

$$R_{\lambda,\lambda}(\lambda, \theta) = E_{P_n}\left[g(X, \theta) g(X, \theta)' \exp(\lambda' g(X, \theta))\right] \tag{D.12}$$

where D_i, N_i were defined above, and their derivatives are computed as follows:

$$\begin{aligned}
\tfrac{\partial N_1(\lambda,\theta)}{\partial\theta'} &= \tfrac{1}{2}E_{P_n}\Big[\Big(\textstyle\sum_{k=1}^m \frac{d^2\hat\lambda_k(\theta)}{d\theta d\theta'}g_k(X,\theta)+\sum_{k=1}^m\frac{\partial^2 g_k(X,\theta)}{\partial\theta\partial\theta'}\lambda_k+\frac{d\hat\lambda(\theta)'}{d\theta}\frac{\partial g(X,\theta)}{\partial\theta'}\\
&\quad +\tfrac{1}{2}\Big(\tfrac{d\hat\lambda(\theta)'}{d\theta}g(X,\theta)+\tfrac{\partial g(X,\theta)'}{\partial\theta}\lambda\Big)\lambda'\tfrac{\partial g(X,\theta)}{\partial\theta'}\Big)\exp\left(\lambda' g(X,\theta)/2\right)\Big]\\
\tfrac{\partial N_2(\lambda,\theta)}{\partial\theta'} &= \tfrac{1}{2}E_{P_n}\Big[\Big(\textstyle\sum_{k=1}^m \frac{d^2\hat\lambda_k(\theta)}{d\theta d\theta'}g_k(X,\theta)+\sum_{k=1}^m\frac{\partial^2 g_k(X,\theta)}{\partial\theta\partial\theta'}\lambda_k+\frac{d\hat\lambda(\theta)'}{d\theta}\frac{\partial g(X,\theta)}{\partial\theta'}\\
&\quad +\Big(\tfrac{d\hat\lambda(\theta)'}{d\theta}g(X,\theta)+\tfrac{\partial g(X,\theta)'}{\partial\theta}\lambda\Big)\lambda'\tfrac{\partial g(X,\theta)}{\partial\theta'}\Big)\exp\left(\lambda' g(X,\theta)\right)\Big]\times D_0\left(\tfrac{\lambda}{2},\theta\right)\\
&\quad +\tfrac{1}{4}E_{P_n}\Big[\Big(\tfrac{d\hat\lambda(\theta)'}{d\theta}g(X,\theta)+\tfrac{\partial g(X,\theta)'}{\partial\theta}\lambda\Big)\exp\left(\lambda g(X,\theta)\right)\Big]\times E_{P_n}\Big[\lambda'\tfrac{\partial g(X,\theta)}{\partial\theta'}\exp\left(\lambda' g(X,\theta)/2\right)\Big]\\
\tfrac{\partial D_1(\lambda,\theta)}{\partial\theta'} &= \tfrac{1}{2}E_{P_n}\Big[\lambda'\tfrac{\partial g(X,\theta)}{\partial\theta'}\exp\left(\lambda' g(X,\theta)\right)\Big]\times D_0(\lambda,\theta)^{-1/2}\\
\tfrac{\partial D_2(\lambda,\theta)}{\partial\theta'} &= \tfrac{3}{2}E_{P_n}\Big[\lambda'\tfrac{\partial g(X,\theta)}{\partial\theta'}\exp\left(\lambda' g(X,\theta)\right)\Big]\times D_0(\lambda,\theta)^{1/2}.
\end{aligned}$$

Also,

$$\begin{aligned}
\tfrac{\partial N_1(\lambda,\theta)}{\partial\lambda'} &= \tfrac{1}{2}E_{P_n}\Big[\Big(\tfrac{\partial g(X,\theta)'}{\partial\theta}+\tfrac{1}{2}\Big(\tfrac{d\hat\lambda(\theta)'}{d\theta}g(X,\theta)+\tfrac{\partial g(X,\theta)'}{\partial\theta}\lambda\Big)g(X,\theta)'\Big)\exp\left(\lambda' g(X,\theta)/2\right)\Big]\\
\tfrac{\partial N_2(\lambda,\theta)}{\partial\lambda'} &= \tfrac{1}{2}E_{P_n}\Big[\Big(\tfrac{\partial g(X,\theta)'}{\partial\theta}+\Big(\tfrac{d\hat\lambda(\theta)'}{d\theta}g(X,\theta)+\tfrac{\partial g(X,\theta)'}{\partial\theta}\lambda\Big)g(X,\theta)'\Big)\exp\left(\lambda' g(X,\theta)\right)\Big]\times D_0\left(\tfrac{\lambda}{2},\theta\right)\\
&\quad +\tfrac{1}{4}E_{P_n}\Big[\Big(\tfrac{d\hat\lambda(\theta)'}{d\theta}g(X,\theta)+\tfrac{\partial g(X,\theta)'}{\partial\theta}\lambda\Big)\exp\left(\lambda' g(X,\theta)\right)\Big]\times E_{P_n}\left[g(X,\theta)'\exp\left(\lambda' g(X,\theta)/2\right)\right]\\
\tfrac{\partial D_1(\lambda,\theta)}{\partial\lambda'} &= \tfrac{1}{2}E_{P_n}\left[g(X,\theta)'\exp\left(\lambda' g(X,\theta)\right)\right]\times D_0(\lambda,\theta)^{-1/2},\\
\tfrac{\partial D_2(\lambda,\theta)}{\partial\lambda'} &= \tfrac{3}{2}E_{P_n}\left[g(X,\theta)'\exp\left(\lambda' g(X,\theta)\right)\right]\times D_0(\lambda,\theta)^{1/2}.
\end{aligned}$$

Let $\text{plim}\overline{R}_n = R$ assumed to be nonsingular; we then get:

$$\begin{aligned}
R\sqrt{n}\begin{pmatrix}\hat\theta-\theta^*\\ \hat\lambda-\lambda^*\end{pmatrix} &= -\sqrt{n}\begin{pmatrix} N_1(\lambda^*,\theta^*)/D_1(\lambda^*,\theta^*)-N_2(\lambda^*,\theta^*)/D_2(\lambda^*,\theta^*)\\ E_{P_n}\left[g(X,\theta^*)\exp(\lambda^{*\prime}g(X,\theta^*))\right]\end{pmatrix}+o_p(1) \qquad \text{(D.13)}\\
&\equiv \sqrt{n}A_n^*+o_p(1)
\end{aligned}$$

with

$$A_n^* = -\begin{pmatrix}A_{n,1}^*\\ A_{n,2}^*\end{pmatrix} = \begin{pmatrix} N_1(\lambda^*,\theta^*)/D_1(\lambda^*,\theta^*)-N_2(\lambda^*,\theta^*)/D_2(\lambda^*,\theta^*)\\ \sum_{i=1}^n[g(X_i,\theta^*)\exp(\lambda^{*'}g(X_i,\theta^*))]/n\end{pmatrix}$$

that is,

$$A_{n,1}^* = E_n^{-1/2}\left[\frac{1}{2n}\sum_{i=1}^n\left(\frac{d\hat\lambda(\theta^*)'}{d\theta}g(X_i,\theta^*)+\frac{\partial g(X_i,\theta^*)'}{\partial\theta}\lambda^*\right)\left(\exp(\lambda^{*'}g(X_i,\theta^*)/2)-\exp(\lambda^{*'}g(X_i,\theta^*))\times\frac{F_n}{E_n}\right)\right]$$

where

$$\begin{aligned}
E_n &= \frac{1}{n}\sum_{i=1}^n\exp(\lambda^{*'}g(X_i,\theta^*)), \quad F_n=\frac{1}{n}\sum_{i=1}^n\exp(\lambda^{*'}g(X_i,\theta^*)/2)\\
\frac{d\hat\lambda(\theta^*)}{d\theta'} &= -\left[\frac{1}{n}\sum_{i=1}^n g(X_i,\theta^*)g(X_i,\theta^*)'\right]^{-1}\frac{1}{n}\sum_{i=1}^n\left(\frac{\partial g(X_i,\theta^*)}{\partial\theta'}+g(X_i,\theta^*)\lambda^{*\prime}\frac{\partial g(X_i,\theta^*)}{\partial\theta'}\right)\exp(\lambda^{*'}g(X_i,\theta^*)).
\end{aligned}$$

Let us define K_i as follows,

$$K_i = \begin{pmatrix} g(X_i,\theta^*)\exp(\lambda^{*'}g(X_i,\theta^*)/2) \\ g(X_i,\theta^*)\exp(\lambda^{*'}g(X_i,\theta^*)) \\ \exp(\lambda^{*'}g(X_i,\theta^*)/2) \\ \exp(\lambda^{*'}g(X_i,\theta^*)) \\ g(X_i,\theta^*)g(X_i,\theta^*)' \\ \left(\frac{\partial g(X_i,\theta^*)}{\partial\theta'} + g(X_i,\theta^*)\lambda^{*\prime}\frac{\partial g(X_i,\theta^*)}{\partial\theta'}\right)\exp(\lambda^{*'}g(X_i,\theta^*)) \end{pmatrix}$$

From Assumption 5, a joint CLT holds for K_i such that,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} K_i - E(K_i)\right) \xrightarrow{n} \mathcal{N}(0,W)$$

We now define $\Omega^* = \texttt{AVar}\,(A_n^*)$ and its explicit expression can be obtained from the previous CLT combined with the Delta-method. Finally, we have:

$$\sqrt{n}\begin{pmatrix} \hat\theta - \theta^* \\ \hat\lambda - \lambda^* \end{pmatrix} \xrightarrow{d} N(0, R^{-1}\Omega^* R^{-1}) \qquad \text{with} \qquad R = \text{plim}\overline{R}_n\,.$$

The expected result directly follows.

Under our maintained i.i.d. assumption, the estimation of the above asymptotic variance-covariance matrix is straightforward: all quantities are replaced by their sample counterparts, and the pseudo-true values (λ^*,θ^*) by their estimators.

(iii) Finally, we show that under correct specification, the expansion (D.13) coincides with (B.11)), that is:

$$\begin{pmatrix} G' & 0 \\ \Omega & G \end{pmatrix}\sqrt{n}\begin{pmatrix} \hat\theta - \theta^* \\ \hat\lambda \end{pmatrix} = \begin{pmatrix} 0 \\ -\sqrt{n}E_{P_n}\left(g(X,\theta^*)\right) \end{pmatrix} + o_p(1).$$

After replacing λ^* by 0, we easily get that

$$\frac{N_1(\lambda^*,\theta^*)}{D_1(\lambda^*,\theta^*)} - \frac{N_2(\lambda^*,\theta^*)}{D_2(\lambda^*,\theta^*)} = 0$$

It remains to show that

$$\text{plim}\overline{R}_n = \begin{pmatrix} 0 & G \\ G' & \Omega \end{pmatrix}$$

After replacing λ^* by 0, we easily get that

$$\begin{aligned}
R_{\theta,\theta}(\lambda^*,\theta^*) &= \frac{\partial N_1(\lambda^*,\theta^*)}{\partial\theta} - \frac{\partial N_2(\lambda^*,\theta^*)}{\partial\theta} \\
&\quad \text{since } D_1(\lambda^*,\theta^*) = D_2(\lambda^*,\theta^*) = 1 \text{ and } \partial D_1(\lambda^*,\theta^*)/\partial\theta = \partial D_2(\lambda^*,\theta^*)/\partial\theta = 0 \\
&= 0 \\
R_{\theta,\lambda}(\lambda^*,\theta^*) &= \frac{\partial N_1(\lambda^*,\theta^*)}{\partial\lambda} - \frac{\partial N_2(\lambda^*,\theta^*)}{\partial\lambda} \\
&\quad \text{since } D_1(\lambda^*,\theta^*) = D_2(\lambda^*,\theta^*) = 1 \text{ and } \partial D_1(\lambda^*,\theta^*)/\partial\lambda = \partial D_2(\lambda^*,\theta^*)/\partial\lambda = 0 \\
&= E_{P_n}\left[\frac{\partial g(X,\theta^*)}{\partial\theta'}\right] \xrightarrow{P} E\left(\frac{\partial g(X,\theta^*)}{\partial\theta'}\right) = G \\
&\quad \text{after using expressions derived in the proof of Theorem 3.3} \\
R_{\lambda,\theta}(\lambda^*,\theta^*) &= E_{P_n}\left[\frac{\partial g'(X,\theta^*)}{\partial\theta}\right] \xrightarrow{P} E\left(\frac{\partial g(X,\theta^*)'}{\partial\theta}\right) = G' \\
R_{\lambda,\lambda}(\lambda^*,\theta^*) &= E_{P_n}\left[g(X,\theta^*)g'(X,\theta^*)\right] \xrightarrow{P} E\left(g(X,\theta^*)g(X,\theta^*)'\right) = \Omega
\end{aligned}$$

and the expected result follows readily □