SMOOTH MIN-DIVERGENCE INFERENCE IN SEMI PARAMETRIC MODELS

Michel Broniatowski, Justin Steward Moutsouka

▶ **To cite this version:**

Michel Broniatowski, Justin Steward Moutsouka. SMOOTH MIN-DIVERGENCE INFERENCE IN SEMI PARAMETRIC MODELS. 2020. hal-02586204

HAL Id: hal-02586204

https://hal.science/hal-02586204v1

Preprint submitted on 15 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SMOOTH MIN-DIVERGENCE INFERENCE IN SEMI PARAMETRIC MODELS

MICHEL BRONIATOWSKI(1), JUSTIN STEWARD MOUTSOUKA(1)

ABSTRACT. This paper considers inference in some semi parametric models through some specific class of statistical procedure, which have proved to be of valuable interest in parametric estimation, namely the power divergence family defined by Basu Hodjt, Harris and Jones (1998).
At a first glance their divergence is not fitted to semiparametric inference. However extending the parametric setting to a smoothed semiparametric one, it is possible to make inference both on θ_T and on the density of P_{θ_T} in semiparametric models defined by moment conditions indexed by some parameter θ, where the data are generated under some unknown θ_T. This question is of interest; indeed usually the estimation of the density of P_{θ_T} with respect to a dominating measure (here the Lebesgue measure) is an open chalenge in the realm of semi parametric models. This is the focus of the present paper.

Key words and phrases :Semi parametric models, Inference, Minimum divergence inference

1. INTRODUCTION

This paper considers inference in some semi parametric models through some specific class of statistical procedure, which have proved to be of valuable interest in parametric estimation. The global paradigm which is considered here consists in the minimization of a pseudo distance between the empirical measure defined by the data set and a model, defined loosely as a collection of probability measures which we consider as candidates for the generic distribution of the data set. This framework is generally referred to as a "divergence based approach"; according to the choice of the divergence (or "pseudo distance"), many classical methods for estimation and testing can be recovered. Before entering into our topics in a more detailed way, let us describe rapidly some of the various divergences which have been discussed in the recent past, and present their specificities.

1.1. **Divergences.** A divergence (or discrepancy) between two probability measures P and Q defined on the same measurable space $\mathcal{X}$ equipped with its Borel field $\mathcal{B}(\mathcal{X})$ is a positive mapping

$$(P, Q) \to D(Q, P)$$

such that $D(Q, P) = 0$ if and only if $Q = P$. No symmetry is assumed, nor any triangular inequality; therefore a divergence need not be a distance. Constructions of such functions D are numerous; we briefly sketch two main schemes, each of

which leading to specific fields of applications in statistics and learning. We refer to Broniatowski and Stummer (2019) or Broniatowski and Vajda (2012) for description and further references. In this paper the space $\mathcal{X}$ is the euclidean space $\mathbb{R}^d$, endowed with its Borel field. In the sequel $\mathcal{M}^1$ designates the class of all probability measures defined on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

1.1.1. *Ali Silvey and Csiszar divergences.* A first class of divergences has been introduced by Ali and Silvey (1966), and by Csiszár (1967); for Q absolutely continuous with respect to P

$$D(Q,P) := \int f\left(\frac{dQ}{dP}\right) dP$$

where f is a non negative convex function defined on $\overline{\mathbb{R}^+}$ which satisfies $f(1) = 0$. When the support of Q is not included in the support of P then $D(Q,P) := +\infty$. Typical examples of functions f are $f_0(x) := -\log x + x - 1$, $f_1(x) := x\log x - x + 1$, $f_{1/2}(x) := 2\left(\sqrt{x}-1\right)^2$, $f_2(x) := (1/2)(x-1)^2$, $f_{-1}(x) := (1/2)\frac{(x-1)^2}{x}$.

In the above list, f_0 induces the likelihood divergence (modified Kulback-Leibler divergence) , f_1 induces the Kulback-Leibler divergences, $f_{1/2}$ defines the Hellinger divergence, while f_2 and f_{-1} respectively define the Pearson (resp. the Neyman) Chi-square divergences. Note that by its very definition, given a sample of iid copies $X_1, .., X_n$ of a generic rv X with continuous distribution on $\mathbb{R}^d$ and a model $\mathcal{M}$ of continuous distributions on $\mathbb{R}^d$ it may hold that the projection of P on $\mathcal{M}$ be defined, although the natural proxy of P defined through the empirical measure

$$P_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

is at infinite "distance" from $\mathcal{M}$ for any n. This drawback can be overpassed and leads to general techniques in parametric inference, encompassing the various classical ones associated to the various names cited hereabove; see Broniatowski and Keziou (2009). We will turn to semiparametric inference a bit later.

A convenient class of such functions f can be defined through the so-called Cressie-Read functions, which are indexed by a real valued parameter

$$f_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma-1)},$$

and the examples listed hereabove are indeed indexed along the value of γ (with limit expansions for the case when $\gamma = 0, 1$).

1.1.2. *The BHHJ power divergences.* This class of divergences has been introduced by Basu Hodjt, Harris and Jones Basu and al. (1998), referred to as BHHJ divergence here, and is defined for distributions which are absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^d$. Given P and Q with respective densities p and q the power divergence with real power index α is defined through

$$D_\alpha(Q,P)=\int \varphi(q(x),p(x))\mathrm{d}x \tag{1.1}$$

where

$$\varphi(u,v)=u^{\alpha+1}-\left(1+\frac{1}{\alpha}\right)u^\alpha\times v+\frac{1}{\alpha}v^{\alpha+1}.$$

We will consider only values of α in $(0;1)$.

The developed form of $D_\alpha(Q,P)$ is therefore

$$D_\alpha(Q,P)=\int\left\{q^{\alpha+1}(v)-\left(1+\frac{1}{\alpha}\right)q^\alpha(v)p(v)+\frac{1}{\alpha}p^{\alpha+1}(v)\right\}\mathrm{d}v \tag{1.2}$$

Motivation for using the BHHJ divergence in parametric inference is considered in Basu and al. (1998). This class of divergences is well suited for the parametric estimation; indeed consider a parametric model $\mathcal{M}:=\{P_\theta,\theta\in\Theta\}$ of absolutely continuous measures where Θ is some parameter space; it then holds

$$D_\alpha(P_\theta,P)=R_\alpha(P_\theta,P)+\beta$$

where

$$\beta:=\frac{1}{\alpha}\int p^{\alpha+1}(v)dv$$

is independent upon θ; therefore minimizing $D_\alpha(P_\theta,P)$ on θ amounts to minimize $R_\alpha(P_\theta,P)$; when dealing with estimation, P is supposed to be P_T , the distribution of the generic observation X, and substitution of the unknown measure P_T by the empirical measure P_n yields the corresponding statistical criterion

$$R_\alpha(P_\theta,P_n):=\int p_\theta^{\alpha+1}(v)dv-\frac{1}{n}\left(1+\frac{1}{\alpha}\right)\sum_{i=1}^n p_\theta^\alpha(X_i) \tag{1.3}$$

which can be minimized upon θ and produce an estimator of θ_T whenever $P_T=P_{\theta_T}$. Whenever the integral in the above display does not depend on the parameter θ, as holds for location models, then minimizing upon θ in $R_\alpha(P_\theta,P_n)$ amounts to smooth the usual likelihood score by a factor p_θ^α which damps the role of outliers in the estimating equation.

This procedure has been developed extensively and leads to classical limit results for estimation and testing; see Theorem 2 in Basu and al. (1998). The performance of this approach has been compared to similar treatments making use of Csiszar divergences, both under the model and under misspecification; globally speaking, performances of either Csiszar divergence approach or power divergence approach are quite similar (same limit distribution of the estimator and of the test statistics as for the maximum likelihood approach (which falls in the field of Csiszar divergences but not in the field of power ones for α in $(0,1)$), nearly similar results in simulation runs on small or medium size samples, Tuning of the parameter α or γ allows to obtain reasonably robust estimators under contamination, as measures through the Influence function; see Toma and Broniatowski (2011).

The main properties of BHHJ divergences are:

Fact 1: $D_\alpha(Q,P)$ is a divergence in that it is non-negative for all absolutely continuous probability measures P and Q and equals 0 iff $P=Q$ a.e..

Fact 2: The mapping $Q \to D_\alpha(Q,P)$ from $\mathcal{P}(\lambda)$ to $\mathbb{R}^+$ is convex.

1.2. **Semi parametric models.** In this paper we extend the power divergence approach to some specific class of semi parametric models. Such models are defined through constraints on moments; define l linearly independent functions

$$(X,\Theta) \ni (x,\theta) \to g_j(x,\theta) \quad 1 \le j \le l. \tag{1.4}$$

For any θ let's denote by $\mathcal{M}_\theta$ the set of all measures in $\mathcal{M}^1$ defined by

$$\mathcal{M}_\theta := \left\{ Q \in \mathcal{M}^1 \text{ such that } \int g_j(x,\theta) dQ(x) = 0, 1 \le j \le l \right\} \tag{1.5}$$

Measures in $\mathcal{M}_\theta$ therefor satisfy l linear constraints. The model $\mathcal{M}$ is defined through

$$\mathcal{M} = \cup_{\theta\in\Theta} \mathcal{M}_\theta \tag{1.6}$$

The inference on θ in the above model can be performed in a natural way for a number of statistical criterions. Indeed for example for Cressie Read criterions, or more generally for Csiszar type ones, a simple plug in of the empirical measure P_n in place of P in the divergence $D(Q,P)$ allows to minimize it on $\mathcal{M}_\theta$ for given θ, and then to optimize upon θ. This is due to the fact that the minimizer of $D(Q,P_n)$ on $\mathcal{M}_\theta$ has support included in the sample points X_i's. Therefore the seemingly formidable search for this minimization problem boils down to a finite dimensional one, on the simplex of $\mathbb{R}^n$. This is the core argument for Empirical Likelihood methods and their extensions. All minimum empirical divergence methods (therefore including EL) aim at assessing whether the model $\mathcal{M}$ is valid and at the estimation of θ_T , the true value of the parameter. so they do not provide any knowledge on the density of P_{θ_T} (whenever $P_0 = P_{\theta_T}$ belongs to $\mathcal{M}$) nor on the density of the projection of P_0 on $\mathcal{M}$ taking into account the very definition of the model.

In the present case, due to the very form of $D_\alpha(Q,P)$ as in (1.3) no plug in of P_n is feasible. Looking at (1.3) it may seem that the divergence D_α is not fitted to semiparametric inference. However extending the parametric setting to a smoothed semiparametric one, it is possible to make inference both on θ_T and on the density of P_{θ_T} . This is the focus of the present paper.

The setting when estimating θ in $\mathcal{M}$ is clearly quite different as in the parametric case, where $\mathcal{M}_\theta$ is not defined by any such condition as above, but merely consists in a single distribution P_θ. Consider the estimation of θ in $\mathcal{M}$ making use of (1.3). Clearly this yields to a two steps minimization; the first one consists in the search for the minimizer Q_θ of $R_\alpha(Q,P_n)$ for Q in $\mathcal{M}_\theta$, and the subsequent minimization should select the value of θ which indeed solves $\min_\theta R_\alpha(Q_\theta,P_n)$ where Q_θ solves the first minimization, whenever possible. Now the first minimization is

indeed difficult, since the class $\mathcal{M}_\theta$ consists in an infinite family of distributions on which the minimization of $R_\alpha(Q, P_n)$ should be performed.

This program can however be experimented, as soon as some appropriate setting is defined. This setting should contain various ingredients; firstly the model should be such that all minimization procedures should be well defined; our basic setting will imply that the mapping $\mathcal{M} \ni Q \to D_\alpha(Q, P)$ be sci in a proper topology for which, for all θ, the convex set $\mathcal{M}_\theta$ should be closed, for any $P.$, and the level sets of the mapping $Q \to D_\alpha(Q, P)$ should be compact. As such its estimator can also be defined. Additional regularity assumptions on the model, with respect to the variation of θ in Θ, will be necessary in order to perform the second optimization.

The problem at hand writes therefore

$$\widehat{\theta} := \arg\min_{\theta\in\Theta} \min_{Q\in\mathcal{M}_\theta} R_\alpha(Q, P_n), \tag{1.7}$$

where for all θ, $\mathcal{M}_\theta$ consists in a family of distributions with densities wrt the Lebesgue measure, with some prescribed regularity. We need to introduce some description on the model; this is done in the next Section.

2. Notation and properties of the semi parametric model

2.1. **Constraints.** All distributions in our model are defined on a compact subset K of $\mathbb{R}^m$.

The linearly independent functions $(g_1, ..., g_l)$ introduced in (1.4) should satisfy some basic requirements. Each of the functions g_l is defined on K with values in $\mathbb{R}$. hence $g := (g_1, ..., g_l)^T$ is defined on $K \times \Theta$ with values in $\mathbb{R}^l$ The parameter space Θ is a compact subset in $\mathbb{R}^d$.
We assume that for all θ the mapping

$$x \longrightarrow g(x, \theta) \quad \text{is continuous on} \quad int(K). \tag{G1}$$

All functions g_l's are uniformly bounded

$$\sup_\theta \sup_{x\in K} \|g(x, \theta)\| < \infty \tag{G2}$$

where $\|x\|$ designates the usual norm in $\mathbb{R}^l$.

We also assume uniform continuity of g in the sense

As $\theta_n \to \underline{\theta}$

$$\lim_{n\to\infty} \sup_{x\in K} \|g(x, \theta_n) - g(x, \underline{\theta})\| = 0 \tag{G3}$$

2.2. **Regularity and smoothness assumptions**
. The semi parametric model $\mathcal{M}$ will be assumed to consist in regular measures, in the sense that they should have density with respect to the Lebesgue measure λ on K, and that their densities should be smooth. This is formalized as follows.

Let $\mathcal{P}$ be the class of all probability measures with support K, and $\mathcal{P}(\lambda)$ the class of all probability measures in $\mathcal{P}$ which are a.c. wrt λ.

We now define a subset E of p.m's in $\mathcal{P}(\lambda)$ which is identified by some smoothness properties pertaining to their densities. Any measure Q in $\mathcal{P}(\lambda)$ is identified with its density q. An element in E will be indifferently identified either by some probability measure Q or by its density q.

The set E is endowed with the metric induced by the sup norm on K; for q and q' in E, denote

$$d(q, q') := \sup_{x \in K} |q(x) - q'(x)| .$$

Four conditions will be assumed on E.
1-Let x_0 be some point in K.

There exists $N > 0$ such that for all Q in E,

$$q(x_0) < N. \tag{E1}$$

2-The class E is equicontinuous: for all $\varepsilon > 0$, there exists $\delta > 0$ such that for all Q in E,

$$\sup_{|x-x'|<\delta} |q(x) - q(x')| < \varepsilon. \tag{E2}$$

3- For all $Q \in E$, the map $x \to q^\alpha(x)$ is a Lipschitz function

$$\sup_{x,y \in K} \frac{|q^\alpha(x) - q^\alpha(y)|}{|x - y|} \leq C \tag{E3}$$

for .some $C > 0$.

Remark 1. Since for any positive δ, $|q(x) - q(y)| < \delta$ implies $|q^\alpha(x) - q^\alpha(y)| < \eta$ for some $\eta > 0$, due to $\alpha \in (0, 1)$, it follows that when E2 holds, $\sup_{|x-y|<\varepsilon} |q^\alpha(x) - q^\alpha(y)| < \eta$, which implies that q^α is equicontinuous. Therefore E3 enforces E2.

For each θ we consider the parametric submodel

$$\mathcal{M}_\theta := \left\{ Q \in \mathcal{P}(\lambda) \text{ such that } \int q(x) \mathrm{d}\lambda(x) = 1, \int g(x, \theta) q(x) \mathrm{d}\lambda(x) = 0 \right\},$$

and its smooth counterpart

$$\mathcal{M}_{\theta_E} := \mathcal{M}_\theta \cap E,$$

which we assume to be non void. We define the model $\mathcal{M}$ through

$$\mathcal{M} = \cup_{\theta\in\Theta}\mathcal{M}_\theta$$

and the smooth version of $\mathcal{M}$ is defined by

$$\mathcal{M}_E = \cup_{\theta\in\Theta}\mathcal{M}_\theta \cap E = \cup_\theta \mathcal{M}_{\theta_E}.$$

The first additional condition is an identifiability property of the model with respect to θ.

We assume that for $\theta \neq \theta'$,

$$\mathcal{M}_\theta \cap \mathcal{M}_{\theta'} = \emptyset \tag{M1}$$

We assume that the collection of submodels $\mathcal{M}_\theta$ is well separated in the sense that

Suppose that

$$(d(\theta,\theta') > \epsilon) \Rightarrow \left(\inf_{\{q\in\mathcal{M}_{\theta_E}, q'\in\mathcal{M}_{\theta'_E}\}} d(q,q') > \delta \right). \tag{M2}$$

As a consequence of those smoothness assumptions we denote indifferently $D_\alpha(Q,P)$ by $D_\alpha(q,P)$. The same notation is adopted for $R_\alpha(Q,P)$ (to be defined further on), etc.

Example 2.1. $g(x) = x - \theta$, $\mathcal{M}_\theta = \{Q : \int x \mathrm{d}Q(x) = \theta\}$ and clearly $\mathcal{M}_\theta \cap \mathcal{M}'_\theta = \emptyset$. Whenever $\left|\int x(q(x) - q'(x))dx\right| > \varepsilon$ then $\int |q(x) - q'(x)|\, dx > \varepsilon/K$, and therefore $d(q,q') > \delta$ for some $\delta > 0$.

2.3. **The estimator.** Given an i.i.d. sample $(X_1, X_2, ..., X_n)$ such that (X_1) has distribution $P_{\theta_0} \in \mathcal{M}_{\theta_0}$ for some $\theta_0 \in \Theta$ we intend to provide an estimator for θ_0 minimizing the pseudo-distance between P_n and $\mathcal{M}_E$ where

$$P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$$

is the empirical measure pertaining to the sample set $(X_1, X_2, ..., X_n)$. Note that the estimation is performed in the smooth model $\mathcal{M}_E$ and not in $\mathcal{M}$.

We introduce the estimator of θ_0 in $\mathcal{M}_E$ by

$$\widehat{\theta} := \arg\inf_\theta \inf_{Q\in\mathcal{M}_{\theta_E}} D_\alpha(Q, P_n). \tag{2.1}$$

Formula (2.1) provides an natural estimate of θ_0 if $P_{\theta_0} \in \mathcal{M}_{\theta_{0_E}}$. Indeed under the identifiability condition $(H3)$ we prove that the above estimator converges to $\theta_0 = \arg\inf_\theta \inf_{Q\in\mathcal{M}_\theta} D_\alpha(Q, P_{\theta_0})$.(see Theorem 1 and Theorem 9).

In the alternative case that $P_{\theta_0} \in \mathcal{M}_{\theta_0}$ but $P_{\theta_0} \notin E$ then formula (2.1) defines an estimator of some $\tilde{\theta} := \arg\inf_\theta \inf_{Q\in\mathcal{M}_{\theta_E}} D_\alpha(Q, P_{\theta_0})$. Hence $\tilde{\theta}$ is the $D_\alpha-$projection

of P_{θ_0} on $\mathcal{M}_E$, and $\tilde{\theta}$ may be different from θ_0 but still represents a proxy of θ_0,in the smooth model. We will consider a natural condition which entrains that $\tilde{\theta} = \theta_0$.(Theorem 1).

3. Projection and regularization

We denote P_0 the distribution of the variable X_1. In this section we consider both cases $P_0 \in \mathcal{M}_{\theta_0}$ and $P_0 \in \mathcal{M}_{\theta_{0_E}}$for some θ_0.

Suppose that the following condition holds

$$\inf_{Q \in \mathcal{M}_{\theta_{0_E}}} D_\alpha(Q, P_0) < \inf_{Q \in \mathcal{M}_{\theta_E}} D_\alpha(Q, P_0) \tag{3.1}$$

for all $\theta \neq \theta_0$, whenever P_0 belongs to $\mathcal{M}_{\theta_0}$ which formalizes the fact that P_0 is approximated smoothly smoothly with a better score in $\mathcal{M}_{\theta_{0_E}}$ than in any $\mathcal{M}_{\theta_E}$, as soon as P_0 belongs to $\mathcal{M}_{\theta_0}$.

Theorem 1. Under (3.1) it holds, whenever P_0 belongs to $\mathcal{M}$ or to $\mathcal{M}_E$,

$$\theta_0 = \arg\inf_\theta \inf_{Q \in \mathcal{M}_{\theta_E}} D_\alpha(Q, P_0) = \arg\inf_\theta \inf_{Q \in \mathcal{M}_\theta} D_\alpha(Q, P_0) \tag{3.2}$$

.

Proof. First case: Suppose that $P_0 = P_{\theta_0} \in \mathcal{M}_E$., i.e. $P_O \in \mathcal{M}_{\theta_{0_E}}$. Then

$$\inf_{Q \in \mathcal{M}_{\theta_{0_E}}} D_\alpha(Q, P_0) = 0.$$

.

Since $\mathcal{M}_E \supset \mathcal{M}_{\theta_{0_E}}$, we have

$$\inf_{Q \in \mathcal{M}_E} D_\alpha(Q, P_0) = 0.$$

Furthermore, θ_0 realizes $\inf_{Q \in \mathcal{M}_{\theta_{0_E}}} D_\alpha(Q, P_0) = 0$.
So

$$\theta_0 \in \arg\inf_\theta \inf_{Q \in \mathcal{M}_{\theta_{0_E}}} D_\alpha(Q, P_0).$$

It must be shown that θ_0 is the only parameter θ that satisfies $\inf_{Q \in \mathcal{M}_{\theta_{0_E}}} D_\alpha(Q, P_0) = 0$.
Suppose that $\theta_1 \neq \theta_0$ such that $\theta_1 \in \arg\inf_\theta \inf_{Q \in \mathcal{M}_{\theta_{0_E}}} D_\alpha(Q, P_0)$.Then

$$\inf_{Q \in \mathcal{M}_{\theta_{1_E}}} D_\alpha(Q, P_0) = \inf_{Q \in \mathcal{M}_{\theta_{0_E}}} D_\alpha(Q, P_0) = 0.$$

Since $\mathcal{M}_{\theta_{1_E}} \subset \mathcal{M}_{\theta_1}$

$$0 = \inf_{Q \in \mathcal{M}_{\theta_{1_E}}} D_\alpha(Q, P_0) \geq \inf_{Q \in \mathcal{M}_{\theta_1}} D_\alpha(Q, P_0) \geq 0.$$

Hence

$$\inf_{Q\in\mathcal{M}_{\theta_1}} D_\alpha(Q,P_0)=0.$$

θ_0 is the only θ who realizes $P_0\in\mathcal{M}_{\theta_0}$ so θ_1 does not exist, otherwise $P_0=P_{\theta_1}$ due to M1.
Second case: Suppose that $P_0=P_{\theta_0}\in\mathcal{M}$ and $P_0\notin\mathcal{M}_E$.. Recall that

$$\theta_0=\arg\inf_\theta \inf_{Q\in\mathcal{M}_\theta} D_\alpha(Q,P_0).$$

We want to show that

$$\theta_0=\arg\inf_\theta \inf_{Q\in\mathcal{M}_{\theta_E}} D_\alpha(Q,P_0)$$

.
We project $P_0=P_{\theta_0}$ on $\mathcal{M}_E$ and define

$$\theta_1\in\arg\inf_\theta \inf_{Q\in\mathcal{M}_{\theta_E}} D_\alpha(Q,P_0).$$

Assume that $\theta_1\neq\theta_0$.
We then have

$$\inf_{Q\in\mathcal{M}_{\theta_{1_E}}} D_\alpha(Q,P_0)\leq \inf_{Q\in\mathcal{M}_{\theta_E}} D_\alpha(Q,P_0)$$

for all θ by definition of θ_1 .So taking $\theta=\theta_0$,we have

$$\inf_{Q\in\mathcal{M}_{\theta_{1_E}}} D_\alpha(Q,P_0)\leq \inf_{Q\in\mathcal{M}_{\theta_{0_E}}} D_\alpha(Q,P_0) \tag{3.3}$$

Under (3.1) it holds . $D_\alpha(\mathcal{M}_{\theta_{0_E}},P_{\theta_0})<D_\alpha(\mathcal{M}_{\theta_E},P_{\theta_0})$, for all $\theta\neq\theta_0$.
Then (3.3) is impossible, so $\theta_1=\theta_0$ .We have proved (3.2).

□

Before handling inference we need to explore some properties of minimum pseudo-distance approximations in $\mathcal{M}_E$. We will make use of a number of definitions, which we quote now. For fixed P in $\mathcal{M}_E$ the divergence $D_\alpha(.,P)|_E$ is the restriction of $Q\to D_\alpha(Q,P)$ on $\mathcal{M}_E$.
For fixed θ, define therefore the projection of P on $\mathcal{M}_{\theta_E}$

$$Q_\theta^*=\arg\inf_{Q\in\mathcal{M}_{\theta_E}} D_\alpha(Q,P)|_E$$

whenever defined.
Since for $Q\in\mathcal{M}_E$

$$D_\alpha(Q,P)|_E=D_\alpha(Q,P)$$

it holds

$$\arg\inf_{Q\in\mathcal{M}_{\theta_E}} D_\alpha(Q,P)=\arg\inf_{Q\in\mathcal{M}_{\theta_E}} D_\alpha(Q,P)|_E=Q_\theta^*.$$

We first set some general definition.

Definition 1. Let Ω be some subset of $\mathcal{P}$. The $\alpha-$divergence between the set Ω and a p.m. P is defined by

$$D_\alpha(\Omega, P) := \inf_{Q\in\Omega} D_\alpha(Q, P).$$

A probability measure $Q^* \in \Omega$, such that $D_\alpha(Q^*, P) < \infty$ and

$$D_\alpha(Q^*, P) \leq D_\alpha(Q, P) \quad \text{for all} \quad Q \in \Omega,$$

is called a projection of P on Ω. This projection may not exist, or may be not defined uniquely.

Definition 2. The sequence of functions $q_n \in E$ tends to q strongly if and if

$$\sup_{x\in K} |q_n(x) - q(x)| \to 0.$$

To $(Q_n)_n \subset \mathcal{M}_E$,we associate (q_n).If there exists some q in E such that

$$\sup_{x\in K} |q_n(x) - q(x)| \to 0, \tag{3.4}$$

then Q_n converges strongly to Q such that $Q(A) = \int 1_A(x)q(x)\mathrm{d}x$ for all $A \in \mathcal{B}(\mathbb{R})$ .Denote $\left(Q_n \underset{st}{\longrightarrow} Q\right)$ when (3.4) holds ; Q may not be a probability measure .

4. Projection:existence and uniqueness

We need some preliminary result pertaining to the properties of $\mathcal{M}_E$.

4.1. **Closure of $\mathcal{M}_E$.** By Arzela-Ascoli Theorem the set E is pre-compact when endowed by the strong topology defined in Definition 1.
Let (Q_n) be a family of probability measures on K; by compactness of K, it holds

Proposition 1. $(Q_n)_{n\geq 0}$ is a tight family.

As a consequence it holds

Proposition 2. Let $(Q_n)_{n\geq 1}$ be a family of p.m's with densities in E.Assume that there exists q in E such that $\lim_{n\to\infty} \sup_{x\in K} |q_n(x) - q(x)| \to 0$. Then $(Q_n)_{n\geq 1}$ is relatively compact.

Proof. Let $\{n_j\} \subset \{n\}$ and $\frac{dQ_{n_j}}{d\lambda}(x) = q_{n_j}(x)$,and $\sup_{x\in K} |q_{n_j}(x) - q(x)| \longrightarrow 0$ then (Q_{n_j}) converges to some p.m Q and $Q(A) = \int_A q(x)\mathrm{d}\lambda(x)$ for all A in $\mathcal{B}(K)$.
.
Indeed

$$\begin{aligned}\left|Q_{n_j}(A) - \int_A q(x)\mathrm{d}\lambda(x)\right| &= \left|\int 1_A(x)q_{n_j}(x)\mathrm{d}\lambda(x) - \int 1_A(x)q(x)\mathrm{d}\lambda(x)\right| \\ &\leq \sup_{x\in K} \left|q_{n_j}(x) - q(x)\right| \lambda(A) \longrightarrow 0.\end{aligned}$$

So $(Q_{n_j})_{j\geq 1}$ converges to Q, such that $q(x) = \frac{dQ}{d\lambda}(x)$ .That Q is a probability measure is a consequence of Prohorov Theorem since $(Q_n)_{n\geq 1}$ is a tight family of p.m's . □

Theorem 2. Under G1, G2 and G3 the set $\mathcal{M}_E$ is closed for the strong topology of convergence defined in Definition 1.

Proof. Assume that $(Q_n)_{n\geq 1} \subset \mathcal{M}_E$ and assume that there exists q such that

$$\sup_E |q_n(x) - q(x)| \longrightarrow 0,$$

with $q_n(x) := (dQ_n/d\lambda)(x)$. Define $Q(A) := \int_A q(x)\mathrm{d}\lambda(x)$for any set A and we have $\left(Q_n \underset{st}{\longrightarrow} Q\right)$ by Proposition 2 . We want to prove that $Q \in \mathcal{M}_E$

(A) q is a density
(B) $\int_K g(x,\theta)q(x)\mathrm{d}x = 0$ for some θ .
(C) q is equicontinuous.

We prove (A); This follows from Prohorov Theorem. We prove (B) Let θ_n be defined by $\int g(x,\theta_n)q_n(x)\mathrm{d}x = 0$; such a θ_n indeed exists since $Q_n \in \mathcal{M}$.
Since Θ is a compact set in $\mathbb{R}^d$,we select $n_j \subset n$ such that the subsequence θ_{n_j} admits a limit $\underline{\theta}$ and $\int g(x,\theta_{n_j})q_{n_j}(x)\mathrm{d}x = 0$.
We prove that $\left|\int_K g(x,\underline{\theta})q(x)\mathrm{d}x\right| = 0$
Indeed

$$\begin{aligned}\left|\int_K g(x,\underline{\theta})q(x)\mathrm{d}x\right| &\leq \left|\int_K g(x,\underline{\theta})q_{n_j}(x)\mathrm{d}x\right| + \left|\int_K g(x,\underline{\theta})q(x)\mathrm{d}x - \int_K g(x,\underline{\theta})q_{n_j}(x))\mathrm{d}x\right| \\ &\leq B + A\end{aligned}$$

$$A = \left|\int_K g(x,\underline{\theta})\left(q(x) - q_{n_j}(x)\right)\mathrm{d}x\right|$$

which tends to 0 by G2.
Next

$$B \leq \int_K \left|g(x,\underline{\theta}) - g(x,\theta_{n_j})\right| q_{n_j}(x)\mathrm{d}x + \left|\int_K g(x,\theta_{n_j})q_{n_j}(x)\mathrm{d}x\right| \leq C + D$$

and $D = 0$ by definition of θ_{n_j}.

Hence

$$\begin{aligned}B &\leq C = \int_K \left|g(x,\underline{\theta}) - g(x,\theta_{n_j})\right| q_{n_j}(x)\mathrm{d}x \\ &\leq \sup_{x\in K}\left|g(x,\underline{\theta}) - g(x,\theta_{n_j})\right| \int_K q_{n_j}(x)\mathrm{d}x \\ &= \sup_{x\in K}\left|g(x,\underline{\theta}) - g(x,\theta_{n_j})\right| \to 0\end{aligned}$$

where we used G3
We have proved that any converging sequence θ_{n_j} satisfies $\int_K g(x,\underline{\theta})q(x)\mathrm{d}x$ when

$\underline{\theta} = \lim_{n_j \to \infty} \theta_{n_j}$.
Consider two converging subsequences n_j and n'_j with $\theta_{n_j} \to \underline{\theta}$ and $\theta'_{n_j} \to \bar{\theta}$,we have

$$\int_K g(x, \underline{\theta}) q(x) \mathrm{d}x = \int_K g(x, \bar{\theta}) q(x) \mathrm{d}x.$$

By $M1$ it follows that $\underline{\theta} = \bar{\theta}$ therefore we have prove that there exists a unique $\theta \in \Theta$ such that

$$\int_K g(x, \theta) q(x) \mathrm{d}x = 0$$

which proves (B).

We prove that there exists some $N > 0$ such that $|q(x_0)| \leq N$. Indeed

$|q(x_0) - q_n(x_0) + q_n(x_0)| \leq |q_n(x_0)| + |q_n(x_0) - q(x_0)| \leq N + |q_n(x_0) - q(x_0)| \leq N + \varepsilon$ for all $\varepsilon > 0$ and therefore $|q(x_0)| \leq N$, since

$$|q_n(x_0) - q(x_0)| \to 0.$$

We prove that q is uniformly equicontinuous on K; indeed

$$|q(x) - q(x')| = |q(x) - q_n(x) + q_n(x) - q(x') + q_n(x') - q_n(x')|.$$

Hence

$$\sup_{|x-x'|<\delta} |q(x) - q(x')| \leq \sup_{|x-x'|<\delta} |q(x) - q_n(x)| + \sup_{|x-x'|<\delta} |q(x') - q_n(x')| + \sup_{|x-x'|<\delta} |q_n(x) - q_n(x')|$$
$$\leq 2 \sup_{x \in K} |q(x) - q_n(x)| + \sup_{|x-x'|<\delta} |q_n(x) - q_n(x')| \leq 2\varepsilon + \eta.$$

The first term in the last display tends to 0 by hypothesis; the second one is smaller than any positive ε for adequate $\delta > 0$. Hence $q \in E$. □

4.2. **Existence and uniqueness of the D_α-projection of P on $\mathcal{M}_E$.**. For any P in $\mathcal{P}(\lambda)$ let $a > 0$ and

$$A_E(a) := \{Q \in \mathcal{M}_E : D_\alpha(Q, P) \leq a\}$$

be some level set of the divergence $Q \to D_\alpha(Q, P)$.

Proposition 3. For any $\alpha \in (0, 1)$ the divergence function $Q \mapsto D_\alpha(Q, P)$ from $\mathcal{P}(\lambda)$ to $[0, +\infty]$ is s.c.i. for the strong topology.

Proof. We prove that $A_E(a)$ is a closed subset in $\mathcal{M}_E$ equipped with the strong topology . Recall that $Q \to D_\alpha(Q, P)$ s.c.i is equivalent to $A_E(a)$ is closed.
Let $Q_n \in A_E(a) \cap \mathcal{M}_E$. Denote $\frac{dQ_n}{d\lambda}(x) = q_n(x)$ with $q_n \in E$,and assume that there exists a function q defined on K such that

$$\sup_{x \in K} |q_n(x) - q(x)| \to 0.$$

Define

$$\frac{dQ}{d\lambda}(x) = q(x)$$

we prove that $q \in E$ and with $Q(A) := \int 1_A(x)q(x)\mathrm{d}\lambda(x)$, it holds $Q \in A_E(a)$.
Since $\mathcal{M}_E$ is closed, (see Theorem 2) the measure Q defined by $Q(A) = \int 1_A(x)q(x)\mathrm{d}\lambda(x)$ for all $A \in \mathcal{B}(\mathbb{R})$ is in $\mathcal{M}_E$.
It remains to prove that $D_\alpha(Q, P) \leq a$.
Consider the concave mapping $t \to t^\alpha$ defined on $\mathbb{R}^+$ which thus satisfies

$$|t^\alpha - s^\alpha| \leq \alpha \max\left(t^{\alpha-1}, s^{\alpha-1}\right)|t - s| \tag{4.1}$$

and set $t := q_n^\beta(x)$ and $s := q^\beta(x)$ with $\beta := (\alpha+1)/\alpha$; we then have

$$\sup_{x\in K}\left|q_n^{\alpha+1}(x) - q^{\alpha+1}(x)\right| \leq \sup_{x\in K}\left\{(\alpha+1)\left[\max\left(q_n^{\alpha+1}(x), q^{\alpha+1}(x)\right)\right]|q_n^\alpha(x) - q^\alpha(x)|\right\} \tag{4.2}$$

It holds similarly

$$\sup_{x\in K}|q_n^\alpha(x) - q^\alpha(x)| \leq \sup_{x\in K}\left\{\alpha\left[\max\left(q_n(x), q(x)\right)\right]|q_n(x) - q(x)|\right\} \to 0. \tag{4.3}$$

Since the function q is bounded on K .
We have

$$\sup_{x\in K}\left|q_n^{\alpha+1}(x) - q^{\alpha+1}(x)\right| \leq (\alpha+1)\sup_{x\in K}|q^\alpha(x)||q_n(x) - q(x)| \to 0$$

Since f_n is bounded on K,

$$|q_n^\alpha(x) - q^\alpha(x)| \leq |q_n^\alpha(x)| + |q^\alpha(x)| < \infty.$$

So $|q_n^\alpha(x) - q^\alpha(x)|$ is bounded.
Consider now the mapping

$$x \to \varphi(q_n(x), p(x)) - \varphi(q(x), p(x)).$$

Since

$$\varphi(q_n(x), p(x)) - \varphi(q(x), p(x)) = q_n^{\alpha+1}(x) - q^{\alpha+1}(x) - \left(1 + \frac{1}{\alpha}\right)p(x)\left(q_n^\alpha(x) - q^\alpha(x)\right)$$

.
using (4.2) and (4.3)

$$\sup_{x\in K}|\varphi(q_n(x), p(x)) - \varphi(q(x), p(x))| \to 0$$

.
Integrating we have

$$\int \varphi(q_n(x), p(x))\mathrm{d}x - \delta \leq \int \varphi(q(x), p(x))\mathrm{d}x = D_\alpha(Q, P) \leq \int \varphi(q_n(x), p(x))\mathrm{d}x + \delta. \tag{4.4}$$

for any $\delta > 0$, for n large. Since $Q_n \in A_E(a)$, $\int \varphi(q_n(x), p(x))\mathrm{d}x \leq a$; the inequality (4.4) becomes

$$\int \varphi(q_n(x), p(x))\mathrm{d}x - \delta \leq \int \varphi(q(x), p(x))\mathrm{d}x \leq \int \varphi(q_n(x), p(x))\mathrm{d}x + \delta \leq a + \delta$$

So $\int \varphi(q(x), p(x))\mathrm{d}x \leq a$; hence $Q \in A_E(a)$ and thus $A_E(a)$ is a closed set in $\mathcal{M}_E$. □

Theorem 3. For all $a > 0$ the set $A_E(a)$ is compact for the strong topology.

Proof. By Arzela-Ascoli Theorem, E has a compact closure . $A_E(a)$ is closed in $Cl(E)$.
$A_E(a)$ is a closed subset of $Cl(E)$,which is compact □

Proposition 4. For any θ in Θ

$$Q^* = \arg \inf_{Q \in \mathcal{M}_{\theta_E}} D_\alpha(Q, P).$$

exists and is unique.

Proof. Let $a_\theta := \inf_{Q \in \mathcal{M}_{\theta_E}} D_\alpha(Q, P)$ and let $\varepsilon > 0$. .Then $A_E(a_\theta + \varepsilon) \neq \emptyset$ $A_E(a_\theta + \varepsilon) \cap \mathcal{M}_{\theta_E} \neq \emptyset$.

It can be observed that for all θ the set $\mathcal{M}_{\theta_E}$ is a closed set, following the same arguments as in Proposition 3. Since $\mathcal{M}_{\theta_E}$ is closed and $A_E(a_\theta + \varepsilon)$ is compact then $A_E(a_\theta + \varepsilon) \cap \mathcal{M}_{\theta_E}$ is compact. Since

$$\arg \inf_{q \in \mathcal{M}_{\theta_E}} D_\alpha(Q, P) = \arg \inf_{q \in A_E(a_\theta + \varepsilon) \cap \mathcal{M}_{\theta_E}} D_\alpha(Q, P)$$

, existence of the projection follows from the lower semi continuity of $Q \to D_\alpha(Q, P)$.Since φ is strictly convex,then the function $Q \in \mathcal{P}(\lambda) \to D_\alpha(Q, P)$ is also strictly convex ,and the projection of P on any closed convex set Ω in $\mathcal{M}_{\theta_E}$ is uniquely defined whenever it exists. □

Consider now the D_α projection of P on a convex subset in $\mathcal{M}_E$. Similarly as in Proposition 4 it holds

Theorem 4. For any convex set Ω in $\mathcal{M}_E$ the D_α projection of P on Ω exists and is unique.

Proof. The proof mimics the one in Proposition 4. Let

$$a := \inf_{Q \in \mathcal{M}_E} D_\alpha(Q, P)$$

and $\varepsilon > 0$. Then $A_E(a + \varepsilon) \cap \mathcal{M}_E \neq \emptyset$ .Since $\mathcal{M}_E$ is closed (see Theorem **??**) and $A_E(a + \varepsilon)$ is compact, existence of the projection follows. Uniqueness is due to convexity. . ■ □

5. Minimum pseudo-distance estimator

Let $X_1, ..., X_n$ denote an i.i.d. sample of a random vector $X \in \mathbb{R}^m$ with distribution P_0. Let $P_n(.)$ be the empirical measure pertaining to this sample, namely

$$P_n(.) := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(.),$$

where $\delta_x(.)$ denotes the Dirac measure at point x. We define

$$\begin{aligned} D_\alpha(\mathcal{M}_{\theta_E}, P_0) &= \inf_{Q\in\mathcal{M}_{\theta_E}} D_\alpha(Q, P_0) \\ &= \inf_{Q\in\mathcal{M}_{\theta_E}} \left\{ \int \left(q^{\alpha+1}(x) - \left(1+\frac{1}{\alpha}\right) q^\alpha(x) p_0(x) + \frac{1}{\alpha} p_0^{\alpha+1}(x) \right) \mathrm{d}x \right\} \end{aligned}$$

Since optimization only pertains to Q define in the following

$$\begin{aligned} R_\alpha(\mathcal{M}_{\theta_E}, P_0) &= \inf_{Q\in\mathcal{M}_{\theta_E}} R_\alpha(Q, P_0) \\ &= \inf_{Q\in\mathcal{M}_{\theta_E}} \left\{ \int \left(q^{\alpha+1}(x) - \left(1+\frac{1}{\alpha}\right) q^\alpha(x) p_0(x) \right) \mathrm{d}x \right\} \end{aligned}$$

the "plug-in" estimate of $R_\alpha(\mathcal{M}_{\theta_E}, P_0)$ through

$$\begin{aligned} \widehat{R}_\alpha(\mathcal{M}_{\theta_E}, P_0) &:= \inf_{Q\in\mathcal{M}_{\theta_E}} R_\alpha(Q, P_n) \\ &= \inf_{Q\in\mathcal{M}_{\theta_E}} \left\{ \int q^{\alpha+1}(x)\mathrm{d}x - \left(1+\frac{1}{\alpha}\right) \int q^\alpha(x)\mathrm{d}P_n(x) \right\} \\ &= \inf_{Q\in\mathcal{M}_{\theta_E}} \left\{ \int q^{\alpha+1}(x)\mathrm{d}x - \left(1+\frac{1}{\alpha}\right) \frac{1}{n}\sum_{i=1}^{n} q^\alpha(X_i) \right\} \end{aligned}$$

In the same way,

$$\begin{aligned} R_\alpha(\mathcal{M}, P_0) &:= \inf_{\theta\in\Theta} \inf_{Q\in\mathcal{M}_{\theta_E}} R_\alpha(Q, P_0) \\ &= \inf_{\theta\in\Theta} \inf_{Q\in\mathcal{M}_{\theta_E}} \left\{ \int q^{\alpha+1}(x)\mathrm{d}x - \left(1+\frac{1}{\alpha}\right) \int q^\alpha(x)\mathrm{d}P_0(x) \right\} \end{aligned}$$

can be estimated by

$$\widehat{R}_\alpha(\mathcal{M}, P_0) := \inf_{\theta\in\Theta} \inf_{Q\in\mathcal{M}_{\theta_E}} \left\{ \int q^{\alpha+1}(x)\mathrm{d}x - \left(1+\frac{1}{\alpha}\right) \frac{1}{n}\sum_{i=1}^{n} q^\alpha(X_i) \right\}$$

Since

$$\arg \inf_{q\in\mathcal{M}_{\theta_E}} D_\alpha(\mathcal{M}_\theta, P_0) = \arg \inf_{q\in\mathcal{M}_{\theta_E}} R_\alpha(\mathcal{M}_\theta, P_0)$$

for any θ

$$\arg \inf_{q\in\mathcal{M}_{\theta_E}} R_\alpha(\mathcal{M}_{\theta_E}, P_0)$$

exists and is unique($P_0 \in \cup \mathcal{M}_{\theta_E}$ or not).
We will consider estimators of θ_0 where $P_0 = P_{\theta_0}$ for a unique $\theta_0 \in \Theta$;this corresponds to the fact that $P_0 \in \mathcal{M}$. In this cases by uniqueness of $\arg\inf_{\theta\in\Theta} R_\alpha(\mathcal{M}_{\theta_E}, P_0)$ and since the infimum is reached at $\theta = \theta_0$ under the model, we estimate θ_0 through

$$\widehat{\theta}_\alpha := \arg\inf_{\theta\in\Theta} \inf_{Q\in\mathcal{M}_{\theta_E}} \left\{ \int q^{\alpha+1}(x)\mathrm{d}x - \left(1+\frac{1}{\alpha}\right)\frac{1}{n}\sum_{i=1}^{n} q^\alpha(X_i) \right\}$$

6. Asymptotic properties

The pseudodistances BHHJ will be applied in the standard statistical estimation model with i.i.d observations $X_1, ..., X_n$ governed by P_0 from a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probability measures on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ indexed by a set parameters $\Theta \subset \mathbb{R}^d$. All distributions in $\mathcal{P}$ are assumed absolutely continuous, and λ denotes the Lebesgue measure on $\mathbb{R}^k$. Denote $p_\theta = dP_\theta/d\lambda$ for $\theta \in \Theta$.

Remark 2. If $P_0 \in \mathcal{M}$ there exists an unique $P_{\theta_0} \in \mathcal{M}$ such that $P_0 = P_{\theta_0} \in \mathcal{M}$ then by identifiability

$$\arg\inf_\theta D_\alpha(P_\theta, P_{\theta_0}) = \theta_0$$

In other words the unknown parameter θ_0 is the unique minimizer of the function $D_\alpha(P_\theta, P_0)$

$$\theta_0 = \arg\min_\theta D_\alpha(P_\theta, P_{\theta_0}) \in \Theta \tag{6.1}$$

The empirical probability measures P_n are known to converge weakly to P_0 as $n \longrightarrow \infty$. Therefore by plugging in 6.1 the measures P_n for P_0 one intuitively expects to obtain the estimator under the form

$$\widehat{\theta}_\alpha(n) = \arg\min_{\theta\in\Theta} M_n(P_\theta, P_n)$$

that converges to θ_0 as $n \to \infty$, where $M_n(P_\theta, P_n)$ is some empirical criterion which estimates the objective function $R_\alpha(P_\theta, P_0)$.

We will repeatedly make use of a basic result which we recall for convenience. Denote $M_n(\tau)$ a family of random functions of a parameter τ which belongs to a space T endowed which a metric denoted d .
Assuming that the sequences M_n converges uniformly to some deterministic function M defined on T, then the following result provides a set of sufficient conditions which entail the weak convergence of minimizers of M_n to the minimizer of M ,if well defined.

Lemma 1. (Van der Vaart (1998), theorem 5.7) Assume that (1)$\sup_{\tau\in T} |M_n(\tau) - M(\tau)| \xrightarrow{P} 0$, (2)For any $\epsilon > 0, \inf_{\{t\in T, d(t,t_0)\geq\epsilon\}} M(t) > M(t_0)$, (3) the sequence t_n satisfies

$$M_n(t_n) \leq M_n(t_0) + 0_p(1)$$

Then the sequence t_n satisfies

$$d(t_n, t_0) \xrightarrow{P} 0.$$

Lemma1 will be used according to the context of minimization at hand.
By (1.7) we consider the inner and the outer minimization problems leading to the estimator. This will be performed in two steps: the inner minimization with respect to Q in $\mathcal{M}_{\theta_E}$ for fixed θ, and the outer minimization wrt θ.
Here we establish the consistency of the minimum pseudodistance estimator on the closed set of measures a.c.w.r.t λ .

6.1. **Inner minimization: convergence of the projection of $P_n$on $\mathcal{M}_{\theta_E}$.** Fix $\theta \in \Theta$. Denote

$$M_n(Q) := R_\alpha(Q, P_n)$$

where $Q \in \mathcal{M}_{\theta_E}$.

Denote

$$q_n(\theta) := \arg \inf_{q \in \mathcal{M}_{\theta_E}} R_\alpha(Q, P_n). \tag{6.2}$$

where,as before,Q is identified with its density q .Existence and uniqueness of a p.m $Q_n(\theta)$ with density $q_n(\theta)$ follows from Proposition 4,following verbatim its proof,substituting P by P_n .
Denote accordingly the unique minimizer of $R_\alpha(Q, P_0)$,

$$q_\theta^* := \arg \inf_{q \in \mathcal{M}_{\theta_E}} R_\alpha(Q, P_0). \tag{6.3}$$

We prove that $q_n(\theta)$ converges to q_θ^* making use of Lemma 1.
Setting

$$M_n(\tau) := R_\alpha(Q, P_n),$$

with $\tau = \frac{dQ}{d\lambda}$,setting $d(\tau, \tau') = \sup_{x \in K} |q(x) - q'(x)|$,it holds.

Lemma 2. Fix θ. Then Condition (1) in Lemma 1 holds

$$\sup_{q \in \mathcal{M}_{\theta_E}} |R_\alpha(Q, P_n) - R_\alpha(Q, P_0)| \to 0 \text{ in probability}$$

Proof. It holds

$$\sup_{q \in \mathcal{M}_{\theta_E}} |R_\alpha(Q, P_n) - R_\alpha(Q, P_0)| \leq \left(1 + \frac{1}{\alpha}\right) \sup_{q \in \mathcal{M}_{\theta_E}} \left| \frac{1}{n} \sum_{i=1}^{n} q^\alpha(X_i) - E_{P_0}(q^\alpha(X)) \right|$$

which tends to 0 almost surely as n tends to infinity, since q^α is a Lipschitz function for all $q \in \mathcal{M}_{\theta_E}$, and a class of Lipschitz function is a Glivenko-Cantelli class .

□

We now prove that the second condition in Lemma 1 holds

Lemma 3. For any $\varepsilon > 0$,

$$\inf_{\{q:\|q-q^*(\theta)\|>\epsilon, q\in\mathcal{M}_{\theta_E}\}} R_\alpha(Q, P_0) > R_\alpha(Q^*_\theta, P_0).$$

where $dQ/dP = q$ and $\mathrm{d}Q^*_\theta/dP = q^*_\theta$.

Proof. We thus prove condition (2) in Lemma 1. By Proposition4

$$Q^*_\theta := \arg\inf_{Q\in\mathcal{M}_{\theta_E}} R_\alpha(Q, P_0)$$

exists with uniqueness. Denote $q^*_\theta := \frac{dQ^*(\theta)}{d\lambda}$ .It holds

$$\inf_{\|q-q^*_\theta\|>\varepsilon, q\in\mathcal{M}_{\theta_E}} R_\alpha(Q, P_0) > R_\alpha(Q^*(\theta), P_0)$$

.Indeed by definition for all Q ,such that $\frac{dQ}{d\lambda}(x) = q(x)$

$$R_\alpha(Q^*(\theta), P_0) \leq R_\alpha(Q, P_0)$$

and therefore

$$R_\alpha(Q^*(\theta), P_0) \leq \inf_{\|q-q^*(\theta)\|>\varepsilon} R_\alpha(Q, P_0).$$

Now let $Q^*(\theta)$ such that $dQ^*(\theta)/d\lambda(x) = q^*(\theta)(x)$ and Q such that $dQ(\theta)/d\lambda(x) = q(\theta)(x)$. We prove that the inequality is strict .From the above display we get

$$R_\alpha(q^*(\theta), P_0) + \frac{1}{\alpha}\int p_0^{\alpha+1}(x)\mathrm{d}x \leq \inf_{\|q-q^*(\theta)\|>\varepsilon}\left\{R_\alpha(q, P_0) + \frac{1}{\alpha}\int p_0^{\alpha+1}(x)\mathrm{d}x\right\}$$

i.e.

$$D_\alpha(\mathcal{M}_{\theta_E}, P_0) \leq \inf_{\|q-q^*(\theta)\|>\varepsilon, q\in\mathcal{M}_{\theta_E}} D_\alpha(Q, P_0)$$

Now if equality holds,there exists $q \in \mathcal{M}_{\theta_E}$, $q \neq q^*(\theta)$ such that

$$D_\alpha(\mathcal{M}_{\theta_E}, P_0) = D_\alpha(q^*(\theta), P_0) = D_\alpha(q, P_0) \tag{6.4}$$

. It hold $Q \neq Q^*(\theta)$ since $q^*(\theta)$ and $q \in E$ .But the projection of P_0 on $\mathcal{M}_{\theta_E}$ is unique, so (6.4) cannot hold.

□

We also prove that the third condition in Lemma 1 holds.

Lemma 4. It holds

$$R_\alpha(q_n(\theta), P_n) \leq R_\alpha(q^*_\theta, P_0) + o_p(1).$$

Proof. This follows from the very definition of $q_n(\theta)$ for which $R_\alpha(q_n(\theta), P_n) \leq R_\alpha(q, P_n)$ for all $q \in \mathcal{M}_{\theta_E}$. □

Making use Lemma 1 we have proved

Theorem 5. For any $\theta \in \Theta$,it holds, with $q_n(\theta)$ defined in (6.2) and q_θ^* defined in (6.3)

$$\sup_{x\in K} |q_n(\theta)(x) - q_\theta^*(x)| \xrightarrow{P} 0.$$

6.2. **Outer minimization.** We now consider the minimization in θ ,with the following notation .Let

$$\hat{\theta}_n := \arg\inf_\theta \inf_{q\in\mathcal{M}_{\theta_E}} R_\alpha(Q, P_n) = \arg\inf_\theta R_\alpha(q_n(\theta), P_n)$$

and

$$\theta_0 := \arg\inf_\theta \inf_{q\in\mathcal{M}_{\theta_E}} R_\alpha(Q, P_0) = \arg\inf_\theta R_\alpha(q_\theta^*, P_0)$$

The parameter θ_0 such that $P_0 = P_{\theta_0}$ is defined in a unique way by the above display; indeed firstly note that θ_0 is well defined ,either when $P_0 \in \mathcal{M}$ (i.e. $P_0 = P_{\theta_0}$) (see Theorem 1) or $P_0 \notin \mathcal{M}$,in which caseP_{θ_0} is the $D_\alpha-$projection of P_0 on $\mathcal{M}_E$.
By the Theorem 5,we have proved that

$$\sup_{x\in K} |q_n(\theta)(x) - q_\theta^*(x)| \xrightarrow{P} 0.$$

where q_θ^* is defined in (6.3).We want to show that

$$\arg\inf_\theta R_\alpha(q_n(\theta), P_n) \xrightarrow{P} \arg\inf_\theta R_\alpha(q_\theta^*, P_0).$$

where $q_\theta^* = \arg\inf_{q\in\mathcal{M}_{\theta_E}} R_\alpha(Q, P_0)$.

By definition

$$\hat{\theta}_n := \arg\inf_\theta R_\alpha(q_n(\theta), P_n);$$

we prove that

$$\arg\inf_\theta R_\alpha(q_\theta^*, P_0) = \theta_0 \tag{6.5}$$

We consider two cases:

(Case 1) If $P_0 \in \mathcal{M}$, i.e. if $\exists!\theta_0$ such that $P_0 = P_{\theta_0}$then (6.5) holds.
(Case 2) If $P_0 \notin \mathcal{M}$,$\theta_0 = \arg\inf_{Q\in\mathcal{M}_E} D_\alpha(Q, P_0)$and under Condition (3.1) ,

$$\theta_0 = \arg\inf_\theta \inf_{Q\in\mathcal{M}_{\theta_E}} D_\alpha(Q, P_0).$$

Therefore (6.5) holds.
We make use of Lemma 1 with

$$\begin{aligned} M_n(\theta) &: = R_\alpha(q_n(\theta), P_n), \\ M(\theta) &: = R_\alpha(Q_\theta^*, P_0). \end{aligned} \tag{6.6}$$

We prove that $\hat{\theta}_n$ converges to θ_0 making use of Lemma 1.
Set

$$M_n(\tau) := R_\alpha(Q_n(\theta), P_n)$$

with $q_n(\theta)(x) = \frac{dQ_n(\theta)}{d\lambda}(x)$ setting

$$d(q_n(\theta), q^*_\theta) = \sup_{x\in K} |q_n(\theta)(x) - q^*_\theta(x)|$$

it holds.

Proposition 5. Suppose that the following condition

$$\sup_{\{q\in\mathcal{M}_{\theta_E}, q'\in\mathcal{M}_{\theta'_E}, d(\theta,\theta')<\delta\}} d(q, q') < K\delta \tag{6.7}$$

holds for some $K > 0$ independent on θ and θ'; then

$$\sup_{\theta\in\Theta}\sup_{x\in K} |q_n(\theta)(x) - q^*_\theta(x)| \xrightarrow{P} 0.$$

Proof. By Theorem 5 for all θ

$$d(q_n(\theta), q^*_\theta) \to 0 \text{ in probability.}$$

We want to prove that uniform convergence upon θ holds. Define θ_n by

$$sup_{\theta\in\Theta} d(q_n(\theta), q^*_\theta) = d(q_n(\theta_n), q^*_{\theta_n}). \tag{6.8}$$

Let $\{n_j\} \subset \{n\}$and suppose $\underline{\theta}$ such that $\theta_{n_j} \to \underline{\theta}$.
We show that $d(q_{n_j}(\theta_{n_j}), q^*_{\theta_{n_j}}) > c > 0$ cannot hold.
Now by definition (6.8)

$$\begin{aligned} sup_{\theta\in\Theta} d(q_{n_j}(\theta), q^*_\theta) &= d(q_{n_j}(\theta_{n_j}), q^*_{\theta_{n_j}}) \\ &\leq d(q_{n_j}(\theta_{n_j}), q_{n_j}(\underline{\theta})) + d(q_{n_j}(\underline{\theta}), q^*_{\underline{\theta}}) + d(q^*_{\theta_{n_j}}, q^*_{\underline{\theta}}) \\ &= : I_1 + I_2 + I_3. \end{aligned}$$

Now $I_1 = d(q_{n_j}(\theta_{n_j}), q_{n_j}(\underline{\theta}))$ and $d(\theta_{n_j}, \underline{\theta}) \to 0$. Hence under (6.7), $I_1 \xrightarrow{P} 0$. Now $I_2 = d(q_{n_j}(\underline{\theta}), q^*_{\underline{\theta}})$; both $q_{n_j}(\underline{\theta})$ and $q^*_{\underline{\theta}}$ belong to $\mathcal{M}_{\underline{\theta}_E}$; By Theorem 5 in $\mathcal{M}_{\underline{\theta}_E}$, $d(q_{n_j}(\underline{\theta}), q^*_{\underline{\theta}}) \xrightarrow{P} 0$ so $I_2 \xrightarrow{P} 0$,
As for $I_3 = d(q^*_{\underline{\theta}}, q^*_{\theta_{n_j}}) \xrightarrow{P} 0$ as for I_1.We have proved that

$$\lim_{j\to\infty}\sup_{\theta\in\Theta} d\left(q_{n_j}(\theta), q^*_\theta\right) = 0 \text{ in probability.} \tag{6.9}$$

Assume now that (6.8)

does not hold. In such a case there exists a subsequence $\{m_k\} \subset \{n\}$ and $\eta > 0$ such that

$$\sup_\theta d(q_{m_k}(\theta), q^*_\theta) > \eta.$$

Let $\theta_{m_k} := \arg\sup_\theta d(q_{m_{kj}}(\theta), q^*_\theta)$, whence

$$d(q_{m_k}(\theta_{n_j}), q^*_{\theta_{m_k}}) > \eta$$

for all k. Extract from $\{m_k\}$ a further subsequence $\{n_j\}$ along which θ_{n_j} converges to some $\underline{\theta}$. Then (6.9) proves our claim, by contradiction. □

Under Condition (6.7) in Proposition 5 , condition (1) in Lemma 1 holds i.e.

$$\sup_{\theta\in\Theta}|M_n(\theta)-M(\theta)|\xrightarrow{P}0$$

with $M_n(\theta)$ and $M(\theta)$ defined in (6.6)

Proof. .Define

$$M_n(\theta)=R_\alpha(q_n(\theta),P_n),\ \text{and}\ M(\theta)=R_\alpha(q_\theta^*,P_0)$$

with

$$R_\alpha(q_n(\theta),P_n)=\int q_n^{\alpha+1}(\theta)(x)\mathrm{d}x-\left(1+\frac{1}{\alpha}\right)\int q_n^{\alpha}(\theta)(x)\mathrm{d}P_n(x)$$

and

$$R_\alpha(q_\theta^*,P_0)=\int q_\theta^{*\alpha+1}(x)\mathrm{d}x-\left(1+\frac{1}{\alpha}\right)\int (x)q_\theta^{*\alpha}(x)\mathrm{d}P_0(x)$$

Hence

$$\begin{aligned}\sup_{\theta\in\Theta}|M_n(\theta)-M(\theta)|\leq&\sup_{\theta\in\Theta}\int\left|q_n^{\alpha+1}(\theta)(x)-q_\theta^{*\alpha+1}(x)\right|\mathrm{d}x+\\&\left(1+\frac{1}{\alpha}\right)\sup_{\theta\in\Theta}\int|q_n^{\alpha}(\theta)(x)-q_\theta^{*\alpha}(x)|\,\mathrm{d}P_n(x)\\&+\left(1+\frac{1}{\alpha}\right)\sup_{\theta\in\Theta}\left|\int q_\theta^{*\alpha}(x)\mathrm{d}(P_n-P_0)\right|\\&\leq R_1+R_2+R_3.\end{aligned}$$

Now

$$\begin{aligned}R_1&=sup_{\theta\in\Theta}\int\left|q_n^{\alpha+1}(\theta)(x)-q_\theta^{*\alpha+1}(x)\right|\mathrm{d}x\\&\leq sup_{\theta\in\Theta}sup_{x\in K}|q_n(\theta)(x)-q_\theta^*(x)|\times Cste\end{aligned}$$

which tends to 0 in Probability by Proposition 5.
Also

$$R_2=\left(1+\frac{1}{\alpha}\right)\sup_{\theta\in\Theta}\int|q_n^{\alpha}(\theta)(x)-q_\theta^{*\alpha}(x)|\,\mathrm{d}P_n(x)\leq\sup_{\theta\in\Theta}\left(1+\frac{1}{\alpha}\right)\frac{1}{n}\sum|q_n^{\alpha}(\theta)(X_i)-q_\theta^{*\alpha}(X_i)|$$

$$\leq cste\times\sup_{\theta\in\Theta}|q_n(\theta)(x)-q_\theta^*(x)|$$

which tends to 0 in Probability, making use of Proposition 5 .
Turn turn to R_3. We prove that the class of functions $q_\theta^{*\alpha}$ indexed by θ satisfies the three following properties: (i)It is indexed by θ in Θ, a compact subset of $\mathbb{R}^d$.(ii)

Secondly it is continuous in θ for all x in K. (iii) Thirdly the function F defined on K by $F(x) := \sup_{\theta\in\Theta} |q_\theta^{*\alpha}(x)|$ is such that

$$\int F(x)dP_0(x) < \infty.$$

Whenever these three facts hold, then

$$R_3 = \left(1 + \frac{1}{\alpha}\right) \sup_{\theta\in\Theta} \left| \int q_\theta^{*\alpha}(x)\mathrm{d}(P_n - P_0) \right|$$

tends to 0 in Probability since $\{q_\theta^{*\alpha}\}_\theta$ is G.C , making use of Lemma 1.1 in Empirical Processes:Glivenko-Cantelli Theorems by Moulinath Banerjee (see also J Wellner, Chapter 1.6, Notes on Empirical Processes, Torgnon Conference).

We now prove the second condition in Lemma 1. □

Lemma 5. For any $\varepsilon > 0$, $\inf_{|\theta-\theta_0|>\epsilon} M(\theta) > M(\theta_0)$. where $M(\theta) = R_\alpha(q_\theta^*, P_0) = \int q_\theta^{*\alpha+1}(x)\mathrm{d}x - \left(1 + \frac{1}{\alpha}\right) \int q_\theta^{*\alpha}(x)\mathrm{dP}_0(x)$

Proof. Denote $q_{\theta_0}^*$ the projection of P_0 on $\mathcal{M}_E$, thus $\theta_0 := \arg\inf_{\theta\in\Theta} R_\alpha(q_\theta^*, P_0)$ .For any $\theta \in \Theta$, let $.q_\theta^*$ be the projection of P_0 on $\mathcal{M}_{\theta_E}$; hence

$$R_\alpha(q_\theta^*, P_0) \geq R_\alpha(q_{\theta_0}^*, P_0)$$

We prove that equality cannot hold in the above display. Let $|\theta - \theta_0| > \epsilon$. Assume that there exists some θ_1 with

$$d(q_{\theta_1}^*, q_{\theta_0}^*) > \delta$$

such that

$$R_\alpha(q_{\theta_1}^*, P_0) = R_\alpha(q_{\theta_0}^*, P_0) \tag{6.10}$$

we can not have equality above because θ_0^* achieves the minimum of $R_\alpha(q_\theta^*, P_0)$ on θ, and $q \longrightarrow R_\alpha(q, P_0)$ is strictly convex. So (6.10) cannot hold. □

We also prove the third condition in Lemma 1.

Lemma 6. $M_n(\theta) \leq M(\theta) + o_p(1)$.

Proof. Define by $M_n(\theta) = R_\alpha(q_n(\theta), P_n)$ and $M(\theta) = R_\alpha(q_\theta^*, P_0)$ .Hence $M_n(\theta) < R_\alpha(q_\theta^*, P_n)$ by definition.
Since $R_\alpha(q_\theta^*, P_n) - R_\alpha(q_\theta^*, P_0) \xrightarrow{P} 0$ by Glivenko Cantelli Theorem, it follows that

$$M_n(\theta) \leq R_\alpha(q_\theta^*, P_0) + \eta_n$$

for n large enough,where $\eta_n \xrightarrow{P} 0$. □

As a consequence of the above arguments, the following convergence result for the minimization of power type divergences on semiparametric models defined by moment conditions holds.

Theorem 6. Under all the above conditions, it holds, whenever P_0 belongs to $\mathcal{M}$ or P_0 belongs to $\mathcal{M}_E$, with corresponding θ_0,

$$\lim_{n\to\infty} D_\alpha(\mathcal{M}, P_n) \to 0$$

and

$$\lim_{n\to\infty} \widehat{\theta}_n = \theta_0$$

Also we get

$$\lim_{n\to\infty} d\left(q_{\widehat{\theta}}, p_{\theta_0}\right) = 0$$

and all convergences above hold in probability.

References

Ali, S. M.; Silvey, S. D. A general class of coefficients of divergence of one distribution from another. J. Roy. Statist. Soc. Ser. B 28 (1966), 131–142.

Basu, Ayanendranath; Harris, Ian R; Hjort, Nils L; Jones, M. C.Robust and efficient estimation by minimising a density power divergence.Biometrika 85 (1998), no. 3, 549–559..

Broniatowski, Michel; Keziou, Amor. Minimization of ■-divergences on sets of signed measures. Studia Sci. Math. Hungar. 43 (2006), no. 4, 403–442.

Broniatowski, Michel; Keziou, AmorParametric estimation and tests through divergences and the duality technique.J. Multivariate Anal. 100 (2009), no. 1, 16–36.

Broniatowski, Michel; Stummer, Wolfgang Some universal insights on divergences for statistics, machine learning and artificial intelligence. Geometric structures of information, 149–211, Signals Commun. Technol., Springer, Cham, 2019. Minimization of ϕ-divergences on sets of signed measures. *Studia Sci. Math. Hungar.; arXiv:1003.5457*, **43**(4), 403–442.

Broniatowski, M. and Keziou, A. (2012). Divergences and duality for estimation and test under moment condition models. *J. Statist. Plann. Inference*, **142**(9), 2554–2573.

Broniatowski, Michel; Vajda, Igor Several applications of divergence criteria in continuous families. Kybernetika (Prague) 48 (2012), no. 4, 600–636.

Chen, J. H. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, **80**(1), 107–116.

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

Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **8**, 85–108.

Csiszár, I. (1967). On topology properties of f-divergences. *Studia Sci. Math. Hungar.*, **2**, 329–339.

Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation. *J. Statist. Plann. Inference*, **22**(2), 137–172. With discussion and a reply by the authors.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Ann. Statist.*, **12**(3), 971–988.

Hansen, L., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, **14**, 462–2800.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**(4), 1029–1054.

Hoff, P. D. (2012). Equivariant estimation. *Preprint*.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Rev. Econom. Stud.*, **64**(3), 359–383.

Jurečková, J. and Picek, J. (2009). Minimum risk equivariant estimator in linear regression model. *Statist. Decisions*, **27**(1), 37–54.

Kuk, A. Y. C. and Mak, T. K. (1989). Median estimation in the presence of auxiliary information. *J. Roy. Statist. Soc. Ser. B*, **51**(2), 261–269.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Liese, F. and Vajda, I. (1987). *Convex statistical distances*, volume 95. BSB B. G. Teubner Verlagsgesellschaft, Leipzig.

McCullagh, P. and Nelder, J. A. (1983). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, **72**(1), 219–255.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**(1), 90–120.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**(2), 237–249.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall, New York.

Pardo, L. (2006). *Statistical inference based on divergence measures*, volume 185 of *Statistics: Textbooks and Monographs*. Chapman & Hall/CRC, Boca Raton, FL.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**(1), 300–325.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton University Press, Princeton, N.J.

Sheehy, A. (1987). Kullback-Leibler constrained estimation of probability measures. *Report, Dept. Statistics, Stanford Univ.*

Smith, R. J. (1997). Alternative semi-parametric likelihood approches to generalized method of moments estimation. *Economic Journal*, **107**, 503–519.

Toma, Aida; Broniatowski, Michel Dual divergence estimators and tests: robustness results. J. Multivariate Anal. 102 (2011), no. 1, 20–36. 62F03

van der Vaart, A. W. Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, Cambridge, 1998. xvi+443 pp.

LPSM, Sorbonne Université, CNRS UMR 8001, Paris, France