Identifiable Uniqueness Conditions for a Large Class of Extremum Estimators

Francisco Blasques*

November 29, 2010

Abstract

Proofs of consistency of extremum estimators usually require assumptions ensuring that there exists a unique well separated (*identifiably unique*) minimizer of the limit criterion function. Unfortunately, these assumptions are sometimes opaque and do not lend themselves to immediate verification. This paper discusses ways of confirming that *identifiable uniqueness* holds for the class of extremum estimators whose limiting criterion function can be appropriately defined as a divergence on a space of probability measures (minimum distance estimators being a special case). In particular, the task of verifying that *identifiable uniqueness* holds is reduced to that of verifying the *strong unicity* of best approximations on an appropriate space of probability measures or regression functions. Sufficient conditions for *strong unicity* of best approximations are often easy to verify.

*The author is thankful to Marco Avarucci, Eric Beutner, Bertrand Candelon, Franz Palm and Jean-Pierre Urbain for the many suggestions from which this paper has benefited considerably. I'm also grateful to Raymond Kan, Michael Jansson, Mika Meitz and other participants of the SETA 2010 conference, in Singapore, for useful comments and suggestions. The usual disclaimer applies. Corresponding Address: Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Email: f.blasques@maastrichtuniversity.nl.

1 Introduction

Building on early work of Doob (1934, 1953), Wald (1949), Le Cam (1953), Cramer (1946) and others that addressed the consistency of maximum likelihood (ML) estimators with independently identically distributed (iid) data, the "classic" consistency proof of extremum estimators originated in the well known contributions of Jennrich (1969) and Malinvaud (1970). These two papers independently addressed the consistency of the least squares estimator in a nonlinear regression framework. They also seem to be at the origin of much of the research on the asymptotic properties of extremum estimators that took place during the following decades. Numerous contributions have since then allowed for the properties of extremum estimators to be well understood in multivariate dynamic settings, misspecified models and under heterogeneity and dependence of the data. The list is extensive. See e.g. Burguete et al. (1982), Amemiya (1983) and Gallant and White (1988) for early reviews of important contributions, as well as Pötscher and Prucha (1991a,b, 1997) for a more recent and complete account of the relevant literature.

Despite the diversity, there is an underlying basic structure of conditions and methodologies that are common to the great majority of consistency results in this literature. In particular, the *uniform convergence* of criterion functions and the *identifiable uniqueness* of the argument that minimizes the limit criterion function seem to have pervasive influence, being present under many guises in most consistency proofs. Here we shall be concerned with the latter of these two conditions, the identifiable uniqueness, which requires fundamentally that the extremum estimator's limit criterion function have a well separated minimum (see e.g. White (1980a) and Domowitz and White (1982)).

Unfortunately, identifiable uniqueness conditions are sometimes opaque, in the sense that they do not seem to lend themselves to immediate verification. The suspicion of failure therefore remains; see e.g. Pötscher and Prucha (1991a, ch.4) for a review of problematic non-trivial cases where identifiable uniqueness fails to hold.

The aim of this paper is to lay down a simple yet general methodology that allows the researcher to verify if the identifiable uniqueness assumption holds true in the context of possibly misspecified models. To follow the tradition of the "classic" results mentioned above we shall also adopt the nonlinear regression framework. For clarity, we consider here a simple prototypical nonlinear regression case that abstracts from the tedious considerations required by a more general result. It will become however clear that extensions to more general cases are often straightforward to achieve. Some trivial extensions to "non-regression" problems are mentioned here. An extension to complex dynamic models is addressed in Blasques (2010).

It should also be made clear from the outset that there is not necessarily a strict relation between imposing an identifiable uniqueness condition and ensuring that the model at hand satisfies the well known *identification condition* (even though this often the case).[1] We do not address the identification condition here, although we discuss the role it plays in the present problem. An important practical implication of this distinction is that the present theory is mostly uninteresting for those special cases (typically involving well-specified models, compact parameter spaces and continuous criterion functions) where model identification implies that identifiable uniqueness holds on the estimator's criterion function.

Finally, it is also important to stress that we will be concerned with providing only sufficient conditions for identifiable uniqueness. Necessity is not addressed here. As such, the conditions under which the methodology remains of practical interest should be as general as possible. Indeed, it is not hard to devise restrictive conditions that once verified, imply immediately identifiable uniqueness (think e.g. of strict convexity of a continuous limit criterion function on a compact domain). Such conditions are however of very limited applicability and become, in that sense, uninteresting. The challenge is thus to achieve generality while at the same time ensuring simple verification.

[1] The researcher can always construct an extremum estimator (albeit possibly an uninteresting one) that satisfies an identifiable uniqueness condition despite having a model at hand that does not satisfy the fundamental identification condition, and vice-versa.

The lack of a general enough methodology allowing researchers to verify if identifiable uniqueness assumptions hold has lead some authors to discuss the adequacy of this assumption in the context of misspecified models and to propose consistency results that do not rely on it; see e.g. Pötscher and Prucha (1991a, section 4.6) and references therein. We shall not follow this trail here.[2] We choose to follow instead the literature aimed at the verification of uniqueness conditions. Some examples include: *(i)* Freedman and Diaconis (1982) analyze inconsistency of redescending M-estimators for location parameters of symmetric distributions using iid data that is caused by failure of the uniqueness assumption; *(ii)* Kabaila (1983) addresses the failure of the uniqueness assumption for estimators of the parameter vector minimizing the one-step-ahead prediction errors in misspecified ARMA models; *(iii)* Clarke (1983) provides verifiable conditions for the uniqueness of ψ-type M-estimators using iid data that rely on somewhat restrictive conditions involving the Frechet differentiability of functional solutions; *(iv)* Rivest (1989) constitutes a failed attempt to prove uniqueness of robust extremum estimators, see Crisp and Burridge (1993); *(v)* Ducharme (1995) shows that the L_1-norm minimizer extremum estimator is generally unique in the context of well specified multivariate response nonlinear-regression models; *(vi)* Donoho and Liu (1988) observe pathologies of minimum-distance estimators related to the failure of uniqueness conditions (these pathologies can be well understood under the general methodology proposed here); and finally; *(vii)* Kent and Tyler (2001) provide conditions for local uniqueness of constrained and redescending M-estimators in the context of well-specified models, by imposing conditions for the estimator's criterion function to be locally well behaved.

In what follows we generalize some of the results just mentioned in that we provide conditions for identifiable uniqueness to hold globally and for a large class of extremum estimators in the context of possibly dependent heterogeneous data and misspecified nonlinear regres-

[2]It often seems desirable to retain the identifiable uniqueness assumption as it provides the researcher with a host of useful properties, e.g. continuous mapping theorems for arg max functionals. Furthermore, this condition seems to play an important role in guaranteeing the economic interpretation of empirical work.

sion models. In particular, we note that typical identifiable uniqueness assumptions can be restated in terms of transparent verifiable conditions on the nature of both the estimator and the model at hand. The idea is to adapt the statistical problem to be amenable to the use of results stemming from the field of Approximation Theory. These results are applicable to the class of extremum estimators whose limiting criterion function can be defined as a divergence on a space probability measures underlying the data. This class includes as a subset the usual minimum distance estimators. The problem of divergence minimization can also be translated to the space of regression functions. Building on Approximation Theory's results, the task of verifying the identifiable uniqueness of the limit minimizer is then reduced to that of verifying the strong uniqueness of best approximations in the space of probability measures or regression functions. Sufficient conditions for strong unicity are often easy to verify, thus giving the researcher the opportunity to check if identifiable uniqueness holds in various applications.

The rest of the paper is structured as follows. Section 2 contains mainly preliminary considerations and lays down the foundations for the remaining sections both in terms of definitions and notation. Section 3 describes briefly the typical framework under which consistency of extremum estimators is obtained. Section 4 restates the estimation exercise in a more useful way by rewriting the limiting estimation problem as that of divergence minimization on the space of probability measures or regression functions. Section 5 introduces some concepts from Approximation Theory and reviews relevant results in this field highlighting the conditions under which approximation problems have (strongly) unique solutions. Section 6 derives identifiable uniqueness from this new set of conditions and provides some consistency results that follow immediately as corollaries. Section 7 illustrates the verification step with a few simple examples of nonlinear regression models and alternative extremum estimators. Section 8 offers some final remarks and concludes.

Finally, a word on notation. In what follows, $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ denote the sets of natural, integer and real numbers. If $\mathcal{A}$ is a set, $\mathfrak{B}(\mathcal{A})$ denotes the Borel σ-algebra over $\mathcal{A}$, and $\times_{t=1}^{t=T}\mathcal{A}$, often denoted $\mathcal{A}_T$, is the Cartesian product of T copies of $\mathcal{A}$. Furthermore, in linear

spaces, boldfaced letters (e.g. $\mathbf{a} \in \mathcal{A}$) denote vectors. Note also that $:=$ denotes *definitional equivalence*, whereas $\equiv$ is used to denote *practical equivalence*. If f and g are maps, then $f \circ g := f(g)$ denotes their composition. The mappings $d_{\mathcal{A}}$ and $d^*_{\mathcal{A}}$ denote a divergence and metric defined on the set $\mathcal{A} \times \mathcal{A}$ respectively, and $\| \cdot \|_{\mathcal{A}}$ denotes a norm on $\mathcal{A}$. Finally, p.m. and a.s. stand for *probability measure* and *almost surely*, respectively.

2 Preliminary Considerations

This section is sometimes dense and the casual reader might prefer to use it exclusively as a reference for notation and definitions, thus proceeding directly to Section 3. Consider the T-period sequence $\{\mathbf{x}_t(\omega)\}_{t=1}^T$, a subset of the realized path of an n_x-variate stochastic sequence $\mathbf{x}(\omega) := \{\mathbf{x}_t(\omega), t \in \mathbb{Z}\}$, for some $\omega \in \Omega$ the event space. Let $\mathbf{x}_t(\omega) \in \mathcal{X} \subseteq \mathbb{R}^{n_x}\ \forall\, (\omega, t) \in \Omega \times \mathbb{Z}$.[3] The random sequence $\mathbf{x}$ is thus an $\mathcal{F}/\mathfrak{B}(\mathcal{X}_\infty)$-measurable mapping $\mathbf{x} : \Omega \to \mathcal{X}_\infty \subseteq \mathbb{R}^{n_x}_\infty$ where $\mathbb{R}^{n_x}_\infty := \times_{t=-\infty}^{t=\infty} \mathbb{R}^{n_x}$ denotes the Cartesian product of infinite copies of $\mathbb{R}^{n_x}$ and $\mathcal{X}_\infty = \times_{t=-\infty}^{t=\infty} \mathcal{X}$ with $\mathfrak{B}(\mathcal{X}_\infty) \equiv \mathfrak{B}(\mathbb{R}^{n_x}_\infty) \cap \mathcal{X}_\infty$ (Billingsley (1995, p.159)) where $\mathfrak{B}(\mathbb{R}^{n_x}_\infty)$ denotes the Borel σ-algebra generated by the finite dimensional product cylinders of $\mathbb{R}^{n_x}_\infty$, $\mathcal{F}$ denotes a σ-field defined on the event space Ω, and together with the p.m. P_0 on $\mathcal{F}$, the triplet $(\Omega, \mathcal{F}, \mathrm{P}_0)$ denotes the complete probability space of interest. For every $\omega \in \Omega$, the stochastic sequence $\mathbf{x}(\omega)$ thus lives on the space $(\mathcal{X}_\infty, \mathfrak{B}(\mathcal{X}_\infty), \mathrm{D}_0^{\mathbf{x}})$ where the p.m. $\mathrm{D}_0^{\mathbf{x}}$ is defined over the elements of $\mathfrak{B}(\mathcal{X}_\infty)$. Following White (1980b) and Domowitz and White (1982), consider now the univariate stochastic sequence,

$$y := \{y_t = h_0(\mathbf{x}_t), t \in \mathbb{Z}\}$$

with $h_0 : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$ an $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$-measurable mapping, $\mathfrak{B}(\mathcal{X}) \equiv \mathcal{X} \cap \mathfrak{B}(\mathbb{R}^{n_x})$ and $\mathfrak{B}(\mathcal{Y}) \equiv \mathcal{Y} \cap \mathfrak{B}(\mathbb{R})$. Blasques (2010) extends the results in this paper to complex high dimensional nonlinear dynamic models with unobserved variables and possibly intractable like-

[3] Properties of the data in terms of dynamics and heterogeneity are addressed in Section 3.

lihood functions.[4] Hence, for every $t \in \mathbb{Z}$, $h_0 \circ \mathbf{x}_t : \Omega \to \mathcal{Y}$ is $\mathcal{F}/\mathfrak{B}(\mathcal{Y})$-measurable. For every $\omega \in \Omega$, the sequence $y(\omega)$ thus lives in the space $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty), \mathrm{D}_0^y)$ where D_0^y is the p.m. induced by h_0 on $\mathfrak{B}(\mathcal{Y}_\infty)$ according to $\mathrm{D}_0^y(B_y) = \mathrm{D}_0^{\mathbf{x}} \circ h_0^{-1}(B_y) \, \forall B_y \in \mathfrak{B}(\mathcal{Y}_\infty)$. Define now the joint process $\mathbf{w} := \{\mathbf{w}_t = (y_t, \mathbf{x}_t), t \in \mathbb{Z}\}$. For every $\omega \in \Omega$, $\mathbf{w}_t(\omega) \in \mathcal{W} \equiv \mathcal{Y} \times \mathcal{X}$ and $\mathbf{w}(\omega) \in \mathcal{W}_\infty \equiv \mathcal{Y}_\infty \times \mathcal{X}_\infty \subseteq \mathbb{R}_\infty^{1+n_x} \equiv \times_{t=-\infty}^{t=\infty} \mathbb{R}^{1+n_x}$. The sequence thus lives in $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), \mathrm{D}_0^{\mathbf{w}})$ where $\mathrm{D}_0^{\mathbf{w}}$ denotes the measure defined on $\mathfrak{B}(\mathcal{W}_\infty) \equiv \mathcal{W}_\infty \cap \mathfrak{B}(\mathbb{R}_\infty^{1+n_x})$.[5] Finally, suppose that for some $\omega \in \Omega$ the T-period sequence $\mathbf{w}_\mathbf{T}(\omega) := (\mathbf{y}_\mathbf{T}(\omega), \mathbf{x}_\mathbf{T}(\omega))$ is observed, where $y_\mathbf{T}(\omega) := \{y_t(\omega)\}_{t=1}^{t=T}$ and $\mathbf{x}_\mathbf{T}(\omega) := \{\mathbf{x}_t(\omega)\}_{t=1}^{t=T}$. Yet, h_0 is unknown.

A postulated parametric regression model takes the form $\hat{y}_t = h(\mathbf{x}_t; \boldsymbol{\theta})$ so that the modeled counterpart of the stochastic sequence y is given by,

$$\hat{y} := \{\hat{y}_t = h(\mathbf{x}_t; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta, t \in \mathbb{Z}\}$$

where $h : \mathcal{X} \times \Theta \to \mathcal{Y}$. Here we deviate slightly from standard notation. The use of the hat over y does not imply that fitted values are obtained at a specific point of Θ (usually some $\hat{\boldsymbol{\theta}}_T(\omega)$, $\omega \in \Omega$). In the present context, the hat is used only to distinguish *modeled data* from *observed data*. Also, we allow Θ to be infinite dimensional (although typically metrizable). By *parametric model* we just mean a set of p.m.s that is indexed by a parameter $\boldsymbol{\theta} \in \Theta$. In this sense, we also deviate somewhat from typical terminology that requires Θ to be finite dimensional. For every $\boldsymbol{\theta} \in \Theta$, let $h(\cdot, \boldsymbol{\theta}) : \mathcal{X} \to \mathcal{Y}$ be $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$-measurable, so that $h(\mathbf{x}_t; \boldsymbol{\theta}) : \Omega \to \mathcal{Y}$ is $\mathcal{F}/\mathfrak{B}(\mathcal{Y})$-measurable $\forall \boldsymbol{\theta} \in \Theta$ and every $t \in \mathbb{Z}$. Define $\mathcal{H}_\Theta(\mathcal{X}) := \{h(\cdot; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ as the space of parametric functions defined on $\mathcal{X}$ generated by Θ under the mapping $h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ where $h_\mathcal{X}(\boldsymbol{\theta}) := h(\cdot; \boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) \, \forall \boldsymbol{\theta} \in \Theta$. The mapping $h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ shall be called a *parameterization mapping*. Immediately, given $\mathrm{D}_0^{\mathbf{x}}$, for

[4] Thus the extension covers what is probably the most common formulation of the nonlinear regression $y_t = h_0(\mathbf{x}_t) + \epsilon_t$ where ϵ_t is unobserved. Here we follow White (1980b) in considering an extremely simple univariate nonlinear regression framework. This allows us to simplify the argument by focusing on what is really essential, therefore avoiding distractions created by unnecessary considerations.

[5] $\mathfrak{B}(\mathcal{W}_\infty) = \mathfrak{B}(\mathcal{X}_\infty) \otimes \mathfrak{B}(\mathcal{Y}_\infty)$ the product σ-algebra; Dudley (2002, p.119).

every $\boldsymbol{\theta} \in \Theta$, $h(\cdot, \boldsymbol{\theta}) \equiv h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$ induces a p.m. $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{y}}$ indexed by $\boldsymbol{\theta}$ on $\mathfrak{B}(\mathcal{Y}_\infty)$ according to $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{y}}(B_y) = \mathrm{D}_0^{\mathbf{x}} \circ h^{-1}(B_y, \boldsymbol{\theta})$ for every $(B_y, \boldsymbol{\theta}) \in \mathfrak{B}(\mathcal{Y}_\infty) \times \Theta$. The triplet $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty), \mathrm{D}_{\boldsymbol{\theta}}^{\hat{y}})$ is thus an element of a family of measure spaces indexed by $\boldsymbol{\theta}$. Now, define accordingly $\hat{\mathbf{w}} := \{\hat{\mathbf{w}}_t = (\hat{y}_t, \mathbf{x}_t), t \in \mathbb{Z}\}$, the counterpart of $\mathbf{w}$, with $\hat{\mathbf{w}}_t(\omega) \in \mathcal{W}\ \forall t \in \mathbb{N}$ and $\hat{\mathbf{w}}(\omega) \in \mathcal{W}_\infty$ which lives in $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$. As a result, given $\mathrm{D}_0^{\mathbf{x}}$, for every $\boldsymbol{\theta} \in \Theta$, $h_{\mathcal{X}}(\boldsymbol{\theta})$ induces also a p.m. $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$ on $\mathfrak{B}(\mathcal{W}_\infty)$ so that $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ is also indexed by $\boldsymbol{\theta}$. For clarity, we let D denote the functional that, given $\mathrm{D}_0^{\mathbf{x}}$ on $\mathfrak{B}(\mathcal{X}_\infty)$, maps elements of $\mathcal{H}_\Theta(\mathcal{X})$ onto the space $\mathcal{D}_\Theta^{\hat{\mathbf{w}}}$ of p.m.s defined on the sets of $\mathfrak{B}(\mathcal{W}_\infty)$ and generated by Θ through h, i.e. $\mathrm{D} : \mathcal{H}_\Theta(\mathcal{X}) \to \mathcal{D}_\Theta^{\hat{\mathbf{w}}}$ (with $\mathcal{D}_\Theta^{\hat{\mathbf{w}}} = \{\mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \mid \mathrm{D}_0^{\mathbf{x}}\}$) satisfies $\mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}) = \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}\ \forall \boldsymbol{\theta} \in \Theta$ with $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}(B_{\mathbf{w}}) \equiv \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}(B_{\mathbf{x}} \times B_y) = \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}(B_{\mathbf{x}} \times \mathcal{Y}_\infty | \mathcal{X}_\infty \times B_y) \cdot \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}(\mathcal{X}_\infty \times B_y) = I_{(B_{\mathbf{x}} = h^{-1}(B_y))} \cdot \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}(\mathcal{X}_\infty \times B_y)$, $B_{\mathbf{x}} \in \mathfrak{B}(\mathcal{X})$ and $B_y \in \mathfrak{B}(\mathcal{X})$ with $I_{(B_{\mathbf{x}} = h^{-1}(B_y))} = 1$ when $B_{\mathbf{x}} = h^{-1}(B_y)$ and $I_{(B_{\mathbf{x}} = h^{-1}(B_y))} = 0$ otherwise.[6] Clearly, since there is no guarantee that $h_0 \in \mathcal{H}_\Theta(\mathcal{X})$, i.e. that $\exists \boldsymbol{\theta}_0 \in \Theta : h(\mathbf{x}_t(\omega); \boldsymbol{\theta}_0) = h_0(\mathbf{x}_t(\omega))\ \forall \mathbf{x}_t(\omega) \in \mathcal{X}$, it might well be the case that $\nexists \boldsymbol{\theta}_0 \in \Theta : \mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}_0) = \mathrm{D}_0^{\mathbf{w}}$ so that $\mathrm{D}_0^{\mathbf{w}} \notin \mathcal{D}_\Theta^{\hat{\mathbf{w}}}$. Note here that the statement $\exists \boldsymbol{\theta}_0 \in \Theta : h(\mathbf{x}_t; \boldsymbol{\theta}_0) = h_0(\mathbf{x}_t)\ \forall \mathbf{x}_t \in \mathcal{X}$ is to be understood in the function equivalence sense (Kolmogorov and Fomin (1975), p.288); i.e. we write $h_{\mathcal{X}}(\boldsymbol{\theta}_0) = h_0$ if and only if $\mathrm{D}_0^{\mathbf{x}}\{B_{\mathbf{x}} \in \mathfrak{B}(\mathcal{X}_\infty) : h_0(B_{\mathbf{x}}) \neq h(B_{\mathbf{x}}; \boldsymbol{\theta})\} \equiv P\{\omega \in \Omega : h_0(\mathbf{x}(\omega)) \neq h(\mathbf{x}(\omega); \boldsymbol{\theta})\} = 0$. The same applies to similar statements throughout the paper. The sets $\mathcal{H}_\Theta(\mathcal{X})$ and Θ are thus naturally partitioned into equivalence classes by the mappings D and $h_{\mathcal{X}}$ respectively, with classes taking the form $\{h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) : \mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}) = \mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}')\}$ and $\{\boldsymbol{\theta} \in \Theta : h_{\mathcal{X}}(\boldsymbol{\theta}) = h_{\mathcal{X}}(\boldsymbol{\theta}')\}$ respectively. This framework is convenient as the identification problem is not the one we which to focus on. We shall address this point later. Finally, let,

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta} \in \Theta} Q_T(y_{\mathbf{T}}, \mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta}) \equiv \arg\min_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{w}_{\mathbf{T}}; \boldsymbol{\theta})$$

denote the extremum estimator of interest, a map $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$. For the moment, let us adopt

[6] By "given $\mathrm{D}_0^{\mathbf{x}}$" we mean that $\mathrm{D} : \mathcal{H}_\Theta(\mathcal{X}) \to \mathrm{D}_\Theta^{\hat{\mathbf{w}}}$ can be obtained from $\mathrm{D}^* : \mathcal{D}^{\mathbf{x}} \times \mathcal{H}_\Theta(\mathcal{X}) \to \mathrm{D}_\Theta^{\hat{\mathbf{w}}}$ as $\mathrm{D} = \mathrm{D}^*(\mathrm{D}_0^{\mathbf{x}}, \cdot) : \mathcal{H}_\Theta(\mathcal{X}) \to \mathrm{D}_\Theta^{\hat{\mathbf{w}}}$ where $\mathrm{D}_0^{\mathbf{x}} \in \mathcal{D}^{\mathbf{x}}$. Also note that every $B_{\mathbf{w}} \in \mathfrak{B}(\mathcal{W}_\infty)$ takes the form $B_{\mathbf{w}} = B_{\mathbf{x}} \times B_y$ with $B_{\mathbf{x}} \in \mathfrak{B}(\mathcal{X}_\infty)$ and $B_y \in \mathfrak{B}(\mathcal{Y}_\infty)$ (Dudley (2002, p.118)); see also Davidson (1994, p.115) for notation.

this notation to stress that Q_T is a function of $\boldsymbol{\theta} \in \Theta$. Hence, we write $Q_T : \mathcal{W}_T \times \Theta \to \mathbb{R}$ where $\mathcal{W}_T := \mathcal{Y}_T \times \mathcal{X}_T$ with $\mathcal{Y}_T := \times_{t=1}^{t=T} \mathcal{Y}$ and $\mathcal{X}_T := \times_{t=1}^{t=T} \mathcal{X}$ so that $\mathbf{w_T}(\omega) \in \mathcal{W}_T$. Note however that we could have written $Q_T(\mathbf{x_T}, \mathbf{h}_0(\mathbf{x_T}), \mathbf{h}(\mathbf{x_T}; \boldsymbol{\theta}))$ where $\mathbf{h}_0(\mathbf{x_T}) := \{h_0(\mathbf{x}_t)\}_{t=1}^T \equiv y_\mathbf{T}$ and $\mathbf{h}(\mathbf{x_T}; \boldsymbol{\theta}) := \{h(\mathbf{x}_t; \boldsymbol{\theta})\}_{t=1}^T \equiv \hat{y}_\mathbf{T}$ to highlight the fact that the criterion Q_T is a function of $\boldsymbol{\theta}$ through $h_\mathcal{X}$, and as a result, that $\hat{\boldsymbol{\theta}}_T$ depends also on the choice of parameterization. For simplicity however, since $h_\mathcal{X}$ is often fixed prior to estimation, an explicit account of this relation is seldom considered. Clearly, nothing is lost in adopting either notational convention as long as these considerations are kept in mind.

Finally, note that we can also address the problem of approximating the true distribution D_0^y of a random variable y_t from a family of parametric distributions $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{y}}$, simply by taking $\mathrm{D}_0^{\mathbf{x}}$ to be known. For example, taking $\mathbf{x}$ to be independently identically distributed, with $n_x = 1$ and $x_t \sim \mathcal{U}([0,1])$ where $\mathcal{U}([0,1])$ denotes the uniform distribution on $[0,1]$, implies that $\mathrm{D}_0^y = h_0^{-1}$ is the true unknown distribution of y_t and that $h_\mathcal{X}^{-1}(\boldsymbol{\theta})$ defines the distribution function $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{y}} = h^{-1}(\cdot; \boldsymbol{\theta})$ of $\hat{y}_t$. Also note that the results in this paper extend trivially to a formulation of the regression model where $y_t = h_0(\mathbf{x}_t) + \epsilon_t$ whenever the distribution of ϵ_t is known, or more generally to $y_t = H(h_0(\mathbf{x}_t), \epsilon_t)$, $\epsilon_t \sim F_\epsilon$ whenever H and F_ϵ are known.

3 Standard Formulation

Following White (1980b) and Domowitz and White (1982), consider for simplicity the regression model $y_t = h_0(\mathbf{x}_t)$ and a postulated parametric counterpart $\hat{y}_t = h(\mathbf{x}_t, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Existence of an estimator $\hat{\boldsymbol{\theta}}_T$ as described above follows immediately from lemma 2 of Jennrich (1969) and Pötscher and Prucha (1997, p.20, lemma 3.4); see also e.g. Brown and Purves (1973) and Stinchcombe and White (1992) for generalizations and extensions.

Assumption 1. *Θ is compact and $Q_T(\mathbf{w_T}(\omega); \cdot) : \Theta \to \mathbb{R}$ is a continuous function of $\boldsymbol{\theta} \in \Theta$ for every $\mathbf{w_T}(\omega) \in \mathcal{W}_T$,(i.e. every $\omega \in \Omega$). Also, $Q_T(\cdot; \boldsymbol{\theta}) : \mathcal{W}_T \to \mathbb{R}$ is a $\mathfrak{B}(\mathcal{W}_T)/\mathfrak{B}(\mathbb{R})$-measurable function of $\mathbf{w_T}$ for every $\boldsymbol{\theta} \in \Theta$.*

Lemma 1. (Existence) *Let Assumption 1 hold. Then there exists a measurable function* $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ *such that for every* $\omega \in \Omega$ *we have* $Q_T(\mathbf{w_T}(\omega); \hat{\boldsymbol{\theta}}_T(\omega)) = \min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w_T}(\omega); \boldsymbol{\theta})$.[7]

Consistency of $\hat{\boldsymbol{\theta}}_T$ has been obtained under conditions that ensure *(i)* the convergence of the sequence of continuous functions $Q_T : \mathcal{W}_T \times \Theta \to \mathbb{R}$ as $T \to \infty$, to a limit deterministic function $Q_\infty : \Theta \to \mathbb{R}$, uniformly on Θ, and *(ii)* the identifiable uniqueness of $\boldsymbol{\theta}_0 := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta})$. Definition 1 is adapted from Bates and White (1985).[8]

Definition 1. (Identifiable Uniqueness) *Suppose that* $\boldsymbol{\theta}_0$ *minimizes* Q_∞ *on* Θ. *Let* $S_0(\epsilon)$ *be an open ball centered at* $\boldsymbol{\theta}_0$ *with radius* $\epsilon > 0$. *Define the neighborhood* $\eta_0(\epsilon) \equiv S_0(\epsilon) \subset \Theta$ *with complement* $\eta_0(\epsilon)^c := \Theta \backslash \eta_0(\epsilon)$. *Then* $\boldsymbol{\theta}_0$ *is said to be identifiable unique on* Θ *if and only if for every* $\epsilon > 0$, $\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[Q_\infty(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}_0)] > 0$.

In general, the identifiable uniqueness of $\boldsymbol{\theta}_0$ allows for alternative formulations of consistency of extremum estimators in terms of non-compact parameter spaces, discontinuous criterion functions, as well as for dependence and heterogeneity of the underlying data. In particular, this condition can be formulated for sequences of minimizers $\boldsymbol{\theta}_0^T$ of a sequence of deterministic functions Q_∞^T to which the random criterion function Q_T converges. For the sake of simplicity however, we shall ignore this possibility. We thus focus only on the case where $Q_\infty^T \equiv Q_\infty \; \forall T$. Lemma 2 below is adapted from Pötscher and Prucha (1997, ch.3).

Assumption 2. $\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\mathbf{w_T}; \boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \overset{a.s.}{\to} 0$.

Assumption 3. $Q_\infty : \Theta \to \mathbb{R}$ *has an identifiably unique minimizer* $\boldsymbol{\theta}_0$.

Lemma 2. (Consistency) *Let Assumptions 2 and 3 hold. Define* $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ *such that* $\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w_T}; \boldsymbol{\theta})$. *Then,* $\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0 \overset{a.s.}{\to} 0$ *as* $T \to \infty$.

[7] Assumption 1 and Lemma 1 can be further generalized to accommodate cases under which Q is continuous on Θ a.s. but not necessarily for all $\omega \in \Omega$; see e.g. Gallant and White (1988, p.14).

[8] The uniform convergence condition is typically stronger than required; see e.g. Van der Vaart and Wellner (1996, p.286) and Pötscher and Prucha (1997, p.24).

Assumption 1 can be added to Assumptions 2 and 3 in the lemma above to ensure that $\hat{\boldsymbol{\theta}}_T$ is a random variable for every T. This however, is not a necessary condition for the measurability of $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$, nor is it necessary to obtain $\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0 \overset{a.s.}{\to} 0$ as the lemma itself testifies. Still, when it is appropriate to work under the influence of Assumption 1, then, given the compactness of Θ and the continuity of Q_∞, the identifiable uniqueness condition turns out to be satisfied as long as the set $\arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta})$ is a singleton, i.e. $\boldsymbol{\theta}_0$ is unique. Sometimes, it will be perfectly fine to consider the set of elementary Assumptions 1, 2 and 4 (below), and to work with the following lemma adapted from Amemiya (1985).

Assumption 4. *$Q_\infty : \Theta \to \mathbb{R}$ attains a unique minimum at $\boldsymbol{\theta}_0$.*

Lemma 3. (Consistency) *Let Assumptions 1, 2 and 4 hold. Define $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ such that $\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w_T}; \boldsymbol{\theta})$. Then $\hat{\boldsymbol{\theta}}_T \overset{a.s.}{\to} \boldsymbol{\theta}_0$ as $T \to \infty$.*

Finally, a few comments on Assumptions 2-4. Well known conditions for $\sup_{\theta\in\Theta} |Q_T(\cdot;\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \to 0$ a.s. on a totally bounded metric space Θ are *(i)* $Q_T(\cdot;\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}) \to 0$ a.s. pointwise for every $\boldsymbol{\theta} \in \Theta$ and *(ii)* $\{Q_T(\cdot, \boldsymbol{\theta}), T \in \mathbb{N}\}$ be strongly asymptotically uniformly stochastically equicontinuous (see e.g. Newey (1991) and Andrews (1992)). When $\{Q_T(\cdot, \boldsymbol{\theta}), T \in \mathbb{N}\}$ is a sequence of normalized partial sums, Assumption 2 boils down to a uniform law of large numbers. These have been achieved under alternative primitive conditions that allow for varying degrees of dependence and heterogeneity in the data; see e.g. Gallant and White (1988, ch.3) and Pötscher and Prucha (1997, ch.5) and references therein.[9]

Statistical tests have been developed that are aimed at verifying whether (at least a part of) the host of assumptions involved in these arguments actually hold in practice. To some extent, this allows researchers to conclude with varying degree of confidence on whether the

[9]When $Q_T(\cdot;\boldsymbol{\theta}) \equiv T^{-1}\sum_{t=1}^{T} q(\mathbf{w}_t;\boldsymbol{\theta})$ uniform convergence is equivalent to $\mathcal{Q} = \{q(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ being a class of Glivenko-Cantelli functions. This requires fundamentally the compactness of Θ, continuity of $q(\mathbf{w_T};\cdot) : \Theta \to \mathbb{R}$ for every $\mathbf{w_T} \in \mathcal{W}_T$ (i.e. every $\omega \in \Omega$) and that $q(\cdot;\boldsymbol{\theta})$ be dominated by an integrable function for every $\boldsymbol{\theta} \in \Theta$.

consistency of any given extremum estimator holds. Unfortunately, in the context of misspecified models, it is often hard to conclude wether the identifiable uniqueness assumption is satisfied. As mentioned in the introduction, some authors have attempted (often successfully) to relax this condition and allow for multiple minima. This might be a fruitful approach in some circumstances, albeit one that we shall not follow here. Below we investigate transparent primitive conditions on both the estimator and the model at hand that imply identifiable uniqueness. These conditions take place in a deterministic setting as they pertain to the limit criterion function. The uniform convergence of the criterion function, established in a probabilistic setting, will be left unaltered.

4 Limit Divergence Functions

As mentioned before, the functional dependence of Q_T on the choice of parameterization mapping $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ is typically omitted for notational convenience. Recall from Section 2 that we could have written $Q_T(\mathbf{x_T}, \mathbf{h}_0(\mathbf{x_T}), \mathbf{h}(\mathbf{x_T}; \boldsymbol{\theta})) \equiv Q_T(\mathbf{w_T}, \hat{\mathbf{w}}_\mathbf{T}(\boldsymbol{\theta}))$ thus having $\hat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w_T}, \hat{\mathbf{w}}_\mathbf{T}(\boldsymbol{\theta}))$. This clarifies the reason why the deterministic limit criterion is often appropriately described as a function $Q_\infty^{\mathcal{D}}$ of the underlying joint p.m.s of $\mathbf{w_T}$ and $\hat{\mathbf{w}}_\mathbf{T}$ (or some of its features, e.g. moments) implicitly defined by the measurable mappings h_0 and $h_{\mathcal{X}}(\boldsymbol{\theta})\ \forall\, \boldsymbol{\theta} \in \Theta$, given $\mathrm{D}_0^{\mathbf{x}}$. Below, we shall restrict attention to limit criterion functions $Q_\infty : \Theta \to \mathbb{R}$ that assume the special form $Q_\infty(\boldsymbol{\theta}) = Q_\infty^{\mathcal{D}}(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})\ \forall\, \boldsymbol{\theta} \in \Theta$ where $\mathrm{D}_0^{\mathbf{w}}$ and $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$ are the p.m.s of the processes $\mathbf{w}$ and $\hat{\mathbf{w}}$ defined in Section 2. When $Q_\infty^{\mathcal{D}}$ is a divergence on a space of probability measures containing $\mathrm{D}_0^{\mathbf{w}}$ and $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}\ \ \forall\, \boldsymbol{\theta} \in \Theta$, then $\boldsymbol{\theta}_0$ is, by definition, the minimizer of that divergence.[10] By establishing a bijection between the space of probability

[10] Given a limit criterion function $Q_\infty : \Theta \to \mathbb{R}$ and a flexible definition of divergence (e.g. a pre-metric), it is often possible to find a divergence $Q_\infty^{\mathcal{D}}$ on the space of p.m.s satisfying $\arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty^{\mathcal{D}}(\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}, \mathrm{D}_0^{\mathbf{w}}) = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta})$. In this sense, the results discussed here are generally applicable to a large number of extremum estimator, even those not initially conceived as minimum divergence estimators.

measures and the space of regression functions containing h_0 and $h_{\mathcal{X}}(\boldsymbol{\theta})\ \forall\, \boldsymbol{\theta} \in \Theta$, we translate the problem of divergence minimization from the space distributions to the space of regression functions. Finally, we also note that it is possible to simplify the argument when there exists a strictly increasing function g such that $g \circ Q_\infty^{\mathcal{D}}$ establishes a metric, norm or inner product on the space of distributions (and regression functions), in which case $\boldsymbol{\theta}_0$ is characterized as the minimizer of this distance between $\mathrm{D}_0^{\mathbf{w}}$ and $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$ (or h_0 and $h_{\mathcal{X}}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$). In Section 5, we review conditions for (strong) uniqueness of best approximations of $\mathrm{D}_0^{\mathbf{w}}$ by $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$ in the space of probability measures (or h_0 by $h_{\mathcal{X}}(\boldsymbol{\theta})$ in the space of regression functions).

Consider the space $\mathcal{H}(\mathcal{X})$ satisfying $\mathcal{H}_\Theta(\mathcal{X}) \subseteq \mathcal{H}(\mathcal{X})$ and $h_0 \in \mathcal{H}(\mathcal{X})$. The smallest $\mathcal{H}(\mathcal{X})$ thus being $\mathcal{H}(\mathcal{X}) = \mathcal{H}_\Theta(\mathcal{X}) \times \{h_0\}$ when $h_0 \notin \mathcal{H}_\Theta(\mathcal{X})$ or simply $\mathcal{H}(\mathcal{X}) = \mathcal{H}_\Theta(\mathcal{X})$ when $\exists \boldsymbol{\theta}_0 \in \Theta : h_{\mathcal{X}}(\boldsymbol{\theta}_0) = h_0$ which implies $h_0 \in \mathcal{H}_\Theta(\mathcal{X})$. Now, define the space of p.m.s $\mathcal{D}^{\mathbf{w}} = \{\mathrm{D}(h), h \in \mathcal{H}(\mathcal{X})\}$ by extending the functional D encountered before to be defined on $\mathcal{H}(\mathcal{X})$ instead of $\mathcal{H}_\Theta(\mathcal{X})$ only; i.e. now $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$, so that in general D is such that $\mathrm{D} \circ h = \mathrm{D}_h^{\mathbf{w}}$ with $\mathrm{D}_h^{\mathbf{w}}$ satisfying $\mathrm{D}_h^{\mathbf{w}}(B_{\mathbf{w}}) \equiv \mathrm{D}_h^{\mathbf{w}}(B_y, B_{\mathbf{x}}) \equiv \mathrm{D}_h^{y|\mathbf{x}}(B_y) \cdot \mathrm{D}_0^{\mathbf{x}}(B_{\mathbf{x}}) \equiv \mathrm{D}^{y|\mathbf{x}}(B_y|h) \cdot \mathrm{D}_0^{\mathbf{x}}(B_{\mathbf{x}})\ \forall (h, B_{\mathbf{w}}) \in \mathcal{H}(\mathcal{X}) \times \mathfrak{B}(\mathcal{W}_\infty)$. It thus follows that $\mathcal{D}^{\mathbf{w}}$ satisfies $\mathcal{D}_\Theta^{\hat{\mathbf{w}}} \subseteq \mathcal{D}^{\mathbf{w}}$ and $\mathrm{D}_0^{\mathbf{w}} \in \mathcal{D}(\mathcal{X})$. The smallest $\mathcal{D}^{\mathbf{w}}$ corresponding to the smallest $\mathcal{H}(\mathcal{X})$ and defined as $\mathcal{D}^{\mathbf{w}} = \mathcal{D}_\Theta^{\hat{\mathbf{w}}} \times \{\mathrm{D}_0^{\mathbf{w}}\}$ for misspecified models or simply $\mathcal{D}^{\mathbf{w}} = \mathcal{D}_\Theta^{\hat{\mathbf{w}}}$ when the model is well specified, i.e. when $\exists\, \boldsymbol{\theta}_0 \in \Theta : \mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}_0) = \mathrm{D} \circ h_0 = \mathrm{D}_0^{\mathbf{w}}$ (which implies $\mathrm{D}_0^{\mathbf{w}} \in \mathcal{D}_\Theta^{\hat{\mathbf{w}}}$). Finally, let the following assumption restrict the class of extremum estimators under consideration.

Assumption 5. *The limit criterion $Q_\infty : \Theta \to \mathbb{R}$ takes the form $Q_\infty(\boldsymbol{\theta}) \equiv Q_\infty^{\mathcal{D}}(\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}, \mathrm{D}_0^{\mathbf{w}})$ $\forall\, \boldsymbol{\theta} \in \Theta$ where $Q_\infty^{\mathcal{D}} : \mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$ is a divergence $d_{\mathcal{D}} \equiv Q_\infty^{\mathcal{D}}$ on $\mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}}$.*

Since under Assumption 5, $Q_\infty^{\mathcal{D}}$ is a function of $\boldsymbol{\theta} \in \Theta$ only through the p.m. $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}} \equiv \mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{D}_\Theta^{\hat{\mathbf{w}}}$, we require that $\nexists (\boldsymbol{\theta}', \boldsymbol{\theta}'') \in \Theta \times \Theta$ satisfying $\boldsymbol{\theta}' \neq \boldsymbol{\theta}''$ and such that $\mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}') = \mathrm{D} \circ h(\cdot; \boldsymbol{\theta}'')$ as a minimal condition for uniqueness. In several contexts, this is called the *identification condition* (see e.g. Hsiao (1983)). As mentioned in the introduction, there is no universal strict relation between identifiable uniqueness and identification. In most cases

of interest however, the absence of observationally equivalent elements in Θ is a necessary condition for identifiable uniqueness to hold. This is also the case in our formulation where the limit criterion $Q_\infty^{\mathcal{D}}$ takes the form of a divergence on $\mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}}$.[11]

In the present context, for $\mathrm{D} \circ h_{\mathcal{X}} : \Theta \to \mathcal{D}^{\mathbf{w}}$ to be injective, a necessary and sufficient condition is that both $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ and $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$ be one-to-one. Now, the injective nature of D is often unverifiable, since it is in the very nature of statistical inference that the true probability measure $\mathrm{D}_0^{\mathbf{w}}$ be not known. In simple cases, depending on the complexity of $\mathcal{H}(\mathcal{X})$, it might be possible to find convincing evidence that $\mathrm{D}_0^{\mathbf{x}}$ is rich enough for D to be injective, based on observed data alone.[12] Yet, this is not always the case and little can be done about it as long as $\mathrm{D}_0^{\mathbf{w}}$ is to remain unknown. There is thus no point in discussing this issue further and we proceed under the common assumption that the data is "rich enough" for different elements of $\mathcal{H}(\mathcal{X})$ to be identified as such.[13] Clearly, the researcher might feel more or less comfortable in imposing this assumption depending on the complexity of $\mathcal{H}(\mathcal{X})$ and on the evidence contained in observed data. Still, imposing some condition on the richness of the data seems simply unavoidable. As mentioned in Section 2, it is important to note that this assumption is already embodied in the function equivalence framework adopted here, so that $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$ is bijective by construction.

It is thus evident that the one-to-one nature of the composition $\mathrm{D} \circ h_{\mathcal{X}} : \Theta \to \mathcal{D}^{\mathbf{w}}$ is to be understood fundamentally as a restriction on the construction of the model (in particular on the parameterization mapping $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$) as it does not concern the estimation procedure nor does it involve considerations about the data generating process beyond those already covered by the function equivalence framework adopted throughout the paper. Also, note that since the parameterization mapping $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ is surjective by construction, and $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$ is bijective (also by construction), the only property of concern to us is

[11] This would not be the case if the limit criterion was instead defined more generally on e.g. $\Omega \times \Theta$.

[12] Think e.g. of a simple linear regression with observed $\mathbf{w_T}$ providing evidence of a rich $\mathrm{D}_0^{\mathbf{w}}$.

[13] A "rich" data setting should exclude e.g. the presence of degenerate and collinear-type random variables.

that $h_{\mathcal{X}}$ be injective. This is generally verifiable for any given class of parametric functions posited by the researcher, and it is controlled by the researcher, so it should be satisfied by an appropriate formulation of the regression model and the parameter space Θ.[14] Still, we let the injective nature of $h_{\mathcal{X}}$ be sated as an assumption for future reference and verification.

Assumption 6. *$h_{\mathcal{X}} : \Theta \to \mathcal{H}_{\Theta}(\mathcal{X})$ is injective.*

This assumption implies by construction that both $h_{\mathcal{X}} : \Theta \to \mathcal{H}_{\Theta}(\mathcal{X})$ and $\mathrm{D} \circ h_{\mathcal{X}} : \Theta \to \mathcal{D}^{\hat{\mathbf{w}}}_{\Theta}$ are bijective. As a result, we can now identify Θ with $\mathcal{H}_{\Theta}(\mathcal{X})$ and $\mathcal{D}^{\hat{\mathbf{w}}}_{\Theta}$. Note also that since $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$ is bijective, we can identify $\mathcal{H}(\mathcal{X})$ with $\mathcal{D}^{\mathbf{w}}$.

The fact that Assumption 6 is sufficient for the identification condition to hold has an important practical implication. Identifiable uniqueness and identification are sometimes equivalent concepts in applications involving well-specified models. For example, when $\mathrm{D}^{\hat{\mathbf{w}}}_0 \in \mathcal{D}^{\mathbf{w}}_{\Theta}$, Θ is compact and $Q^{\mathcal{D}}_{\infty}$ is a continuous pre-metric, then identification is both necessary and sufficient for the identifiable uniqueness of $\boldsymbol{\theta}_0$.[15] The results discussed here are thus especially relevant for misspecified models. They are not necessarily interesting otherwise (since identifiable uniqueness would require only verification of 6).

Indeed, it is precisely when $h_0 \notin \mathcal{H}_{\Theta}(\mathcal{X}) \Leftrightarrow \mathrm{D}^{\mathbf{w}}_0 \notin \mathcal{D}^{\hat{\mathbf{w}}}_{\Theta}$ that the present formulation of the problem becomes advantageous. In particular, it is useful to note that given $\mathrm{D}^{\mathbf{x}}_0$, then there exists a functional $Q^{\mathcal{H}}_{\infty}$ that maps pairs of elements from $\mathcal{H}(\mathcal{X})$ to $\mathbb{R}$, such that $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \Theta} Q^{\mathcal{D}}_{\infty}(\mathrm{D}^{\mathbf{w}}_0, \mathrm{D}^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}}) \equiv \arg\min_{\boldsymbol{\theta} \in \Theta} Q^{\mathcal{H}}_{\infty}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$. Writing $Q^{\mathcal{H}}_{\infty} : \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X}) \to \mathbb{R}^+_0$ is convenient because it conveys the notion of the limiting criterion establishing a divergence $d_{\mathcal{H}}$ on $\mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$. Clearly, $d_{\mathcal{H}}$ is induced by $d_{\mathcal{D}}$ on $\mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$ through D according to $d_{\mathcal{H}}(h_1, h_2) = d_{\mathcal{D}}(\mathrm{D}(h_1), \mathrm{D}(h_2)) \; \forall (h_1, h_2) \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$. Given $h_{\mathcal{X}}$ and $\mathrm{D}^{\mathbf{x}}_0$, the limit $\boldsymbol{\theta}_0$ is thus to be seen as the element in Θ that minimizes the divergence $d_{\mathcal{H}}$ between $h_0 \in \mathcal{H}(\mathcal{X})$

[14] As we shall see in Section 7, verification of Assumption 6 is often a straightforward exercise.

[15] The pre-metric is associated here with a divergence that satisfying non-negativity $d_{\mathcal{H}}(h_1, h_2) \geq 0 \; \forall (h_1, h_2) \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$ and identity of indiscernibles $d_{\mathcal{H}}(h_1, h_2) = 0$ if and only if $h_1 = h_2 \; \forall (h_1, h_2) \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$.

and $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_{\Theta}(\mathcal{X}) \subseteq \mathcal{H}(\mathcal{X})$. This is stated concisely as $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$ where $d_{\mathcal{H}}(h_0, h_{\mathcal{X}}) : \Theta \to \mathbb{R}_0^+$. The employed notion of divergence can be quite general, such as e.g. coinciding with that of a pre-metric, pseudo-metric or quasi-metric. As mentioned before, even though there is no guarantee that $h_0 \in \mathcal{H}_{\Theta}(\mathcal{X})$, we shall see that under certain conditions $\exists \boldsymbol{\theta}_0 \in \Theta : d_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}_0)) < d_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta})) \; \forall (\boldsymbol{\theta} \neq \boldsymbol{\theta}_0) \in \Theta$, and hence, that $h_{\mathcal{X}}(\boldsymbol{\theta}_0)$ is the unique best approximation from $\mathcal{H}_{\Theta}(\mathcal{X})$ to h_0 in $\mathcal{H}(\mathcal{X})$ w.r.t. $d_{\mathcal{H}}$. This implies, under Assumption 6, that $\boldsymbol{\theta}_0$ is the unique minimizer of $Q_{\infty}^{\mathcal{H}}$.

Finally, we assume that an appropriate transformation of the limit criterion function yields us with a metric or norm. We emphasize that the only purpose of this assumption is that of retaining the simplicity of the argument, keeping technical requirements to a minimum and allowing us to focus on what is essential. This assumption allows us to make use of the "classical" theorems on existence and uniqueness of best approximations produced in the field of Approximation Theory, which have been naturally obtained for metric, normed and inner product spaces; see Cheney (1982) for a detailed list of existence and uniqueness (and other) accomplishments in the field. Even though equivalent results exist for non-metric divergences such as e.g. semi-metrics, pseudo-metrics or quasi-norms, clarity dictates that we consider here only the simpler results available for standard distances.[16] A sufficient requirement in this context is hence that there exists a continuous strictly increasing function g such that $Q_{\infty}^{\mathcal{H}}$ induces a metric on $\mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$.[17]

Assumption 7. *There exists a continuous strictly increasing function $g : \mathbb{R} \to \mathbb{R}_0^+$ such that $d_{\mathcal{D}}^* \equiv g \circ Q_{\infty}^{\mathcal{D}} : \mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$ is a metric.*

[16] These results shed some light on the pathologies identified by Donoho and Liu (1988) concerning the consistency of minimum distance estimators.

[17] As we shall see in section 7, it is often straightforward to verify if Assumption 7 holds.

5 Strong Unicity of Best Approximations

This section reviews some important results stemming from the field of Approximation Theory. The reader already familiar with this literature might find it preferable to proceed directly to Section 6. Observe first the following useful definitions available e.g. in Cheney (1974), Ahuja et al. (1977), Nurberger (1979) and Narang (1981). Let $(\mathcal{B}, d_{\mathcal{B}})$ be a linear metric space. Consider a subset $\mathcal{A} \subset \mathcal{B}$. A *projection mapping* is a set valued map $P_{d_{\mathcal{B}}}^{\mathcal{A}} : \mathcal{B} \to 2^{\mathcal{A}}$ satisfying $P_{d_{\mathcal{B}}}^{\mathcal{A}}(b) := \{a_0 \in \mathcal{A} : d_{\mathcal{B}}(b, a_0) \leq d_{\mathcal{B}}(b, a), a \in \mathcal{A}\} \; \forall \, b \in \mathcal{B}$, where $2^{\mathcal{A}}$ denotes the power set of $\mathcal{A}$. Note that $P_{d_{\mathcal{B}}}^{\mathcal{A}}(b)$ is the set of elements of best approximation of $b \in \mathcal{B}$ in $\mathcal{A}$, under $d_{\mathcal{B}}$. A set $\mathcal{A} \subset \mathcal{B}$ is then called *proximinal* if $P_{d_{\mathcal{B}}}^{\mathcal{A}}(b)$ is non-empty for every $b \in \mathcal{B}$ and *semi-Chebyshev* if $P_{d_{\mathcal{B}}}^{\mathcal{A}}(b)$ contains at most one element for every $b \in \mathcal{B}$. A set that is both proximinal and semi-Chebyshev is called *Chebyshev*. Note furthermore that a metric space $(\mathcal{B}, d_{\mathcal{B}})$ is said to be *strongly convex* if for every $(b_1, b_2) \in \mathcal{B} \times \mathcal{B}$ and every $t \in [0, 1]$ there exists a unique $b \in \mathcal{B}$ such that $d_{\mathcal{B}}(b_1, b) = (1-t) d_{\mathcal{B}}(b_1, b_2)$ and $d_{\mathcal{B}}(b, b_2) = t d_{\mathcal{B}}(b_1, b_2)$, i.e. each $t \in [0, 1]$ determines a unique element of the segment $[b_1, b_2] := \{b \in \mathcal{B} : d_{\mathcal{B}}(b_1, b) + d_{\mathcal{B}}(b, b_2) = d(b_1, b_2)\}$. Also, a strongly convex metric space $(\mathcal{B}, d_{\mathcal{B}})$ is said to be *strictly convex* if for every $(b_1, b_2) \in \mathcal{B} \times \mathcal{B}$ and $r > 0$, $d_{\mathcal{B}}(b_1, b_0) \leq r$, $d_{\mathcal{B}}(b_2, b_0) \leq r$ implies $d_{\mathcal{B}}(b, b_0) < r$ every $b \in]b_1, b_2[:= [b_1, b_2] \backslash \{b_1, b_2\}$ and fixed $b_0 \in \mathcal{B}$.[18]

When a function g exists that satisfies the properties postulated in Assumption 7, then, the following lemmas adapted from Cheney (1974), Ahuja et al. (1977), Powell (1981, p.4), Narang (1981) and Cheney (1982, p.4), are available to judge on the existence and uniqueness of a best approximation.

Lemma 4. (Existence on Metric Spaces) *Let $(\mathcal{B}, d_{\mathcal{B}})$ be a metric space and $\mathcal{A} \subseteq \mathcal{B}$ be compact. Then $\mathcal{A}$ is proximinal; i.e. for every $b \in \mathcal{B}$ there exists an element $a^* \in \mathcal{A}$, a best approximation to b from $\mathcal{A}$, satisfying $d_{\mathcal{B}}(a^*, b) \leq d_{\mathcal{B}}(a, b) \; \forall \, a \in \mathcal{A}$.*

[18] In a strictly convex metric space $(\mathcal{B}, d_{\mathcal{B}})$ if $(b_1, b_2) \in \mathcal{B} \times \mathcal{B}$ are two points in the boundary of a sphere, then the open line segment $]b_1, b_2[$ lies strictly inside the sphere.

Lemma 5. (Uniqueness on Metric Spaces) (i) *Let* $(\mathcal{B}, d_{\mathcal{B}})$ *be a strongly convex metric space and* $\mathcal{A} \subseteq \mathcal{B}$ *be convex. Then* $\mathcal{A}$ *is semi-Chebyshev; i.e. there exists at most one element* $a^* \in \mathcal{A}$ *such that* $d_{\mathcal{B}}(a^*, b) \leq d_{\mathcal{B}}(a, b)\ \forall\, a \in \mathcal{A}$. (ii) *Let* $(\mathcal{B}, d_{\mathcal{B}})$ *be a strictly convex metric space. Then* $\mathcal{A}$ *is semi-Chebyshev.*

The following lemma then follows from combining Lemmas 4 and 5 above, and theorem 2 in Ahuja et al. (1977).

Lemma 6. (Uniqueness on Metric Spaces) (i) *Let* $(\mathcal{B}, d_{\mathcal{B}})$ *be a strongly convex metric space and* $\mathcal{A}$ *be a compact convex subset of* $\mathcal{B}$. *Then* $\mathcal{A}$ *is Chebyshev; i.e. there exists a unique element* $a^* \in \mathcal{A}$ *such that* $d_{\mathcal{B}}(a^*, b) \leq d_{\mathcal{B}}(a, b)\ \forall\, a \in \mathcal{A}$. (ii) *Let* $(\mathcal{B}, d_{\mathcal{B}})$ *be a strictly convex metric space and* $\mathcal{A} \subset \mathcal{B}$ *compact. Then* $\mathcal{A}$ *is Chebyshev.*

Given the linearity of the function spaces considered under the usual definition of addition and multiplication by scalars, it is often beneficial to work on normed vector spaces. Some estimators might have limiting criterion functions $Q_\infty^{\mathcal{H}}$ for which $g \circ Q_\infty^{\mathcal{H}}$ is a metric on $\mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$ but not a norm (since the latter requires also homogeneity and translation invariance). When $g \circ Q_\infty^{\mathcal{H}}$ is a norm on $\mathcal{H}(\mathcal{X})$ however, simpler results from Approximation Theory are available for the uniqueness of best approximations. For this reason the following assumption is also introduced.

Assumption 8. *There exists a continuous strictly increasing function* $g : \mathbb{R} \to \mathbb{R}_0^+$ *such that* $g \circ Q_\infty^{\mathcal{D}}(\mathrm{D}, \mathrm{D}') \equiv \| \mathrm{D} - \mathrm{D}' \|_{\mathcal{D}}\ \ \forall\, (\mathrm{D}, \mathrm{D}') \in \mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}}$ *where* $\| \cdot \|_{\mathcal{D}} \equiv : \mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$ *is a norm.*

Consider now the natural extensions of the definition of strictly convex metric space to normed vector spaces. Let $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$ be a normed vector space. Then $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$ is said to be strictly convex if for every $(b_1, b_2) \in \mathcal{B} \times \mathcal{B}$ satisfying $\| b_1 \|_{\mathcal{B}} = \| b_2 \|_{\mathcal{B}} = 1$ the inequality $\| (1-t)b_1 + tb_2 \|_{\mathcal{B}} < 1$ holds for every $t \in]0, 1[$.

The following lemmas, which follow from those above for metric spaces, are adapted from Powell (1981, p.6,13-15) and Cheney (1982, p.20,23), and establish a few useful results on the existence and uniqueness of best approximations.

Lemma 7. (Existence on Normed Spaces) *Let* $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ *be a normed space and* $\mathcal{A}$ *a finite-dimensional subset of* $\mathcal{B}$*. Then* $\mathcal{A}$ *is proximinal.*

Lemma 8. (Uniqueness on Normed Spaces) (i) *Let* $\mathcal{A} \subset B$ *be a compact and strictly convex set in a normed linear space* $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$*. Then* $\mathcal{A}$ *is Chebyshev.* (ii) *Let* $\mathcal{A} \subset B$ *be a convex set in a strictly convex normed linear space* $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$*. Then* $\mathcal{A}$ *is semi-Chebyshev.* (iii) *Let* $\mathcal{A} \subset B$ *be a finite dimensional subspace of* $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$*. Then* $\mathcal{A}$ *is Chebyshev.*

Here it is important to point out that e.g. the well known L^1 and sup norms do not satisfy the strict convexity property of Lemma 8 (nor that of Lemma 5 in the induced metrics). Fortunately, the well known Haar condition allows us to overcome this limitation.

Definition 2. (Haar Condition) *A system of functions* $\{\psi_1, ..., \psi_n\}$ *with* $\psi_i : \mathcal{A} \subset \mathbb{R} \to \mathbb{R}$*,* $i = 1, ..., n$ *is said to satisfy the Haar condition on* $\mathcal{A}$ *if each* $\psi_i \in C(\mathcal{A})$*, the space of continuos functions on* $\mathcal{A}$*, for* $i = 1, ..., n$*, and if every set of* n *vectors of the form* $[\psi_1(a), ..., \psi_n(a)]$*,* $a \in \mathcal{A}$ *is independent; i.e. if for any given collection* $(a_1, ..., a_n) \in \times_{i=1}^{n} \mathcal{A}$*,* $a_i \neq a_j \; \forall \, i \neq j, i = 1, ...n, j = 1, ...n$*, the system has non-vanishing Vandermonde's determinant.*

A subspace $\mathcal{H}_\Theta(\mathcal{X}) \subset C(\mathcal{X})$ of generalized polynomials spanned by a system of functions $\{\psi_1, ..., \psi_n\}$ satisfying the Haar condition is called a *Haar subspace* of $C(\mathcal{X})$. The following lemma is adapted from Cheney (1982, p.81,219) and Powell (1981, 80,170). It is suitable for both L^1 and sup norm approximations.

Lemma 9. (Haar's Unicity theorem) *Let* $\mathcal{H}_\Theta(\mathcal{X})$ *be a Haar subspace of* $(C(\mathcal{X}), \|\cdot\|_1)$ *or* $(C(\mathcal{X}), \|\cdot\|_\infty)$ *and* $\mathcal{X}$ *a compact Hausdorff space. Then,* $\mathcal{H}_\Theta(\mathcal{X})$ *is Chebyshev.*

The Haar condition offers more than just a unicity characterization of best approximations on normed linear subspaces of $C(\mathcal{X})$. Under certain conditions, the element of best approximation from a Haar subspace is characterized by the strong unicity property. This property is relevant in the present context since the identifiable uniqueness condition in Assumption 3 can be derived from it. Following Newman and Shapiro (1963) and

Cheney (2000, p.80), let $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$ be a normed linear space and $a \in \mathcal{A} \subseteq \mathcal{B}$ an element of best approximation to $b_0 \in \mathcal{B}$ from $\mathcal{A}$. Then, a is said to be *strongly unique* if $\exists\, \gamma(b_0) > 0 : \| b_0 - a' \| > \| b_0 - a \| + \gamma \| a - a' \| \ \forall\, a' \in \mathcal{A}$.

Lemma 10. (Strong Unicity in Normed Linear Spaces) *Let $\mathcal{H}_\Theta(\mathcal{X})$ be a Haar subspace of $(C(\mathcal{X}), \| \cdot \|_\infty)$ and $\mathcal{X}$ a compact Hausdorff space. Then, for every $h_0 \in C(\mathcal{X})$ the element $h \in \mathcal{H}_\Theta(\mathcal{X})$ of best approximation to $h_0 \in C(\mathcal{X})$ is strongly unique; i.e. there exists a generalized polynomial $h \in \mathcal{H}_\Theta(\mathcal{X})$, $h = \sum_{i=1}^n \theta_i \psi_i$ where $\{\psi_1, ..., \psi_n\}$ satisfy the Haar condition, such that $\exists\, \gamma(h_0) > 0 : \| h_0 - h' \|_\infty > \| h_0 - h \|_\infty + \gamma \| h - h' \|_\infty \ \forall\, h' \in \mathcal{H}_\Theta(\mathcal{X})$.*

Unfortunately, Lemma 10 is available only under the sup norm. Furthermore, it is known since Wulbert (1971) that strong unicity of elements of best approximation is generally not available in smooth Banach spaces. This holds in particular in $L^p(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mu_{\mathcal{E}})$ spaces, with $1 < p < \infty$, where $(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mu_{\mathcal{E}})$ is a given measure space. Fortunately, the identifiable uniqueness property of Assumption 3 can also be derived from the concept of *strong unicity of order* α. Following Angelos and Egger (1984) and Lin (1989), let $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$ be a Banach space and $a \in \mathcal{A} \subseteq \mathcal{B}$ be an element of best approximation to $b_0 \in \mathcal{B}$ from $\mathcal{A}$. Then, a is said to be *strongly unique of order* α ($\alpha > 1$) if $\exists\, \gamma(b_0) > 0 : \| b_0 - a' \| > \| b_0 - a \| + \gamma \| a - a' \|^\alpha \ \forall\, a' \in \mathcal{A}$. The following lemma, adapted from Angelos and Egger (1984) and Lin (1989), reveals that this strong unicity property holds for finite-dimensional subspaces of $L^p(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mu_{\mathcal{E}})$ smooth Banach spaces ($1 < p < \infty$), or general subspaces of uniformly convex Banach spaces of type p. Note that a Banach space $(\mathcal{A}, \| \cdot \|)$ is said to be *uniformly convex* (Clarkson (1936)) if for every $0 < \epsilon \leq 2$ there exists a $\delta(\epsilon) > 0$ such that having $\| a_1 \| = \| a_2 \| = 1$ and $\| a_1 - a_2 \| \geq \epsilon$ implies $\| (a_1 + a_2)/2 \| \leq 1 - \delta(\epsilon)$. The function $\delta(\epsilon) : (0, 2] \to [0, 1]$ defined as $\delta(\epsilon) = \inf\{1 - \| a_1 + a_2 \| /2 | \ \| a_1 \| \leq 1, \| a_2 \| \leq 1, \| a_1 - a_2 \geq \epsilon\}$ is called the *modulus of convexity* of the Banach space $(\mathcal{A}, \| \cdot \|)$, and this space is said to be *uniformly convex of power type* p if there exists $\Delta > 0$ such that $\delta(\epsilon) \geq \Delta \epsilon^p$. The following lemma uses also a result of Hanner (1956) showing that L^p spaces with $1 < p < \infty$ are uniformly convex of power type

$\max\{2, p\}$, and the fact that strictly convex normed linear spaces are also uniformly convex; see e.g. Cheney (1974) or Cheney (1982, p.23).

Lemma 11. (Strong Unicity of Order α in Normed Linear Spaces) *(i) Let $\mathcal{A}$ be a finite-dimensional subspace of an $L^p(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mu_{\mathcal{E}})$ space with $1 < p < \infty$ and $\mathcal{E} \subseteq \mathbb{R}^{n_\varepsilon}$. Then, the element $a \in \mathcal{A}$ of best approximation to $b \in L^p(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mu_{\mathcal{E}})$, when it exists, is strongly unique of order $\alpha = \max\{p, 2\}$. (ii) Let $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ be a uniformly convex Banach space of power type p and let $\mathcal{A}$ be a subspace of $\mathcal{B}$. Then, an element $a \in \mathcal{A}$ of best approximation to $b \in \mathcal{B}$, when it exists, is strongly unique of order p.*

Also, similar results to those obtained above are available under weaker conditions on the employed notion of distance. Examples include Romaguera and Sanchis (2000) that deal with quasimetric spaces and Cobzas and Mustata (2006) that work with asymmetric normed linear spaces. While these formulations might offer more generality, we manage to achieve a significant simplification by restricting ourselves to the former case where $Q_\infty^{\mathcal{H}}$ induces a metric or norm on $\mathcal{H}(\mathcal{X})$. The reader should nevertheless bear in mind the limitations introduced by the simplifying assumption just mentioned. This is important as this restriction might prove to be relevant in several applications.

6 Consistency Restated

Finally, we are ready to restate the consistency results of Section 3 using alternative conditions. We note in particular that Assumption 3 (identifiable uniqueness of $\boldsymbol{\theta}_0$) and Assumption 4 (uniqueness of $\boldsymbol{\theta}_0$) used in Lemmas 2 and 3 to obtain the consistency of $\hat{\boldsymbol{\theta}}_T$ can now be substituted by sets of sufficient conditions that make use of the problem formulation discussed in Section 4 and the lemmas of Section 5 on the unicity of best approximations. Under the more restrictive assumptions of Lemma 3, which impose the compactness of Θ and continuity of Q_∞, showing the uniqueness of $\boldsymbol{\theta}_0$ is enough to obtain the consistency of $\hat{\boldsymbol{\theta}}_T$ since in this

setting, a unique $\boldsymbol{\theta}_0$ is automatically identifiably unique. In this simpler case, we will need only to make use of those lemmas establishing the uniqueness of best approximations covered in Section 5. It is under the less restrictive conditions of Lemma 2 that the results on strong unicity of best approximations become important since, in that case, $Q_\infty(\boldsymbol{\theta}_0)$ must be shown to be well separated without the aid of the compactness of Θ or the continuity of Q_∞.

As we have seen in the previous section, the uniqueness of $\boldsymbol{\theta}_0$ can be established either in the context of metric spaces or that of normed linear spaces. Depending on the problem, each formulation will be more or less advantageous in terms of verification.[19] Assumptions 9 and 10 below establish the conditions from which the uniqueness of $\boldsymbol{\theta}_0$ will be derived. These make use of the fact that every convex proximinal set is Chebyshev and are stated for future reference. Assumptions 11, 12 and 13 establish useful conditions for directly deriving the identifiable uniqueness of $\boldsymbol{\theta}_0$.

Assumption 9. (i) $(\mathcal{H}(\mathcal{X}), d^*_{\mathcal{H}})$ *is a strongly convex metric space and* $\mathcal{H}_\Theta(\mathcal{X})$ *a compact convex subset of* $\mathcal{H}(\mathcal{X})$*; or* (ii) $(\mathcal{H}(\mathcal{X}), d^*_{\mathcal{H}})$ *is a strictly convex metric space and* $\mathcal{H}_\Theta(\mathcal{X})$ *a compact subset of* $\mathcal{H}(\mathcal{X})$*.*

Assumption 10. (i) $(\mathcal{H}(\mathcal{X}), \|\cdot\|_{\mathcal{H}})$ *is a normed linear space and* $\mathcal{H}_\Theta(\mathcal{X})$ *a compact strictly convex subset of* $\mathcal{H}(\mathcal{X})$*; or* (ii) $(\mathcal{H}(\mathcal{X}), \|\cdot\|_{\mathcal{H}})$ *is a strictly convex normed vector space and* $\mathcal{H}_\Theta(\mathcal{X})$ *a finite dimensional convex subset of* $\mathcal{H}(\mathcal{X})$*.*

Assumption 11. $(\mathcal{H}(\mathcal{X}), \|\cdot\|_{\mathcal{H}}) = (C(\mathcal{X}), \|\cdot\|_\infty)$ *where* $\|\cdot\|_\infty$ *denotes the supremum norm, and for every* $\boldsymbol{\theta} \in \Theta$*, the elements* $h(\cdot;\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$ *accept a generalized polynomial representation* $h(\cdot,\boldsymbol{\theta}) = \sum_{i=1}^{n_h} \boldsymbol{\theta}_i h_i$ *where* $\{h_1, ..., h_n\}$ *satisfies the Haar condition.*

Assumption 12. $(\mathcal{H}(\mathcal{X}), \|\cdot\|_{\mathcal{H}}) = L^p(\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mu_{\mathcal{X}})$ *with* $1 < p < \infty$*, so that* $\|\cdot\|_{\mathcal{H}}$ *satisfies* $\|h\|_{\mathcal{H}} = \left(\int_{\mathcal{X}} |h|^p d\mu\right)^{1/p} \forall h \in \mathcal{H}(\mathcal{X})$ *with* $1 < p < \infty$*. Furthermore,* $\mathcal{H}_\Theta(\mathcal{X})$ *is a finite dimensional subspace of* $\mathcal{H}(\mathcal{X})$*.*

[19] In particular, while simpler results are available for norms, the limiting criterion $Q^{\mathcal{D}}_\infty$ that induces a metric on $\mathcal{D}^{\mathbf{w}}$ must also be homogeneous and translation invariant to establish a norm on the vector space.

Assumption 13. $(\mathcal{H}(\mathcal{X}), \|\cdot\|_{\mathcal{H}})$ *is a uniformly convex Banach space of power type* p *and* $\mathcal{H}_\Theta(\mathcal{X})$ *is a closed convex subspace of* $\mathcal{H}(\mathcal{X})$.

Finally, we derive the uniqueness of $\boldsymbol{\theta}_0$ from the properties of the limiting criterion function $Q_\infty^{\mathcal{H}}$ and the space of parametric functions $\mathcal{H}_\Theta(\mathcal{X})$ implied by both the parameterization mapping $h_{\mathcal{X}}$ and the parameter space Θ. Theorem 1 below addresses uniqueness in the context of metric-inducing limiting criteria $Q_\infty^{\mathcal{H}}$.

Theorem 1. (Uniqueness for Metric Limit Criteria) *Let Assumptions 5, 6, 7 and 9 hold. Then* $Q_\infty : \Theta \to \mathbb{R}$ *attains a unique minimum at* $\boldsymbol{\theta}_0$.

Proof. See Appendix A1. □

Now, in light of Lemma 3, the a.s. convergence of $\hat{\boldsymbol{\theta}}_T$ to $\boldsymbol{\theta}_0$ follows immediately as corollary under the added influence of Assumptions 1 and 2.

Corollary 1. *Let Assumptions 1, 2, 5, 6, 7 and 9 hold. Define* $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ *such that* $\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(y_{\mathbf{T}}, \mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta})$. *Then* $\hat{\boldsymbol{\theta}}_T \overset{a.s.}{\to} \boldsymbol{\theta}_0$ *as* $T \to \infty$.[20]

Accordingly, Theorem 2 below addresses the uniqueness of $\boldsymbol{\theta}_0$ in the context of norm-inducing limiting criteria $Q_\infty^{\mathcal{H}}$.

Theorem 2. (Uniqueness for Norm Limit Criteria) *Let Assumptions 5, 6, 8 and 10 hold. Then* $Q_\infty : \Theta \to \mathbb{R}$ *attains a unique minimum at* $\boldsymbol{\theta}_0$.

Proof. See Appendix A2. □

Again, $\hat{\boldsymbol{\theta}}_T \overset{a.s.}{\to} \boldsymbol{\theta}_0$ follows immediately as a corollary when Assumptions 1 and 2 also hold.

[20] It is well known that standard consistency proofs apply also to approximate extremum estimators, thus eliminating the need to impose the existence conditions postulated in Assumption 1 and substituting it by more general conditions for the existence of measurable approximate minimizers of the criterion function of interest (see e.g. Brown and Purves (1973)).

Corollary 2. *Let Assumptions 1, 2, 5, 6, 8 and 10 hold. Define* $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ *such that* $\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(y_{\mathbf{T}}, \mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta})$. *Then* $\hat{\boldsymbol{\theta}}_T \overset{a.s.}{\to} \boldsymbol{\theta}_0$ *as* $T \to \infty$.

When the assumptions of Lemma 3 are too restrictive, it is possible to work with those of Lemma 2 instead by verifying that identifiable uniqueness follows essentially from the stricter conditions of Assumptions 6 and 8, plus either 11, 12 or 13. In particular, it is possible to relax the assumptions of compactness of Θ and continuity of $Q_\infty : \Theta \to \mathbb{R}$. This however, is not to be done without the further qualification stated in Assumption 14 below.

Assumption 14. $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ *is an open map.*[21]

The following theorem establishes the relation between the concepts of strong unicity found in the previous section and that of identifiable uniqueness used in Lemma 2.

Theorem 3. (Strong Unicity Implies Identifiable Uniqueness) *Let Assumptions 5, 6, 8 and 14 be satisfied. Then* $Q_\infty : \Theta \to \mathbb{R}$ *has an identifiably unique minimizer* $\boldsymbol{\theta}_0$ *if either Assumption 11, 12 or 13 hold.*

Proof. See Appendix A3. □

This time $\hat{\boldsymbol{\theta}}_T \overset{a.s.}{\to} \boldsymbol{\theta}_0$ follows immediately as corollary of Theorem 3 and Lemma 2.

Corollary 3. *Let Assumptions 2, 5, 6, 8 and 14 be satisfied. Define* $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ *such that* $\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(y_{\mathbf{T}}, \mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta})$. *Then* $\hat{\boldsymbol{\theta}}_T \overset{a.s.}{\to} \boldsymbol{\theta}_0$ *as* $T \to \infty$ *if either Assumption 11, 12 or 13 hold.*

Finally, we use a number of simple examples that illustrate how to verify that the conditions for uniqueness and identifiable uniqueness postulated in Assumptions 3 and 4 hold.

[21] A sufficient condition for the openess of $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ is that its inverse $h_{\mathcal{X}}^{-1} : \mathcal{H}_\Theta(\mathcal{X}) \to \Theta$ be continuous in $h \in \mathcal{H}_\Theta(\mathcal{X})$. Also, note that *(i)* the existence of the inverse function $h_{\mathcal{X}}^{-1}$ is assured by the bijectiveness of $h_{\mathcal{X}}$, and that *(ii)* in the special case where $h_{\mathcal{X}}$ is also continuous, then $h_{\mathcal{X}}$ is an homeomorphism.

7 Some Examples

In the previous sections of this paper we obtained the desired results essentially by decomposing the mapping of elements from Θ to $\mathbb{R}$, compounded in the limiting objective function $Q_\infty : \Theta \to \mathbb{R}$, into three sub-mappings that are easier to handle. We thus obtained a more transparent account of the structure of the extremum estimation problem in nonlinear regression models. The three sub-maps are: *(i)* the so-called *parameterization mapping* $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$, *(ii)* the *probability measure map* $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$, and finally, *(iii)* the divergence criterion function $Q_\infty^{\mathcal{D}} : \mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$.

Simple conditions on each of these sub-maps, as well as the sets Θ, $\mathcal{H}_\Theta(\mathcal{X})$ and $\mathcal{D}_\Theta^{\hat{\mathbf{w}}}$, were shown to ensure the identifiable uniqueness of $\boldsymbol{\theta}_0$. We now review very briefly simple examples of regression models and extremum estimators satisfying the above mentioned properties. The purpose of this section is only that of clarifying the nature of the Assumptions 6 to 14. To remain short and concise, we discuss only a few cases for which verification is straightforward. The interesting cases are likely to be those requiring a more intricate argument. These however are left to be found by researchers having specific applications in mind.

7.1 The Parameterization Mapping: Illustrative Regression Models

Several immediate examples of regression models can be devised for which the injective and open properties of the parameterization mapping $h_{\mathcal{X}}$ hold (Assumptions 6 and 14) and where properties such as compactness, convexity, closedness, finite dimensionality and Haar characterization of $\mathcal{H}_\Theta(\mathcal{X})$ (in Assumptions 9, 10, 11, 12 or 13) are trivially satisfied. As we shall see, it is generally easy to derive the properties of $\mathcal{H}_\Theta(\mathcal{X})$ from those of Θ, whose qualities are defined by the researcher in any given application.

Note first that *the bijective nature of* $h_{\mathcal{X}}$ (implied by Assumption 6) is generally easily derived in this simple regression framework. This is true for instance in models involving polynomial, exponential, logarithmic, trignometric or power functions, that satisfy simple

regularity conditions. Note for example that for regression functions that are analytic on the domain of interest, i.e. $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) \equiv C^\omega_\Theta(\mathcal{X})$, the bijective nature of $h_{\mathcal{X}}$ follows immediately from the fact that each element of $C^\omega_\Theta(\mathcal{X})$ has a power-series representation. The uniqueness of this representation, and hence the bijective nature of $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$, then follows immediately from the uniqueness of power series.[22]

The finite dimensionality of $\mathcal{H}_\Theta(\mathcal{X})$ (stated in Assumptions 10 and 12) is implied by the finite dimensionality of Θ (which holds in several applications) given the identification of $\mathcal{H}_\Theta(\mathcal{X})$ with Θ (a consequence of $h_{\mathcal{X}}$ being bijective). This is true e.g. for the case of polynomial regressions $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) \equiv \mathcal{P}^k_\Theta$, $k \in \mathbb{N}$.

The compactness of $\mathcal{H}_\Theta(\mathcal{X})$ (Assumptions 9 and 10) is easily obtained, for instance, under the continuity of $h_{\mathcal{X}}$ and the compactness of Θ. Here note that, for example, given a regression model of the form $h(x_t; \theta_1, \theta_2, \theta_3) = \theta_1 + \theta_2 \exp(-\theta_3 x_t)$, the continuity of $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ holds for a large class of metric or norm functions with which Θ and $\mathcal{H}_\Theta(\mathcal{X})$ are possibly equipped, and it is immediately satisfied for polynomial regression functions $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{P}^k_\Theta$ regardless of the metric or norm defined on these spaces.

The convexity of $\mathcal{H}_\Theta(\mathcal{X})$ (used in Assumption 9, 10 and 13) can be easily obtained from the convexity of Θ for a large class of parameterization mappings. For example, in the case of a polynomial regression of order k, when $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) \equiv \mathcal{P}^k_\Theta$, we have that, for every $(h_{\mathcal{X}}(\boldsymbol{\theta}_1), h_{\mathcal{X}}(\boldsymbol{\theta}_2)) \in \mathcal{H}_\Theta(\mathcal{X}) \times \mathcal{H}_\Theta(\mathcal{X})$ and every $\tau \in [0,1]$, the function $(\tau h_{\mathcal{X}}(\boldsymbol{\theta}_1) + (1-\tau) h_{\mathcal{X}}(\boldsymbol{\theta}_2))$ belongs to $\mathcal{H}_\Theta(\mathcal{X})$ and takes the form $h_{\mathcal{X}}(\boldsymbol{\theta}_3)$ with $\boldsymbol{\theta}_3 = \tau\boldsymbol{\theta}_1 + (1-\tau)\boldsymbol{\theta}_2$.

The closedness of $\mathcal{H}_\Theta(\mathcal{X})$ (used in Assumption 13) can be easily obtained, for instance, under the closedness of Θ and the continuity of $h^{-1}_{\mathcal{X}} : \mathcal{H}(\mathcal{X}) \to \Theta$.[23] The continuity of the inverse parameterization mapping $h^{-1}_{\mathcal{X}}$ is easily obtained for a large class of regression models.

[22] In multi-index notation (see e.g. Krantz and Parks (1992, p.25)), let $h(\mathbf{x}_t; \boldsymbol{\theta}) = \sum_{|\boldsymbol{\alpha}|=0}^{\infty} \boldsymbol{\theta}_{\boldsymbol{\alpha}} \mathbf{x}_t^{\boldsymbol{\alpha}}$ $\forall\, \mathbf{x}_t \in \mathcal{X}$ and $h(\mathbf{x}_t; \boldsymbol{\theta}') = \sum_{|\boldsymbol{\alpha}|=0}^{\infty} \boldsymbol{\theta}'_{\boldsymbol{\alpha}} \mathbf{x}_t^{\boldsymbol{\alpha}}$ $\forall\, \mathbf{x}_t \in \mathcal{X}$. Then, $h(\mathbf{x}_t; \boldsymbol{\theta}) = h(\mathbf{x}_t; \boldsymbol{\theta}')$ $\forall\, \mathbf{x}_t \in \mathcal{X}$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

[23] Existence of $h^{-1}_{\mathcal{X}}$ is assured by the bijective nature of $h_{\mathcal{X}}$. A bijective map is closed if and only if it is open. The inverse of a continuous map is open.

It holds, for example, on regressions models based on power functions $h(x_t;\theta_1,\theta_2)=\theta_1 x_t^{\theta_2}$ under a large class of norms on Θ and $\mathcal{H}_\Theta(\mathcal{X})$. Once more, it also holds for polynomial regressions under arbitrary norms.

The openess of $h_\mathcal{X}:\Theta\to\mathcal{H}_\Theta(\mathcal{X})$ (postulated in Assumption 14) is also implied by the continuity of the inverse map $h_\mathcal{X}^{-1}$. Hence, the previous argument holds as well.[24]

The Haar characterization of $\mathcal{H}_\Theta(\mathcal{X})$ (Assumption 11) has been obtained for large classes of functions. For example, $h_\mathcal{X}(\boldsymbol{\theta})\in\mathcal{P}_\Theta^k(\mathcal{X})$ satisfies trivially the Haar condition.[25]

7.2 The Limit Divergence Criterion: Illustrative Estimators

We now observe how the properties of the divergence map $Q_\infty^\mathcal{H}:\mathcal{H}(\mathcal{X})\times\mathcal{H}(\mathcal{X})\to\mathbb{R}$ implicitly defined in Assumptions 7, 8 and 9-13 are directly obtained from those of the estimation procedure employed. In particular, we discuss the verification of the simplifying assumption that $g\circ Q_\infty^\mathcal{H}$ (g strictly increasing) be a metric or norm, and that it be either, strongly convex, strictly convex, uniformly convex, of the L^p type ($p<\infty$), or the supremum norm.

The existence of a metric/norm $g\circ Q_\infty^\mathcal{H}$ on $\mathcal{H}(\mathcal{X})$ (established in Assumptions 7 and 8 and used in Assumptions 9-13) is immediate for the class of minimum distance estimators (e.g. the minimum Hellinger distance estimator), since by definition, these estimators are such that $Q_\infty^\mathcal{D}$ takes the form of a distance on $\mathcal{D}^\mathbf{w}\times\mathcal{D}^\mathbf{w}$. As observed in Section 4, a metric or norm is then induced on $\mathcal{H}(\mathcal{X})$ by the bijective mapping D. For many other estimators, in particular those that are not directly obtained as distance minimizers, it is often easy to find a strictly increasing function g such that $g\circ Q_\infty^\mathcal{H}$ defines a distance on $\mathcal{H}(\mathcal{X})\times\mathcal{H}(\mathcal{X})$. For example, it is well known that under appropriate regularity conditions, the least squares

[24] An obvious sufficient condition is that $h_\mathcal{X}$ be a homeomorphism, i.e. that $h_\mathcal{X}$ be bijective, continuous with continuous $h_\mathcal{X}^{-1}$. Note that the homeomorphic nature of $h_\mathcal{X}$ can be obtained by letting (Θ,d_Θ^*) be a metric space with d_Θ^* induced by $h_\mathcal{X}^{-1}$ so that $h_\mathcal{X}$ is automatically isometric and also an isometric isomorphism.

[25] Power monomials satisfy the Haar condition. The system $\{1,\mathbf{x}_t,...,\mathbf{x}_t^\mathbf{k}\}$ has non-vanishing Vandermonde's determinant $VD[a_1,...,a_k]\neq 0$ $(a_1,...,a_k)\in\times_{i=1}^k\mathbb{R}_i^{n_x}$ and hence it satisfies the Haar condition.

estimator,

$$\hat{\boldsymbol{\theta}}_T^{LS} := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(y_{\mathbf{T}}, \mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta}) := \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{T}\sum_{t=1}^{T} e_t^2$$

with $\sum_{t=1}^{T} e_t^2 \equiv \sum_{t=1}^{T} \left[y_t - \hat{y}_t\right]^2 \equiv \sum_{t=1}^{T} \left[h_0(\mathbf{x}_t) - h(\mathbf{x}_t;\boldsymbol{\theta})\right]^2$ is such that,

$$\boldsymbol{\theta}_0^{LS} := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta}) := \arg\min_{\boldsymbol{\theta}\in\Theta} \int_{\mathcal{X}} \mathrm{D}_0^{\mathbf{x}}(\mathbf{x}_t)\left[h_0(\mathbf{x}_t) - h(\mathbf{x}_t;\boldsymbol{\theta})\right]^2 d\mathbf{x}_t,$$

see e.g. White (1980b). This implies that $\boldsymbol{\theta}_0^{LS} = \arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$ where $d_{\mathcal{H}}$ is a divergence.[26] Immediately, taking $g \circ d_{\mathcal{H}}(h_1, h_2) = \sqrt{d_{\mathcal{H}}(h_1, h_2)} \equiv \| h_1 - h_2 \|_{\mathcal{H}}$ for every $(h_1, h_2) \in \mathcal{H}(\mathcal{X})$ implies that $\| \cdot \|_{\mathcal{H}}$ is the well known L^2 norm where $\| h(\mathbf{x}) \|_{\mathcal{H}} = \left(\int_{\mathcal{X}} |h|^2 d_{\mathbf{x}}\right)^{1/2}$. Hence, Assumption 8 holds (and 7 as well by the induced metric) and $\boldsymbol{\theta}_0$ can be described as minimizer of $\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}$ on $(\mathcal{H}(\mathcal{X}), \| \cdot \|_{\mathcal{H}}) \equiv (\mathcal{H}(\mathcal{X}), L^2)$, i.e.,

$$\boldsymbol{\theta}_0^{LS} = \arg\min_{\boldsymbol{\theta}\in\Theta} \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left(\int_{\mathcal{X}} \mathrm{D}_0^{\mathbf{x}}(\mathbf{x}_t)\left[h_0(\mathbf{x}_t) - h(\mathbf{x}_t;\boldsymbol{\theta})\right]^2 d\mathbf{x}_t \right)^{1/2}.$$

The strict convexity of $d_{\mathcal{H}}^*$ *or* $\| \cdot \|_{\mathcal{H}} \equiv g \circ Q_\infty^{\mathcal{H}}$ *on* $\mathcal{H}_\Theta(\mathcal{X})$ (Assumption 10) is generally easy to verify and it holds e.g. for the minimum Hellinger distance and least squares estimators just mentioned above (see e.g. Donoho and Liu (1988) and Powell (1981) respectively). Note also that in this case the *strong convexity of* $d_{\mathcal{H}}^* \equiv g \circ Q_\infty^{\mathcal{H}}$ (used in Assumption 9) is immediately obtained since the later is by construction implied by the former (see Cheney (1974), Ahuja et al. (1977) and Narang (1981)). This is also true of *uniform convexity of power type p of* $\| \cdot \|_{\mathcal{H}} \equiv g \circ Q_\infty^{\mathcal{H}}$ (Assumption 13) and L^p *representation of* $\| \cdot \|_{\mathcal{H}} \equiv g \circ Q_\infty^{\mathcal{H}}$ (Assumption 12) in the case of least squares estimation (see Cheney (1974) or Cheney (1982, p.23)).

The supremum representation of $\| \cdot \|_{\mathcal{H}} \equiv g \circ Q_\infty^{\mathcal{H}}$ (Assumption 11) is considerably more restrictive (as mentioned before) and holds for minimax estimators.

[26] The least squares divergence, $d_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$ satisfies non-negativity $d_{\mathcal{H}}(h_1, h_2) \geq 0 \; \forall (h_1, h_2) \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$ and identity of indiscernibles $d_{\mathcal{H}}(h_1, h_2) = 0$ if and only if $h_1 = h_2 \; \forall (h_1, h_2) \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$, but not symmetry or sub-additivity.

8 Final Remarks

In this paper we have illustrated the possibility of using results from Approximation Theory to verify the assumption of identifiable uniqueness commonly used to obtain consistency of extremum estimators. We made use only of simple intuitive results on the (strong) uniqueness of best approximations. Clearly, much more can be done in extending these results to a larger class of extremum estimators and regression models. Here, generality was sacrificed in favor of conciseness and simplicity, but it should be kept in mind that, in this context, we could be as general as Approximation Theory allows us to be. In particular, these results extend immediately to various models outside the regression framework and the notion of distance function can be easily weakened to include non-metric divergences.

A Proofs

A.1 Theorem 1

Proof. Assumption 5 implies that $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta}) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty^{\mathcal{D}}(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ and according to Assumption 7,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{D}}^*(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$$

where $d_{\mathcal{D}}^* : \mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$ is a metric defined on $\mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}}$ as $d_{\mathcal{D}}^* \equiv g \circ Q_{\mathcal{D}}$ with $g : \mathbb{R} \to \mathbb{R}_0^+$ a strictly increasing function. Now given Assumption 6, we have $d_{\mathcal{D}}^*(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}) \equiv d_{\mathcal{H}}^*(h_0, h(\cdot, \boldsymbol{\theta}))\ \forall\, \boldsymbol{\theta} \in \Theta$ by construction since $d_{\mathcal{H}}^* : \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X}) \to \mathbb{R}_0^+$ is a metric defined on $\mathcal{H}(\mathcal{X})$ according to $d_{\mathcal{H}}^*(h, h') \equiv d_{\mathcal{D}}^*(\mathrm{D}(h), \mathrm{D}(h'))\ \forall\, (h, h') \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$ and hence $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{H}}^*(h_0, h(\cdot, \boldsymbol{\theta}))$ holds true. Finally, according to Lemmas 4, 5 and 6, Assumption 9 implies that for every $h_0 \in \mathcal{H}(\mathcal{X})$, there exists a unique $h \in \mathcal{H}_\Theta(\mathcal{X})$ satisfying,

$$d_{\mathcal{H}}^*(h_0, h) \leq d_{\mathcal{H}}^*(h_0, h')\ \forall\, h' \in \mathcal{H}_\Theta(\mathcal{H}).$$

Given the bijective nature of the parameterization mapping $h_{\mathcal{X}} : \Theta \to \mathcal{H}_{\Theta}(\mathcal{X})$ postulated in Assumption 6, it follows that there exists a unique $\boldsymbol{\theta} \in \Theta$ satisfying,

$$d^*_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta})) \leq d^*_{\mathcal{H}}(h_0, h(\cdot; \boldsymbol{\theta}'))\ \forall\, \boldsymbol{\theta}' \in \Theta;$$

i.e. $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \Theta} d^*_{\mathcal{H}}(h_0, h(\cdot, \boldsymbol{\theta})) \equiv \arg\min_{\boldsymbol{\theta} \in \Theta} Q^{\mathcal{D}}_{\infty}(\mathrm{D}^{\mathbf{w}}_0, \mathrm{D}^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})$ is unique. □

A.2 Theorem 2

Proof. The argument follows essentially that of Theorem 1. Assumption 5 ensures that attention is restricted to the class of extremum estimators satisfying $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \Theta} Q_{\infty}(\boldsymbol{\theta}) \equiv \arg\min_{\boldsymbol{\theta} \in \Theta} Q^{\mathcal{D}}_{\infty}(\mathrm{D}^{\mathbf{w}}_0, \mathrm{D}^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})$ and by Assumption 8,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \Theta} \parallel \mathrm{D}^{\mathbf{w}}_0 - \mathrm{D}^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}} \parallel_{\mathcal{D}}$$

where $\parallel \cdot \parallel_{\mathcal{D}}: \mathcal{D}^{\mathbf{w}} \to \mathbb{R}^+_0$ is a norm defined on $\mathcal{D}^{\mathbf{w}}$ as $\parallel \cdot \parallel_{\mathcal{D}} \equiv g \circ Q^{\mathcal{D}}_{\infty}$ with $g : \mathbb{R} \to \mathbb{R}^+_0$ a strictly increasing transformation. Now given Assumption 6, we have $\parallel \mathrm{D}^{\mathbf{w}}_0 - \mathrm{D}^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}} \parallel_{\mathcal{D}} \equiv \parallel h_0 - h(\cdot, \boldsymbol{\theta}) \parallel_{\mathcal{H}}$ $\forall\, \boldsymbol{\theta} \in \Theta$ by construction, since $\parallel \cdot \parallel_{\mathcal{H}}: \mathcal{H}(\mathcal{X}) \to \mathbb{R}^+_0$ is a norm defined on $\mathcal{H}(\mathcal{X})$ according to $\parallel h \parallel_{\mathcal{H}} = \parallel \mathrm{D}(h) \parallel_{\mathcal{D}}\ \ \forall h \in \mathcal{H}(\mathcal{X})$, and hence $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \Theta} \parallel h_0 - h(\cdot, \boldsymbol{\theta}) \parallel_{\mathcal{H}}$ holds true. Finally, according to Lemmas 7 and 8, Assumption 10 implies that for every $h_0 \in \mathcal{H}(\mathcal{X})$, there exists a unique $h \in \mathcal{H}_{\Theta}(\mathcal{X})$ satisfying,

$$\parallel h_0 - h \parallel_{\mathcal{H}} \leq \parallel h_0 - h' \parallel_{\mathcal{H}}\ \ \forall h' \in \mathcal{H}_{\Theta}(\mathcal{H}).$$

Given the bijective nature of the parameterization mapping $h_{\mathcal{X}} : \Theta \to \mathcal{H}_{\Theta}(\mathcal{X})$ postulated in Assumption 6, it follows that there exists a unique $\boldsymbol{\theta} \in \Theta$ satisfying,

$$\parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_{\mathcal{H}} \leq \parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}') \parallel_{\mathcal{H}}\ \ \forall\, \boldsymbol{\theta}' \in \Theta;$$

i.e. $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \Theta} \parallel h_0 - h(\cdot, \boldsymbol{\theta}) \parallel_{\mathcal{H}} \equiv \arg\min_{\boldsymbol{\theta} \in \Theta} Q^{\mathcal{D}}_{\infty}(\mathrm{D}^{\mathbf{w}}_0, \mathrm{D}^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})$ is unique. □

A.3 Theorem 3

Proof. In what follows, we first take some initial steps that are similar to those of Theorems 1 and 2 and then specialize the discussion to the cases of *(i)* strong unicity obtained under Assumption 11, and *(ii)* strong unicity of order α obtained under either Assumption 12 or 13. As before, Assumption 5 guarantees the formulation,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty^{\mathcal{D}}(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}).$$

Furthermore, according to Assumption 8, $\boldsymbol{\theta}_0$ also satisfies $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \parallel \mathrm{D}_0^{\mathbf{w}} - \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}} \parallel_{\mathcal{D}}$ where $\| \cdot \|_{\mathcal{D}}$: $\mathcal{D}^{\mathbf{w}} \rightarrow \mathbb{R}_0^+$ is a norm defined on $\mathcal{D}^{\mathbf{w}}$ as $\| \cdot \|_{\mathcal{D}} \equiv g \circ Q_\infty^{\mathcal{D}}$ with $g : \mathbb{R} \rightarrow \mathbb{R}_0^+$ a strictly increasing function. Now given Assumption 6, we have $\| \mathrm{D}_0^{\mathbf{w}} - \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}} \|_{\mathcal{D}} \equiv \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_{\mathcal{H}}$ by construction since $\| \cdot \|_{\mathcal{H}}$: $\mathcal{H}(\mathcal{X}) \rightarrow \mathbb{R}_0^+$ is a norm defined on $\mathcal{H}(\mathcal{X})$ according to $\| h \|_{\mathcal{H}} \equiv \| \mathrm{D}(h) \|_{\mathcal{D}} \ \forall h \in \mathcal{H}(\mathcal{X})$ and hence $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_{\mathcal{H}}$ holds true. Finally, we split this proof into three parts and obtain the desired identifiable uniqueness of $\boldsymbol{\theta}_0$, under either Assumption 11, 12 or 13 respectively.

Part I. Let Assumption 11 hold. Then,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_\infty .$$

Furthermore, for h_0 and $h_{\mathcal{X}}(\boldsymbol{\theta})$ satisfying the conditions of Assumption 11 we have by Lemma 10 that for every $h_0 \in \mathcal{H}(\mathcal{X})$, there exists a unique $h \in \mathcal{H}_\Theta(\mathcal{X})$ satisfying the strong unicity property, $\| h_0 - h' \|_\infty > \| h_0 - h \|_\infty + \gamma \| h - h' \|_\infty \ \forall h' \in \mathcal{H}_\Theta(\mathcal{X})$, with $\gamma > 0$, thus conveniently restated as,

$$\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty > \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_\infty + \gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty \ \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{H})$$

since every element $h \in \mathcal{H}_\Theta(\mathcal{X})$ has a parametric representation of the form $h_{\mathcal{X}}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$. Now, clearly, $\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty > \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_\infty + \gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty \quad \Leftrightarrow$

$\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_\infty > \gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty \ \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$, and

hence, $\inf_{\boldsymbol{\theta}\in\Theta^*}[\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_\infty] \geq \inf_{\boldsymbol{\theta}\in\Theta^*}[\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty]$ holds for any $\Theta^* \subseteq \Theta$. We now show that when $\Theta^* = \eta_0(\epsilon)^c$, then,

$$\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_\infty] \geq \inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty] > 0.$$

Indeed, note first that,

$$\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty] = \inf_{h_{\mathcal{X}}(\boldsymbol{\theta})\in h_{\mathcal{X}}(\eta_0(\epsilon)^c)}[\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty]$$

and hence that it is enough to show that,

$$\inf_{h_{\mathcal{X}}(\boldsymbol{\theta})\in h_{\mathcal{X}}(\eta_0(\epsilon)^c)}[\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty] > 0.$$

It is elementary that for every $\gamma > 0$, having,

$$\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty > c > 0 \, \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in h_{\mathcal{X}}(\eta_0(\epsilon)^c)$$

for some $c > 0$ independent of $\boldsymbol{\theta}$, implies $\inf_{h_{\mathcal{X}}(\boldsymbol{\theta})\in h_{\mathcal{X}}(\eta_0(\epsilon)^c)} \gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty > 0$, and that, $\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty > c > 0 \, \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in h_{\mathcal{X}}(\eta_0(\epsilon)^c)$ holds true whenever $h_{\mathcal{X}}(\eta_0(\epsilon))$ is an open set with $h_{\mathcal{X}}(\boldsymbol{\theta}_0) \in h_{\mathcal{X}}(\eta_0(\epsilon))$, because then, $\exists \, \delta > 0$ such that $S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)$ is an open ball of radius δ centered at $h_{\mathcal{X}}(\boldsymbol{\theta}_0)$ satisfying $S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta) \subseteq h_{\mathcal{X}}(\eta_0(\epsilon))$, and hence, by definition, $\| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty \geq \delta > 0 \, \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)^c$ where $S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)^c := \mathcal{H}_\Theta(\mathcal{X}) \backslash S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)$. This implies that, for every $\gamma > 0$,

$$\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty > c > 0 \, \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)^c$$

holds uniformly in $\boldsymbol{\theta} \in \Theta$ for every $0 < c < \delta/\gamma$. Thus, the desired result,

$$\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_\infty] > 0$$

is implied by Assumption 14 which ensures the openess of $h_{\mathcal{X}}(\boldsymbol{\theta}_0) \in h_{\mathcal{X}}(\eta_0(\epsilon))$ and hence that $\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty] > 0$. Finally, since $Q_\infty(\boldsymbol{\theta}) \equiv Q_\infty^{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$ satisfies,

$$g \circ Q_\infty^{\mathcal{H}}(h_0, h(\cdot;\boldsymbol{\theta})) \equiv \| h_0 - h(\cdot;\boldsymbol{\theta}) \|_{\mathcal{H}} \equiv \| h_0 - h(\cdot;\boldsymbol{\theta}) \|_\infty \ \forall \boldsymbol{\theta} \in \Theta$$

with strictly increasing g, it follows that,

$$\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\| h_0 - h(\cdot;\boldsymbol{\theta}) \|_\infty - \| h_0 - h(\cdot;\boldsymbol{\theta_0}) \|_\infty] > 0 \Leftrightarrow \inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[Q_\infty(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}_0)] > 0.$$

We thus conclude that strong unicity implies identifiable uniqueness under Assumptions 6, 8, 11 and 14, i.e. that $\| h_0 - h_\mathcal{X}(\boldsymbol{\theta}) \|_\infty > \| h_0 - h_\mathcal{X}(\boldsymbol{\theta}_0) \|_\infty + \gamma \| h - h_\mathcal{X}(\boldsymbol{\theta}) \|_\infty \ \forall \boldsymbol{\theta} \in \Theta \Rightarrow \inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[Q_\infty(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}_0)] > 0 \, \forall \boldsymbol{\theta} \in \Theta$.

Part II. Let Assumption 12 hold instead of 11. Then, except for some trivial minor details, the same argument holds. In particular, we now have,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \| h_0 - h(\cdot,\boldsymbol{\theta}) \|_\mathcal{H}$$

where $\| \cdot \|_\mathcal{H}$ satisfies,

$$\| h \|_\mathcal{H} = \Big(\int_\mathcal{X} |h|^p d\mu \Big)^{1/p} \ \forall h \in \mathcal{H}(\mathcal{X})$$

with $1 < p < \infty$. Since $h_0 \in L^p(\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mu_\mathcal{X})$ with $1 < p < \infty$ and $h_\mathcal{X}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$ where $\mathcal{H}_\Theta(\mathcal{X})$ is a finite dimensional subset of $\mathcal{H}(\mathcal{X})$, we have by Lemma 11 that for every $h_0 \in \mathcal{H}(\mathcal{X})$, when there exists a unique best approximation $h \in \mathcal{H}_\Theta(\mathcal{X})$ to $h_0 \in \mathcal{H}(\mathcal{X})$ then it is strongly unique of order $\alpha = \max\{p, 2\}$. In other words, $\exists\, \gamma(h_0) > 0 : \| h_0 - h' \|_\mathcal{H} > \| h_0 - h \|_\mathcal{H} + \gamma \| h - h' \|_\mathcal{H}^\alpha \ \forall h' \in \mathcal{H}_\Theta(\mathcal{X})$. This property is conveniently restated as

$$\| h_0 - h_\mathcal{X}(\boldsymbol{\theta}) \|_\mathcal{H} > \| h_0 - h_\mathcal{X}(\boldsymbol{\theta}_0) \|_\mathcal{H} + \gamma \| h - h_\mathcal{X}(\boldsymbol{\theta}) \|_\mathcal{H}^\alpha \ \forall h_\mathcal{X}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$$

since every element $h \in \mathcal{H}_\Theta(\mathcal{X})$ has a parametric representation of the form $h_\mathcal{X}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$. The existence of an element of best approximation follows from Lemma 7 by noting that every uniformly convex normed vector space is strictly convex (Cheney (1982, p.23)). As before, the elementary step $h_0 - h_\mathcal{X}(\boldsymbol{\theta}) \|_\mathcal{H} > \| h_0 - h_\mathcal{X}(\boldsymbol{\theta}_0) \|_\mathcal{H} + \gamma \| h_\mathcal{X}(\boldsymbol{\theta}_0) - h_\mathcal{X}(\boldsymbol{\theta}) \|_\mathcal{H}^\alpha \quad \Leftrightarrow$
$\| h_0 - h_\mathcal{X}(\boldsymbol{\theta}) \|_\mathcal{H} - \| h_0 - h_\mathcal{X}(\boldsymbol{\theta}_0) \|_\mathcal{H} > \gamma \| h_\mathcal{X}(\boldsymbol{\theta}_0) - h_\mathcal{X}(\boldsymbol{\theta}) \|_\mathcal{H}^\alpha \ \forall h_\mathcal{X}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$ implies,

$$\inf_{\boldsymbol{\theta}\in\Theta^*}[\| h_0 - h_\mathcal{X}(\boldsymbol{\theta}) \|_\mathcal{H} - \| h_0 - h_\mathcal{X}(\boldsymbol{\theta}_0) \|_\mathcal{H}] \geq \inf_{\boldsymbol{\theta}\in\Theta^*}[\gamma \| h_\mathcal{X}(\boldsymbol{\theta}_0) - h_\mathcal{X}(\boldsymbol{\theta}) \|_\mathcal{H}^\alpha]$$

holds for any $\Theta^* \subseteq \Theta$. Again, we are interested in the case $\Theta^* = \eta_0(\epsilon)^c$, and to obtain

$$\inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_{\mathcal{H}}] > 0$$

it is enough to show that $\inf_{h_{\mathcal{X}}(\boldsymbol{\theta}) \in h_{\mathcal{X}}(\eta_0(\epsilon)^c)} [\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^{\alpha}] > 0$. Since for every $\gamma > 0$ and $\alpha > 1$, having,

$$\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^{\alpha} > c > 0 \, \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in h_{\mathcal{X}}(\eta_0(\epsilon)^c)$$

for some $c > 0$ constant, implies,$\inf_{h_{\mathcal{X}}(\boldsymbol{\theta}) \in h_{\mathcal{X}}(\eta_0(\epsilon)^c)} \gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^{\alpha} > 0$, and that, $\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^{\alpha} > c > 0 \, \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in h_{\mathcal{X}}(\eta_0(\epsilon)^c)$ holds true if $h_{\mathcal{X}}(\eta_0(\epsilon))$ is an open set satisfying $h_{\mathcal{X}}(\boldsymbol{\theta}_0) \in h_{\mathcal{X}}(\eta_0(\epsilon))$ for every $\epsilon > 0$, by the same argument as before. Thus, for every $\gamma > 0$ it holds true that,

$$\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^{\alpha} > c > 0 \, \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)^c$$

uniformly in $\boldsymbol{\theta} \in \Theta$, for every $0 < c < (\delta/\gamma)^{1/\alpha}$ where $S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)^c := \mathcal{H}_{\Theta}(\mathcal{X}) \backslash S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)$. Hence, $\inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_{\mathcal{H}}] > 0$ is implied by Assumption 14 which ensures the openess of $h_{\mathcal{X}}(\boldsymbol{\theta}_0) \in h_{\mathcal{X}}(\eta_0(\epsilon))$ and hence that $\inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^{\alpha}] > 0$. Finally, since $Q_{\infty}(\boldsymbol{\theta}) \equiv Q_{\infty}^{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$ satisfies,

$$g \circ Q_{\infty}^{\mathcal{H}}(h_0, h(\cdot; \boldsymbol{\theta})) \equiv \| h_0 - h(\cdot; \boldsymbol{\theta}) \|_{\mathcal{H}} \equiv \| h_0 - h(\cdot; \boldsymbol{\theta}) \|_{\mathcal{H}} \ \ \forall \boldsymbol{\theta} \in \Theta$$

with strictly increasing g, it follows that,

$$\inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [\| h_0 - h(\cdot; \boldsymbol{\theta}) \|_{\mathcal{H}} - \| h_0 - h(\cdot; \boldsymbol{\theta_0}) \|_{\mathcal{H}}] > 0 \ \Leftrightarrow \ \inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [Q_{\infty}(\boldsymbol{\theta}) - Q_{\infty}(\boldsymbol{\theta}_0)] > 0.$$

We thus conclude that strong unicity of order α implies identifiable uniqueness under Assumptions 6, 8, 12 and 14, i.e. that $\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} > \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_{\mathcal{H}} + \gamma \| h - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^{\alpha}$ $\forall \boldsymbol{\theta} \in \Theta \Rightarrow \inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [Q_{\infty}(\boldsymbol{\theta}) - Q_{\infty}(\boldsymbol{\theta}_0)] > 0 \, \forall \boldsymbol{\theta} \in \Theta$.

Part III. Finally, let Assumption 13 hold instead of 11 or 12. We now have,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_{\mathcal{H}}$$

where $\| \cdot \|_{\mathcal{H}}$ is such that $(\mathcal{H}(\mathcal{X}), \| \cdot \|_{\mathcal{H}})$ is a uniformly convex Banach space of power type $p > 1$. Since $h_0 \in \mathcal{H}(\mathcal{X})$ and $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$ where $\mathcal{H}_\Theta(\mathcal{X})$ is a closed convex subspace of $\mathcal{H}(\mathcal{X})$, we have by Lemma 11 that for every $h_0 \in \mathcal{H}(\mathcal{X})$, when there exists a unique best approximation $h \in \mathcal{H}_\Theta(\mathcal{X})$ to $h_0 \in \mathcal{H}(\mathcal{X})$, then, it is strongly unique of order p. As we have already seen, this form of strong unicity implies the identifiable uniqueness of $\boldsymbol{\theta}_0$. The existence of an element of best approximation follows from the fact that a closed convex subset of a uniformly convex Banach space is proximinal (Cheney (1982, p.22)).

□

References

Ahuja, G. C., Narang, T. D., and Trehan, S. (1977). Best approximation on convex sets in metric linear spaces. *Mathematische Nachrichten*, 78:125–130.

Amemiya, T. (1983). Non-linear regression models. *Handbook of Econometrics*, 1:333–389.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Andrews, D. W. (1992). Generic uniform convergence. *Econometric Theory*, 8:241–257.

Angelos, J. and Egger, A. (1984). Strong uniqueness in L^p spaces. *Journal of Approximation Theory*, 42:14–26.

Bates, C. and White, H. (1985). A unified theory of consistent estimation for parametric models. *Econometric Theory*, 1:151–178.

Billingsley, P. (1995). *Probability and Measure*. Wiley-Interscience.

Blasques, F. (2010). Semi-nonparametric indirect inference. *mimeo*.

Brown, L. D. and Purves, R. (1973). Measurable selections of extrema. *Annals of Statistics*, 1:902–912.

Burguete, J., Gallant, A. R., and Souza, G. (1982). On unification of the asymptotic theory of nonlinear econometric models. *Econometric Reviews*, 1(2):151–190.

Cheney, E. (1982). *Approximation Theory*. American Mathematical Society Chelsea Publishing, 2nd edition.

Cheney, E. W. (1974). Letter to the editor: Best approximation on convex sets in a metric space. *Journal of Approximation Theory*, 12:94–97.

Clarke, B. (1983). Uniqueness and frechet differentiability of functional solutions to maximum likelihood. *The Annals of Statistics*, 11(4):1196–1205.

Clarkson, J. A. (1936). Uniformly convex spaces. *Transactions of the American Mathematical Society*.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton UNiverstity Press.

Crisp, A. and Burridge, J. (1993). A note on the uniqueness of m-etimators in robust regression. *The Canadian Journal of Statistics*, 21(2):205–208.

Davidson, J. (1994). *Stochastic Limit Theory*. Advanced Texts in Econometrics. Oxford University Press.

Domowitz, I. and White, H. (1982). Misspecified models with dependent observations. *Journal of Econometrics*, 20(1):35–58.

Donoho, D. L. and Liu, R. C. (1988). Pathologies of some minimum distance estimators. *The Annals of Statistics*, 16(2):587–608.

Doob, J. L. (1934). Probability and statistics. *Transactions of the American Mathematical Society*, 36(4):759–775.

Doob, J. L. (1953). *Stochastic Processes*. John Wiley and Sons.

Ducharme, G. R. (1995). Uniqueness of the least-distances estimator in regression models with multivariate response. *The Canadian Journal of Statistics*, 23(4):421–424.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Freedman, D. A. and Diaconis, P. (1982). On inconsistent m-estimators. *The Annals of Statistics*, 10(2):454–461.

Gallant, R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Cambridge University Press.

Hanner, O. (1956). On the uniform convexity of L^p and l^p. *Arkiv för Matematik*, 3(19):239–244.

Hsiao, C. (1983). *Handbook of Econometrics*, volume 1, chapter Identification. North Holland Publishing Company.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

Kabaila, P. (1983). Parameter values of arma models minimising the one-step-ahead prediction error when the true system is not in the model set. *Journal of Applied Probability*, 20(2):405–408.

Kent, J. T. and Tyler, D. E. (2001). Regularity and uniqueness for constrained m-estimates and redescending m-estimates. *The Annals of Statistics*, 29(1):252–265.

Kolmogorov, A. N. and Fomin, S. V. (1975). *Introductory Real Analysis*. Dover Publications.

Krantz, S. and Parks, H. (1992). *A Primer of Real Analytic Functions*. Birkhauser Advanced Texts, second edition edition.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related bayes' estimates. *Universtity of California Publications in Statistics*, 2:23–53.

Lin, P. K. (1989). Strongly unique best approximation in uniformly convex banach spaces. *Journal of Approximation Theory*, 56:101–107.

Malinvaud, E. (1970). The consistency of nonlinear regressions. *The Annals of Mathematical Statistics*, 41(3):956–969.

Narang, T. D. (1981). Best approximation and strict convexity of metric spaces. *Archivum Mathematicum*, 017(2):87–90.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4):1161–67.

Newman, D. J. and Shapiro, H. S. (1963). Some theorems on chebyshev approximation. *Duke Mathematical Journal*, 30:673–681.

Nurberger, G. (1979). Unicity and strong unicity in approximation theory. *Journal of Approximation Theory*, 26:54–70.

Pötscher, B. M. and Prucha, I. R. (1991a). Basic structure of the asymptotic theory in dynamic nonlinear econometric models, part i: consistency and approximation concepts. *Econometric Reviews*, 10(2):125–216.

Pötscher, B. M. and Prucha, I. R. (1991b). Basic structure of the asymptotic theory in dynamic nonlinear econometric models, part ii: Asymptotic normality. *Econometric Reviews*, 10(3):253–325.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models: Asymptotic Theory*. Springer-Verlag.

Powell, M. J. D. (1981). *Approximation Theory and Methods*. Cambridge University Press.

Rivest, L. P. (1989). De l'unicite des estimateurs robustes en regression lorsque le parametre d'echelle et le parametre de la regression sont estimes simultanement. *Canadian Journal of Statistics*, 17(2):141–153.

Romaguera, S. and Sanchis, M. (2000). Semi-lipschitz functions and best approximation in quasi-metric spaces. *Journal of Approximation Theory*, 103:292–301.

Stinchcombe, M. B. and White, H. (1992). Some measurability results for extrema of random functions over random sets. *Review of Economic Studies*, 59(3):495–514.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer-Verlag, New York.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Institute of Mathematical Statistics*.

White, H. (1980a). Nonlinear regression on cross-section data. *Econometrica*, 48(3):721–46.

White, H. (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, 21(1):149–70.

Wulbert, D. E. (1971). Uniqueness and differential characterization of approximations from manifolds of functions. *American Journal of Mathematics*, 18:350–366.