INFERENCE IN φ-FAMILIES OF DISTRIBUTIONS

Bruno PELLETIER

IRMAR, Université Rennes 2, CNRS, UEB
Campus Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France

bruno.pelletier@.univ-rennes2.fr

Abstract

This paper is devoted to the study of the parametric family of multivariate distributions obtained by minimizing a convex functional under linear constraints. Under certain assumptions on the convex functional, it is established that this family admits an affine parametrization, and parametric estimation from an i.i.d. random sample is studied. It is also shown that the members of this family are the limit distributions arising in inference based on empirical likelihood. As a consequence, given a probability measure μ_0 and an i.i.d. random sample drawn from μ_0, nonparametric confidence domains on the generalized moments of μ_0 are obtained.

Index Terms — Parametric statistics, Maximum entropy, φ-divergence, empirical likelihood, generalized moment.

AMS 2000 Classification: 62F10, 62G05.

1 Introduction

Exponential families of distributions cover a large number of useful distributions and their properties have been widely studied (see e.g., Barndorff Nielsen, 1978). It is well known that an exponential family of distributions may be derived by maximizing the entropy under several moments constraints. The entropy, also called the *relative entropy* or the *Shannon entropy* $I(\mu)$, of a probability measure μ on a space $\mathcal{X}$ is defined by

$$I(\mu) = -\int_{\mathcal{X}} \log \frac{d\mu}{d\mu_0}(x)\mu_0(dx),$$

where μ_0 is a reference measure. In this definition, the entropy may take infinite values when μ is not absolutely continuous with respect to μ_0.

The negative entropy, i.e. $-I(\mu)$, is a convex functional in its argument μ. Several types of other (negative) entropy-like convex functionals have been defined and used mainly in the context of linear inverse problems and moments problems (Borwein and Lewis, 1991, 1993a, 1993b; Dacunha-Castelle and Gamboa, 1990; Decarreau *et al*, 1992; Gamboa and Gassiat, 1997). In these problems, the objective is to reconstruct an unknown measure μ from the observation y of *generalized moments* of μ, or Φ-*moments* of μ, i.e., the data y is related to μ by

$$y = \int_{\mathcal{X}} \Phi(x)\mu(dx), \tag{1.1}$$

where Φ is a known map from $\mathcal{X}$ to $\mathbb{R}^k$. Recovering the measure μ from the data y is an ill-posed inverse problem in the sense that a solution may not exist for every y in $\mathbb{R}^k$ (e.g., in the case of perturbed data), and if a solution exists, it may not be unique nor depends continuously on the data. In the field of inverse problems, regularization methods are very popular to cope with these issues. In particular, regularization by entropy amounts at minimizing a negative entropy-like convex functional $I_\varphi(\mu)$ over all measures μ subject to the constraint (1.1). The convex functional I_φ is defined by

$$I_\varphi(\mu) = \int_{\mathcal{X}} \varphi\left(\frac{d\mu}{d\mu_0}(x)\right)\mu_0(dx), \tag{1.2}$$

where φ is a convex function on $\mathbb{R}$. Under certain conditions on φ and the data y, Borwein and Lewis (1991, 1993a, 1993b) have shown that the problem of minimizing $I_\varphi(\mu)$ subject to the constraint (1.1) admits a unique solution $\hat{\mu}$ which may be written as

$$\hat{\mu} = \varphi^{*'}\left(\langle\omega, \Phi(x)\rangle\right)\mu_0, \tag{1.3}$$

where $\varphi^{*'}$ is the derivative of the Fenchel-Legendre transform of φ, and where ω is a vector of scalar parameters obtained as the unique solution to a dual optimization problem; see also Csiszár (1995) and Léonard (2003).

The present paper focuses on the family of probability measures which are in the form of (1.3), further referred to as a φ-*family*. These measures have been considered in Dacunha-Castelle and Gamboa (1988) in the case where $\mathcal{X}$ is a compact set and where φ is of negative type. Since φ-families contain exponential families for an appropriate choice of φ, these distributions could be used to form generalized linear models (McCullagh and Nelder, 1983); see also Pardo and Pardo (2008) for inference in generalized linear models with φ-divergences. They also arise as the limit distributions in inference based on empirical likelihood, under certain conditions on the function φ which turn the functional (1.2) into a φ-divergence (Liese and Vajda, 1987; Keziou, 2003; Broniatowski and Keziou, 2006, Pardo, 2006). To see this, let μ_0 be a probability measure, and suppose that we are interested in μ_0 only through its Φ-moment $y_0 = \int_{\mathcal{X}} \Phi(x)\mu_0(dx)$. The classical nonparametric estimator of y_0 is $\int_{\mathcal{X}} \Phi(x)\mathbb{P}_n(dx)$, where $\mathbb{P}_n$ is the empirical measure of the random sample. Its asymptotic properties are well known and can be used to construct a confidence domain, yet this latter depends on the estimation of y_0. One alternative is the method of empirical likelihood proposed by Owen (1988, 2001), by considering empirical likelihood ratios with respect to the empirical measure. As exposed for instance in Bertail (2006), the method of empirical likelihood amounts at minimizing the Kullback-Leibler divergence $K(\mu; \mathbb{P}_n)$ between the empirical measure $\mathbb{P}_n$ of the random sample, and a measure $\mu \ll \mathbb{P}_n$ satisfying the contraints of the model. In this display, the statistic

$$T_n(y) = \inf \left\{ K(\mu; \mathbb{P}_n) \; : \; \mu \ll \mathbb{P}_n \text{ and } \int_{\mathcal{X}} \Phi(x)\mu(dx) = y \right\} \qquad (1.4)$$

is used to test for y_0 as well as to construct a nonparametric confidence domain on y_0. Recently, several authors (Keziou, 2003; Broniatowski, 2004; Bertail, 2006; Browniatowski and Keziou, 2006) have proposed to use other convex statistical divergences in the form of (1.2) rather than the Kullback-Leibler divergence. This leads to alternative statistics in the form of (1.4) which are intimately related to the φ-family considered herein. Indeed, as exposed further in the paper, for a feasible y, the infimum in (1.4) is attained by a random discrete measure which converges to a member of the φ-family, i.e., a probability measure in the form of (1.3).

The paper is organized as follows. The φ-family of distributions is introduced in Section 2. In Section 3, we show that the φ-family admits an affine parametrization. Section 4 is devoted to the estimation of the affine parameter of a member of the family from an i.i.d. random sample. In Section 5, we show that the φ-family is the limit family of distributions arising in empirical likelihood. Next, nonparametric confidence domains on the Φ-moment of the underlying probability measure are derived. Technical results are postponed in an Appendix, at the end of the paper.

2 Notation and definitions

Let $(\mathcal{X}, \mu_0)$ be a finite measure space, where $\mathcal{X}$ is a measurable subset of $\mathbb{R}^d$. Let $\Phi_1, \ldots, \Phi_k$ be k functions in $L_2(\mathcal{X}, \mu_0)$ such that the maps $1, \Phi_1, \ldots, \Phi_k$ are linearly independent. We shall denote by $\Phi = (\Phi_1, \ldots, \Phi_k)$ the map $\mathcal{X} \to \mathbb{R}^k$, and by $\tilde{\Phi} = (1, \Phi_1, \ldots, \Phi_k)$ the map $\mathcal{X} \to \mathbb{R}^{k+1}$. The set of finite measures and probability measures on $\mathcal{X}$ will be denoted respectively by $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}_1^+(\mathcal{X})$.

Let $\varphi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be an extended function satisfying the following assumption.

Assumption 1

(i) $\mathrm{dom}(\varphi) = (0, +\infty)$,

(ii) φ is strictly convex and essentially smooth,

(iii) φ is $\mathcal{C}^2$ on the interior of $\mathrm{dom}(\varphi)$.

We recall that a proper convex function φ is said to be *essentially smooth* if it is differentiable on the interior of its domain, supposed non empty, and if $|\nabla\varphi(x_i)| \to \infty$ whenever x_i is a sequence converging to a boundary point of $\mathrm{dom}(\varphi)$ (Rockafellar, 1970, Chap. 26). Note that since $\mathrm{dom}(\varphi) = (0, +\infty)$, we have $\varphi(x) = +\infty$ for all $x < 0$, and that the Fenchel-Legendre transform of φ, further denoted by φ^*, may be written as

$$\varphi^*(u) = \sup_{x \geq 0}\{xu - \varphi(x)\}.$$

From this definition, it follows that φ^* is monotone increasing, so that its derivative $\varphi^{*'} \geq 0$. Under Assumption 1, we have $\mathrm{dom}(\varphi^*) = (-\infty; \kappa)$, where κ is the real number defined by

$$\kappa := \lim_{x \to \infty} \frac{\varphi(x)}{x}.$$

The limit in the equation above exists; see e.g., Lemma 2.1 in Borwein and Lewis (1993a). The essential smoothness of φ implies that φ^* is strictly convex. At last,

${\varphi^*}'$ is invertible with $({\varphi^*}')^{-1} = \varphi'$.

As explained in the Introduction, the aim of this paper is to study the family of measures minimizing the convex functional I_φ defined in (1.2) under the moments constraints (1.1). Solutions to this problem have been obtained by Borwein and Lewis (1991) (see also Borwein and Lewis, 1993a, 1993b). More precisely, we have the following result.

Theorem 2.1 *Let φ be a strictly convex function satisfying Assumption 1, and let $\tilde{y} \in \mathbb{R}^{k+1}$. Consider the following primal problem:*

$$\begin{aligned} \text{Minimize} \quad & I_\varphi(\mu) := \int_{\mathcal{X}} \varphi\left(\frac{d\mu}{d\mu_0}(x)\right)\mu_0(dx) \\ \text{subject to} \quad & \mu \in \mathcal{M}(\mathcal{X}) \quad \mu \ll \mu_0 \\ \text{and} \quad & \int_{\mathcal{X}} \tilde{\Phi}(x)\mu(dx) = \tilde{y}. \end{aligned}$$

Suppose that there exists at least one solution $\bar{\mu}$ with $I_\varphi(\bar{\mu})$ finite. Let $\bar{u}$ be the unique solution of the dual problem:

$$\begin{aligned} \text{Maximize} \quad & \langle \tilde{y}, u\rangle - \int_{\mathcal{X}} \varphi^*\left(\langle u, \tilde{\Phi}(x)\rangle\right)\mu_0(dx) \\ \text{subject to} \quad & u \in \mathbb{R}^{k+1}. \end{aligned}$$

Suppose that $\operatorname{ess\,sup} \langle \bar{u}, \tilde{\Phi}(x)\rangle < \kappa$*. Then the unique optimal solution of the primal problem is given by*

$$\bar{\mu} = {\varphi^*}'\left(\langle \bar{u}, \tilde{\Phi}(x)\rangle\right)\mu_0,$$

with dual attainment.

We are now in a position to define the φ-family of probability measures. To this aim, consider the parametric family $\tilde{\mathcal{F}}$ of finite measures on $\mathcal{X}$ defined by

$$\tilde{\mathcal{F}} = \left\{\tilde{\mu}_{\tilde{\xi}} := {\varphi^*}'\left(\langle \tilde{\xi}, \tilde{\Phi}(x)\rangle\right)\mu_0\,;\, \tilde{\xi} \in \tilde{\Xi}\right\}, \tag{2.1}$$

where

$$\tilde{\Xi} = \left\{\tilde{\xi} \in \mathbb{R}^{k+1} \,:\, \operatorname{ess\,sup} \langle \tilde{\xi}, \tilde{\Phi}(x)\rangle < \kappa\right\}, \tag{2.2}$$

where the essential supremum is taken with respect to μ_0. For all $\tilde{\xi}$ in $\tilde{\Xi}$, the Radon-Nikodym derivative of $\tilde{\mu}_{\tilde{\xi}}$ with respect to μ_0 is in $L_\infty(\mathcal{X}, \mu_0)$ by Lemma A.1. Then we define the *φ-family* $\mathcal{F}$ as the set of probability measures in $\tilde{\mathcal{F}}$, i.e., we set

$$\mathcal{F} = \tilde{\mathcal{F}} \cap \mathcal{M}_1^+(\mathcal{X}). \tag{2.3}$$

Some examples of possible choices for the convex function φ satisfying Assumption 1 are provided below.

Example 2.1 *Consider the function φ defined by*

$$\varphi(x) = \begin{cases} x\log(x) - x + 1, & \text{if } x > 0, \\ 1 & \text{if } x = 0, \\ +\infty & \text{if } x < 0. \end{cases}$$

We have $\mathrm{dom}(\varphi) = (0, +\infty)$ *and* $\kappa = +\infty$*. The convex conjugate of φ is given by* $\varphi^*(u) = \exp(u) - 1$ *and* $\mathrm{dom}(\varphi^*) = \mathbb{R}$*. Then* $\varphi^{*'}(u) = \exp(u)$ *and the family $\mathcal{F}$ is therefore an exponential family. Also in this case, the functional I_φ corresponds to the Kullback-Leibler divergence when restricted to probability measures arguments.*

Example 2.2 *Consider the function φ defined by*

$$\varphi(x) = \begin{cases} 2\left(\sqrt{x} - 1\right)^2 & \text{if } x \geq 0, \\ +\infty & \text{if } x < 0. \end{cases}$$

We have $\mathrm{dom}(\varphi) = (0, +\infty)$ *and* $\kappa = 2$*. The convex conjugate of φ is given by*

$$\varphi^*(u) = \begin{cases} \frac{2u}{2-u} & \text{if } u < 2, \\ +\infty & \text{if } u \geq 2. \end{cases}$$

We have $\mathrm{dom}(\varphi^*) = (-\infty, 2)$*, and* $\varphi^{*'}(u) = \frac{4}{(2-u)^2}$ *on* $(-\infty, 2)$*. When restricted to probability measures arguments, I_φ corresponds to the Hellinger distance, up to a multiplicative constant: the squared Hellinger distance between two probability measures μ_1 and μ_2 is defined by* $d_H(\mu_1, \mu_2)^2 = \frac{1}{2}\int \left(\sqrt{\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_0}} - \sqrt{\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_0}}\right)^2 \mathrm{d}\mu_0$*, where μ_0 dominates μ_1 and μ_2. Moreover,* $d_H(\mu_1, \mu_2)$ *does not depend on the choice of the dominating measure.*

3 Parametrization of $\mathcal{F}$

Consider the set $\tilde{S}$ of $\tilde{\Phi}$-moments of the measures in $\tilde{\mathcal{F}}$, i.e.,

$$\tilde{S} = \left\{ \int_{\mathcal{X}} \tilde{\Phi}(x)\tilde{\mu}_{\tilde{\xi}}(dx) \,:\, \tilde{\xi} \in \tilde{\Xi} \right\}. \tag{3.1}$$

Theorem 3.1 *Suppose that Assumption 1 holds. The map $\tilde{\Psi} : \tilde{\Xi} \to \tilde{S}$ defined by*

$$\tilde{\Psi}(\tilde{\xi}) = \int_{\mathcal{X}} \tilde{\Phi}(x)\tilde{\mu}_{\tilde{\xi}}(dx)$$

is a diffeomorphism from $\tilde{\Xi}$ to $\tilde{S}$.

Proof. Clearly $\tilde{\Psi}$ is surjective, and differentiable from Lemma A.2. Now we proceed to show that $\tilde{\Psi}$ is injective. Consider the map $U : \tilde{\Xi} \to \mathbb{R}$ defined by

$$U(\tilde{\xi}) = \int_{\mathcal{X}} \varphi^* \left(\langle \tilde{\xi}, \tilde{\Phi}(x) \rangle \right) \mu_0(dx).$$

Note that $U(\tilde{\xi})$ is well-defined for all $\tilde{\Xi}$ by Lemma A.1, and differentiable from Lemma A.2. Then the $\tilde{\Phi}$-moments of $\tilde{\mu}_{\tilde{\xi}}$ are obtained by differentiating U, i.e., we have

$$\tilde{\Psi}(\tilde{\xi}) = \int_{\mathcal{X}} \tilde{\Phi}(x)\tilde{\mu}_{\tilde{\xi}}(dx) = \nabla U(\tilde{\xi}).$$

Clearly, U is strictly convex since φ^* is strictly convex. Consequently the gradient map $\tilde{\xi} \to \nabla U(\tilde{\xi})$ is injective and so is $\tilde{\Psi}$.

There remains to show that $\tilde{\Psi}^{-1}$ is differentiable. To this aim, consider the map $H : \tilde{\Xi} \times \tilde{S} \to \mathbb{R}^{k+1}$ defined by

$$H(\tilde{\xi}; \tilde{y}) = \nabla U(\tilde{\xi}) - \tilde{y},$$

so that $\tilde{\psi}^{-1}(\tilde{y})$ is the unique solution (in $\tilde{\xi}$) of the equation $H(\tilde{\xi}, \tilde{y}) = 0$. Differentiating H with respect to $\tilde{\xi}$, we obtain

$$\frac{\partial}{\partial \tilde{\xi}_i} H_j(\tilde{\xi}; \tilde{y}) = \int_{\mathcal{X}} \tilde{\Phi}_i(x)\tilde{\Phi}_j(x)\varphi^{*''} \left(\langle \tilde{\xi}, \tilde{\Phi}(x) \rangle \right) \mu_0(dx),$$

where $(H_j)_{j=1,\ldots,k+1}$ are the components of H. Note that, for all $\tilde{\xi} \in \tilde{\Xi}$, the integral above is finite since the map $x \mapsto \varphi^{*''} \left(\langle \tilde{\xi}, \tilde{\Phi}(x) \rangle \right)$ is in $L_\infty(\mathcal{X}, \mu_0)$ by

Lemma A.1, and since the components of $\tilde{\Phi}$ are in $L_2(\mathcal{X}, \mu_0)$. Furthermore, $\varphi^{*''}$ is strictly positive by the strict convexity of φ^*, so that the matrix $\left(\frac{\partial}{\partial \tilde{\xi}_i} H_j(\tilde{\xi}; \tilde{y})\right)_{i,j}$ is the Gram matrix of the scalar products of the maps $1, \Phi_1, \ldots, \Phi_k$ w.r.t. the measure $\varphi^{*''}\left(\langle \tilde{\xi}, \tilde{\Phi}(x) \rangle\right) \mu_0$. Since these latter are linearly independent, the above matrix is positive-definite. Consequently, for all $(\tilde{\xi}, \tilde{y})$, $D_{\tilde{\xi}} H(\tilde{\xi}, \tilde{y})$ is a linear invertible map. The continuity and differentiability of $\tilde{\Psi}^{-1}$ then follow from the Implicit Function Theorem (see e.g., Bredon, 1993, Chap. 2). □

Now let

$$\Xi = \left\{ \tilde{\xi} \in \tilde{\Xi} \,:\, \tilde{\mu}_{\tilde{\xi}}(\mathcal{X}) = 1 \right\}. \tag{3.2}$$

and let $i_\Xi : \Xi \to \mathbb{R}^{k+1}$ be the canonical embedding of Ξ in $\mathbb{R}^{k+1}$. Then we may rewrite the family $\mathcal{F}$ as

$$\mathcal{F} = \left\{ \mu_\xi := \varphi^{*'}\left(\langle i_\Xi(\xi), \tilde{\Phi}(x) \rangle\right) \mu_0 \,:\, \xi \in \Xi \right\}. \tag{3.3}$$

Let

$$S = \left\{ \int_{\mathcal{X}} \Phi(x) \mu_\xi(dx) \,:\, \xi \in \Xi \right\}. \tag{3.4}$$

As an immediate consequence of the Theorem above, we obtain the following result.

Theorem 3.2 *Suppose that Assumption 1 holds. The map $\Psi : \Xi \to S$ defined by*

$$\Psi(\xi) = \int_{\mathcal{X}} \Phi(x) \mu_\xi(dx)$$

is a diffeomorphism from Ξ to S.

We are now in a position to provide an affine parametrization of the family $\mathcal{F}$.

Theorem 3.3 *Suppose that Assumption 1 holds. There exists a unique subset Θ of $\mathbb{R}^k$ diffeomorphic to Ξ and a unique differentiable map $g : \Theta \to \mathbb{R}$ such that*

$$\mathcal{F} = \left\{ \mu_\theta := \varphi^{*'}\left(g(\theta) + \langle \theta, \Phi(x) \rangle\right) \mu_0 \,;\, \theta \in \Theta \right\}.$$

Proof Let us write $\tilde{\xi} \in \tilde{\Xi} \subset \mathbb{R}^{k+1}$ as $\tilde{\xi} = (\alpha, \beta)$, with $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^k$ such that we have

$$\varphi^{*'}\big(\langle \tilde{\xi}, \tilde{\Phi}(x)\rangle\big) = \varphi^{*'}\big(\alpha + \langle \beta, \Phi(x)\rangle\big).$$

Furthermore, let π_1 and π_2 be the projections on respectively $\mathbb{R}$ and $\mathbb{R}^k$, i.e., $(\alpha, \beta) = (\pi_1(\tilde{\xi}), \pi_2(\tilde{\xi}))$ and let $F : \pi_1(\tilde{\Xi}) \times \pi_2(\tilde{\Xi}) \to \mathbb{R} \cup \{+\infty\}$ be the map defined by

$$F(\alpha, \beta) = \int_{\mathcal{X}} \varphi^{*'}\big(\alpha + \langle \beta, \Phi(x)\rangle\big)\mu_0(dx) - 1.$$

Note that F takes infinite values on the complement of $\tilde{\Xi}$ in $\pi_1(\tilde{\Xi}) \times \pi_2(\tilde{\Xi})$ and that we have

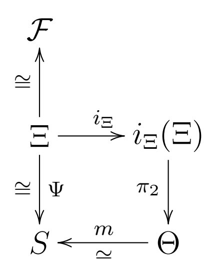
$$\Xi = \big\{(\alpha, \beta) \,:\, F(\alpha, \beta) = 0\big\}.$$

First we have

$$\begin{aligned} \frac{\partial}{\partial \alpha} F(\alpha, \beta) &= \int_{\mathcal{X}} \varphi^{*''}\big(\alpha + \langle \beta, \Phi(x)\rangle\big)\mu_0(dx) \\ &> 0 \end{aligned}$$

since φ^* is strictly convex. Hence for all (α, β), $D_\alpha F(\alpha, \beta)$ is a linear invertible map from $\pi_1(\tilde{\Xi})$ to itself. Second, Ξ is connected since Ξ is homeomorphic to S by Theorem 3.2 and S is connected. The existence and uniqueness of the map g now follows from a global version of the Implicit Function Theorem (see e.g., Dieudonné, 1972, pp. 265-266, or Blot, 1991) and is defined on $\Theta := \pi_2(\tilde{\Xi})$ which is diffeomorphic to Ξ. $\square$

As in the proof of Theorem 3.3, we shall write $\mathbb{R}^{k+1}$ as $\mathbb{R} \times \mathbb{R}^k$ and denote by π_1 and π_2 the projections from $\mathbb{R}^{k+1}$ on $\mathbb{R}$ and $\mathbb{R}^k$, respectively. Then we have the following diagram:



where i_Ξ denotes the canonical embedding of Ξ in $\mathbb{R}^{k+1}$, and where $\cong$ denotes a diffeomorphism. In this diagram, the map m is a diffeomorphism from Θ to S

and is defined by

$$m(\theta) = \int_{\mathcal{X}} \Phi(x)\mu_\theta(dx)\,, \tag{3.5}$$

i.e., m is the inverse of map of $\pi_2 \circ i_\Xi \circ \Psi^{-1}$.

4 Inference in $\mathcal{F}$

In this section, we consider the estimation of a parameter $\theta_0 \in \Theta$ based on an i.i.d. random sample $X_1, \ldots, X_n$ drawn from μ_{θ_0}, which may be written as $\mu_{\theta_0} = \varphi^{*'}\big(g(\theta_0) + \langle \theta_0, \Phi(x)\rangle\big)\mu_0$ from Theorem 3.3.

Let us start by drawing some consequences of the results in Section 3. If we denote y_{θ_0} the Φ-moments of μ_{θ_0}, i.e.,

$$y_{\theta_0} = \int_{\mathcal{X}} \Phi(x)\mu_{\theta_0}(dx).$$

then we have $\theta_0 = m^{-1}(\theta_0)$. In practice, though, and depending on the choice of φ, it may be difficult to derive explicit expressions for the maps g and m, apart from the special case of an exponential family. However, the results of Borwein and Lewis (1991, 1993a, 1993b) exposed in Theorem 2.1 provide one with a convenient algorithm to compute the value of θ_0 given the moment y_{θ_0}, without explicit expressions for the maps g and m. First of all, we may write $S = \pi_2\left(\tilde{S} \cap \{1\} \times \mathbb{R}^k\right)$. Consider the vector $\tilde{y}_{\theta_0} = (1, y_{\theta_0})$ in $\tilde{S}$. Then $\tilde{\Psi}^{-1}\left(\tilde{y}_{\theta_0}\right)$ lies in $i_\Xi(\Xi) \subset \tilde{\Xi}$ so we obtain

$$\theta_0 = \left(\pi_2 \circ \tilde{\Psi}^{-1}\right)\left(\tilde{y}_{\theta_0}\right).$$

Second, from the proof of Theorem 3.1, for all $\tilde{y}$ in $\tilde{S}$, $\tilde{\Psi}^{-1}(\tilde{y})$ is the unique solution to the following minimization problem:

$$\begin{array}{ll} \text{Minimize} & \displaystyle\int_{\mathcal{X}} \varphi^*\left(\langle u, \tilde{\Phi}(x)\rangle\right)\mu_0(dx) - \langle \tilde{y}, u\rangle \\ \text{subject to} & u \in \mathbb{R}^{k+1}. \end{array}$$

Consequently, θ_0 may be evaluated by taking the k last components of the unique minimum over $\mathbb{R}^{k+1}$ of the map

$$u := (u_0, ..., u_k) \mapsto \int_{\mathcal{X}} \varphi^* \left(u_0 + \sum_{i=1}^{k} u_i \Phi_i(x) \right) \mu_0(dx) - \left(u_0 + \sum_{i=1}^{k} u_i y_{\theta_0,i} \right), \tag{4.1}$$

i.e., letting $\bar{u} := (\bar{u}_0, \ldots, \bar{u}_k)$ be the unique minimum in (4.1), then $\theta_0 = (\bar{u}_1, \ldots, \bar{u}_k)$. In addition, we also have $g(\theta_0) = \bar{u}_0$. Another interest of this procedure is that the map in (4.1) is convex. So evaluating θ_0 from y_{θ_0} requires solving an unconstrained convex minimization problem for which efficient numerical algorithms are available.

These observations lead us to estimate θ_0 by minimizing the empirical version of (4.1). More precisely, let $\hat{y}_n$ be the empirical Φ-moment of μ_{θ_0} associated with the sample $X_1, \ldots, X_n$, i.e.,

$$\hat{y}_n = \frac{1}{n} \sum_{i=1}^{n} \Phi(X_i), \tag{4.2}$$

set $\tilde{y}_n = (1, \hat{y}_n)$, and let $\mathbb{P}_n$ be the empirical measure associated with the random sample. Then we define the estimate $\hat{\theta}_n$ as a minimizer over $\mathbb{R}^{k+1}$ of the map

$$u \mapsto \int_{\mathcal{X}} \varphi^* \left(\langle u, \tilde{\Phi}(x) \rangle \right) \mathbb{P}_n(dx) - \langle \tilde{y}_n, u \rangle,$$

which is the empirical version of (4.1). Indeed, $\hat{\theta}_n$ is an M-estimator, and on the probability event that $\hat{y}_n$ lies in the set S, we may write

$$\hat{\theta}_n = m^{-1}(\hat{y}_n). \tag{4.3}$$

Next, by the law of large numbers, almost surely, there exists n_0 such that for $n \geq n_0$, $\hat{y}_n$ belongs to S. Consequently, since m is a diffeomorphism from Θ to S, it follows that $\hat{\theta}_n$ converges in probability to θ_0, and since $\hat{y}_n$ is asymptotically normally distributed, it follows that $\hat{\theta}_n$ is in turn asymptotically normal. Finally, we have the following Theorem.

Theorem 4.1 *Suppose that Assumption 1 holds. The sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a normal distribution with mean 0 and covariance matrix given by*

$$\Sigma = \left[\gamma(\theta_0)\right]^{-2} \left[\mathrm{Cov}_{\mu^{\dagger}_{\theta_0}} \left(\Phi(X) \right) \right]^{-1} \mathrm{Cov}_{\mu_{\theta_0}} \left(\Phi(X) \right) \left[\mathrm{Cov}_{\mu^{\dagger}_{\theta_0}} \left(\Phi(X) \right) \right]^{-1},$$

where

$$\gamma(\theta) = \int_{\mathcal{X}} \varphi^{*''}\big(g(\theta) + \langle \theta, \Phi(x)\rangle\big)\mu_0(dx)\,,$$

and where $\mu^{\dagger}_{\theta_0}$ is the measure defined by

$$\mu^{\dagger}_{\theta_0} = \gamma(\theta_0)^{-1}\varphi^{*''}\big(g(\theta) + \langle \theta, \Phi(x)\rangle\big)\mu_0\,.$$

Proof Since $\hat{y}_n$ is asymptotically normal, and since m is a diffeomorphism, it follows from standard arguments on moment estimators (see e.g. Van der Vaart, 1998, Theorem 4.1, p. 36), that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to a normal distribution with mean 0 and covariance matrix

$$\Sigma = m'^{-1}_{\theta_0}\,\mathrm{Cov}_{\mu_{\theta_0}}\big(\Phi(X)\big)\,\big(m'^{-1}_{\theta_0}\big)^t\,,$$

where m'_{θ_0} is the derivative of m at θ_0. We have

$$\frac{\partial m_j}{\partial \theta_i}(\theta) = \int_{\mathcal{X}} \Phi_j(x)\left(\frac{\partial g}{\partial \theta_i}(\theta) + \Phi_i(x)\right)\varphi^{*''}\big(g(\theta) + \langle \theta, \Phi(x)\rangle\big)\mu_0(dx). \tag{4.4}$$

and

$$\frac{\partial g}{\partial \theta_i}(\theta) = -\frac{\int_{\mathcal{X}} \Phi_i(x)\varphi^{*''}\big(g(\theta) + \langle \theta, \Phi(x)\rangle\big)\mu_0(dx)}{\int_{\mathcal{X}} \varphi^{*''}\big(g(\theta) + \langle \theta, \Phi(x)\rangle\big)\mu_0(dx)} \tag{4.5}$$

since $\int_{\mathcal{X}} \varphi^{*'}\big(g(\theta) + \langle \theta, \Phi(x)\rangle\big)\mu_0(dx) = 1$. Reporting (4.5) in (4.4) yields the desired result. □

5 Nonparametric inference on the Φ-moment

Let $X_1, \ldots, X_n$ be an i.i.d. random sample drawn from a probability measure μ_0 on $\mathcal{X}$. Suppose that we are interested in μ_0 only through its Φ-moment $y_0 = \int_{\mathcal{X}} \Phi(x)\mu_0(dx)$. As exposed in the Introduction, the method of empirical likelihood (Owen, 1988, 2001) amounts at minimizing the Kullback-Leibler divergence between the empirical measure $\mathbb{P}_n$ of the random sample, and a measure μ satisfying the contraints of the model and absolutely continuous with respect to $\mathbb{P}_n$. Replacing the Kullback-Leibler divergence by a φ-divergence provides one with an alternative statistic to test for y_0, as well as to construct a confidence domain on y_0.

First of all, let $\mathbb{P}_n$ be the empirical measure associated with the random sample $X_1, \ldots, X_n$. Define the functional $I_\varphi^n(\mu)$ over $\mathcal{M}(\mathcal{X})$ by

$$I_\varphi^n(\mu) = \int_{\mathcal{X}} \varphi\left(\frac{d\mu}{d\mathbb{P}_n}(x)\right)\mathbb{P}_n(dx),$$

whenever $\mu \ll \mathbb{P}_n$ and set $I_\varphi^n(\mu) = +\infty$ otherwise. Observe that if $I_\varphi^n(\mu)$ is finite then μ is a discrete measure concentrated on the X_i's. Additional conditions on φ are needed to ensure that I_φ is a divergence between probability measure. More precisely, we shall need the following asumption.

Assumption 2

$\varphi(1) = 0.$

For all $y \in S$, we shall let $\tilde{y} = (1, y)$, and we consider the following primal problem:

$$\begin{aligned} \text{Minimize} \quad & I_\varphi^n(\mu) \\ \text{subject to} \quad & \mu \in \mathcal{M}(\mathcal{X}), \quad \mu \ll \mathbb{P}_n, \\ \text{and} \quad & \int_{\mathcal{X}} \tilde{\Phi}(x)\mu(dx) = \tilde{y}. \end{aligned}$$

The dual optimization problem is:

$$\begin{aligned} \text{Maximize} \quad & \langle \tilde{y}, \tilde{v}\rangle - \int_{\mathcal{X}} \varphi^*\big(\langle \tilde{v}, \tilde{\Phi}(x)\rangle\big)\mathbb{P}_n(dx) \\ \text{subject to} \quad & \tilde{v} \in \mathbb{R}^{k+1}. \end{aligned}$$

Let Ω_n be the probability event that a solution to the dual problem exists, solution further denoted by $\tilde{\xi}_n$. Then, by Theorem 2.1, on Ω_n, the unique primal solution is given by

$$\tilde{\mu}_n = \frac{1}{n}\sum_{i=1}^n \varphi^{*\prime}\Big(\langle \tilde{\xi}_n, \tilde{\Phi}(X_i)\rangle\Big)\delta_{X_i}. \tag{5.1}$$

The convergence of $\tilde{\mu}_n$ may be analysed using known results on M-estimators (see e.g., van de Geer, 2000, Chap. 12, and van der Vaart, 1998, Chap. 5). In essence, the concavity of the objective function in the dual program (i.e., the convexity of the negative objective function) is sufficient to establish the convergence of $\tilde{\xi}_n$ to

$\tilde{\xi}$ in probability, where $\tilde{\xi} = \tilde{\Psi}^{-1}(\tilde{y})$.

More precisely, since $y \in S$, we have $\tilde{y} = \int_{\mathcal{X}} \tilde{\Phi}(x)\tilde{\mu}_{\tilde{\xi}}(dx)$. Consequently, by the law of large numbers, it follows that $\mathbb{P}(\Omega_n) \to 1$ as $n \to \infty$. So on Ω_n, $\tilde{\xi}_n$ is the point of minimum of the map $\tilde{v} \mapsto \int_{\mathcal{X}} h_{\tilde{v}}(x)\mathbb{P}_n(dx)$, where

$$h_{\tilde{v}}(x) = \varphi^*\big(\langle \tilde{v}, \tilde{\Phi}(x)\rangle\big) - \langle \tilde{v}, \tilde{y}\rangle.$$

Since $\tilde{v} \mapsto h_{\tilde{v}}(x)$ is continuous and convex for μ_0-almost every x, and since by Lemma A.2, for $\varepsilon > 0$ small enough,

$$\int_{\mathcal{X}} \sup_{\tilde{v} \in B_\varepsilon(\tilde{\xi})} |h_{\tilde{v}}(x)|\tilde{\mu}_{\tilde{\xi}}(dx) < \infty,$$

where $B_\varepsilon(\tilde{\xi})$ is the Euclidean ball centered at $\tilde{\xi}$ and of radius ε, it follows that

$$\tilde{\xi}_n \to \tilde{\xi} \quad \text{in probability as}\, n \to \infty. \tag{5.2}$$

As a consequence, we obtain the convergence of $\tilde{\mu}_n$ to the member of the family $\mathcal{F}$ having Φ-moment y, which is stated below without proof.

Theorem 5.1 *Suppose that Assumption 1 and Assumption 2 hold. Then for all $y \in S$, $\tilde{\mu}_n$ converges weakly to the probability measure $\tilde{\mu}_{\tilde{\xi}}$, in probability, where $\tilde{\xi} = \tilde{\Psi}^{-1}(\tilde{y})$.*

Additionally, since $\tilde{\xi}_n$ converges in probability to ξ, by applying Theorem 5.23 in van der Vaart (1998, p. 53), we obtain:

$$\sqrt{n}\big(\tilde{\xi}_n - \tilde{\xi}\big) = -V_{\tilde{\xi}}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\tilde{\Phi}(X_i)\varphi^{*'}\big(\langle\tilde{\xi}, \tilde{\Phi}'X_i\rangle\big) - \tilde{y}\Big] + o_P(1), \tag{5.3}$$

where $V_{\tilde{\xi}}$ is the matrix defined by

$$V_{\tilde{\xi}} = \Big[\int_{\mathcal{X}} \tilde{\Phi}_i(x)\tilde{\Phi}_j(x)\varphi^{*''}\big(\langle\tilde{\xi}, \tilde{\Phi}(x)\rangle\big)\mu_0(dx)\Big]_{i,j}. \tag{5.4}$$

Now consider the statistic $T_n(y)$ defined by

$$T_n(y) = \inf\Big\{I_\varphi^n(\mu) \,:\, \int_{\mathcal{X}} \tilde{\Phi}(x)\mu(dx) = \tilde{y}\Big\}. \tag{5.5}$$

Then we have the following result, which proves that a confidence domain on the Φ-moment y_0 and a convergent test for y_0 may be based on the statistic $T_n(y)$.

Theorem 5.2 *Suppose that Assumption 1 and Assumption 2 hold. Suppose in addition that φ^* is $\mathcal{C}^3$ on $\mathbb{R}$ and that, for all j, k, l, there exists $\varepsilon > 0$ such that*

$$\sup_{\tilde{v}\in B_\varepsilon(\tilde{\xi})} \left| {\varphi^*}'''\big(\langle \tilde{v}, \tilde{\Phi}(x)\rangle\big)\tilde{\Phi}_i(x)\tilde{\Phi}_j(x)\tilde{\Phi}_l(x)\right| \leq h_{jkl}(x)$$

for some μ_0-integrable functions h_{jkl}, and where $B_\varepsilon(\tilde{\xi})$ denotes the ball centered at $\tilde{\xi}$ and of radius ε.

(i) If $y \neq y_0$, then

$$\sqrt{n}\big(T_n(y) - I_\varphi(y)\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

as $n \to \infty$, where

$$\sigma^2 = \int_{\mathcal{X}} {\varphi^*}^2\big(\langle\tilde{\xi}, \tilde{\Phi}(x)\rangle\big)\mu_0(dx) - \Big[\int_{\mathcal{X}} \varphi^*\big(\langle\tilde{\xi}, \tilde{\Phi}(x)\rangle\big)\mu_0(dx)\Big]^2.$$

(ii) If $y = y_0$, then

$$\frac{2n}{\varphi''(1)} T_n(y) \xrightarrow{\mathcal{D}} \chi^2(k),$$

as $n \to \infty$.

The second statement of Theorem 5.2 implies that an asymptotic confidence domain on y_0 of level $1-\alpha$ is given by $\{y \,:\, \frac{2n}{\varphi''(1)}T_n(y) \leq q_{1-\alpha}\}$, where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of a χ^2 distribution with k degrees of freedom. Moreover, the corresponding test for y_0 based on the statistics $\frac{2n}{\varphi''(1)}T_n(y)$ is consistent, by the first statement of Theorem 5.2.

Proof By dual attainment, we have

$$T_n(y) = \langle\tilde{\xi}_n, \tilde{y}\rangle - \frac{1}{n}\sum_{i=1}^n \varphi^*\big(\langle\tilde{\xi}_n, \tilde{\Phi}(X_i)\rangle\big).$$

Let us start with the following decomposition of the sum in the preceding equation:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\varphi^*\big(\langle\tilde{\xi}_n,\tilde{\Phi}(X_i)\rangle\big) &= \frac{1}{n}\sum_{i=1}^{n}\varphi^*\big(\langle\tilde{\xi},\tilde{\Phi}(X_i)\rangle\big)\\ &+\frac{1}{n}\sum_{i=1}^{n}\varphi^{*'}\big(\langle\tilde{\xi},\tilde{\Phi}(X_i)\rangle\big)\langle\tilde{\Phi}(X_i),\tilde{\xi}_n-\tilde{\xi}\rangle\\ &+\frac{1}{n}\sum_{i=1}^{n}\varphi^{*''}\big(\langle\tilde{\xi},\tilde{\Phi}(X_i)\rangle\big)\langle\tilde{\Phi}(X_i),\tilde{\xi}_n-\tilde{\xi}\rangle^2\\ &+R_n,\end{aligned}$$

where

$$R_n=\frac{1}{n}\sum_{i=1}^{n}\varphi^{*'''}\big(\langle\tilde{\xi}+\alpha_n(\tilde{\xi}_n-\tilde{\xi}),\tilde{\Phi}(X_i)\rangle\big)\langle\tilde{\Phi}(X_i),\tilde{\xi}_n-\tilde{\xi}\rangle^3,$$

for some $\alpha_n\in(0;1)$. Since the sequence $\sqrt{n}(\tilde{\xi}_n-\xi)$ is uniformly tight, and since for all j,k,l, the functions $x\mapsto\sup_{\tilde{v}\in B_\varepsilon(\tilde{\xi})}\left|\varphi^{*'''}\big(\langle\tilde{v},\tilde{\Phi}(x)\rangle\big)\tilde{\Phi}_i(x)\tilde{\Phi}_j(x)\tilde{\Phi}_l(x)\right|$ are dominated by some μ_0-integrable functions by assumption, we conclude that

$$nR_n=o_P(1). \tag{5.6}$$

First, suppose that $y\neq y_0$. In this case, it suffices to consider the decomposition at the order two. Set

$$\tilde{z}_n=\frac{1}{n}\sum_{i=1}^{n}\varphi^{*'}\big(\langle\tilde{\xi},\tilde{\Phi}(X_i)\rangle\big)\tilde{\Phi}(X_i)^t.$$

The Central Limit Theorem entails that the sequence $\sqrt{n}\big(\tilde{z}_n-\tilde{y}\big)$ is uniformly tight. Then we may write

$$
\begin{aligned}
T_n(y) - I_\varphi(y) &= \langle \tilde{\xi}_n, \tilde{y} \rangle - \frac{1}{n} \sum_{i=1}^{n} \varphi^*\big(\langle \tilde{\xi}_n, \tilde{\Phi}(X_i) \rangle\big) - \langle \tilde{\xi}, \tilde{y} \rangle \\
&\quad + \int_{\mathcal{X}} \varphi^*\big(\langle \tilde{\xi}, \tilde{\Phi}(x) \rangle\big) \mu_0(dx) \\
&= \langle \tilde{\xi}_n, \tilde{y} \rangle - \frac{1}{n} \sum_{i=1}^{n} \varphi^*\big(\langle \tilde{\xi}, \tilde{\Phi}(X_i) \rangle\big) - \langle \tilde{z}_n, \tilde{\xi}_n - \tilde{\xi} \rangle - o_P(1/\sqrt{n}) \\
&\quad - \langle \tilde{\xi}, \tilde{y} \rangle + \int_{\mathcal{X}} \varphi^*\big(\langle \tilde{\xi}, \tilde{\Phi}(x) \rangle\big) \mu_0(dx) \\
&= \langle \tilde{\xi}_n - \tilde{\xi}, \tilde{y} - \tilde{z}_n \rangle - \frac{1}{n} \sum_{i=1}^{n} \varphi^*\big(\langle \tilde{\xi}, \tilde{\Phi}(X_i) \rangle\big) \\
&\quad + \int_{\mathcal{X}} \varphi^*\big(\langle \tilde{\xi}, \tilde{\Phi}(x) \rangle\big) \mu_0(dx).
\end{aligned}
$$

But $\sqrt{n}\langle \tilde{\xi}_n - \tilde{\xi}, \tilde{y} - \tilde{z}_n \rangle \to 0$ in probability, and so the first statement follows from the Central Limit Theorem.

Second, suppose that $y = y_0$. Then $\tilde{\xi} = \tilde{\xi}_0$, and for all $i = 1, \dots, n$, the following relations hold:

$$
\begin{aligned}
\varphi^*\big(\langle \tilde{\xi}_0, \tilde{\Phi}(X_i) \rangle\big) &= \varphi^*\big(\varphi'(1)\big) = \varphi'(1), \\
\varphi^{*\prime}\big(\langle \tilde{\xi}_0, \tilde{\Phi}(X_i) \rangle\big) &= \varphi^{*\prime}\big(\varphi'(1)\big) = 1, \\
\varphi^{*\prime\prime}\big(\langle \tilde{\xi}_0, \tilde{\Phi}(X_i) \rangle\big) &= \varphi^{*\prime\prime}\big(\varphi'(1)\big) = \frac{1}{\varphi''(1)}.
\end{aligned}
$$

Let $\hat{y}_n = \frac{1}{n} \sum_{i=1}^{n} \Phi(X_i)$ and set $\tilde{y}_n = (1, \hat{y}_n)$. Let $\bar{V}_n$ be the matrix defined by

$$
\bar{V}_n = \frac{1}{n} \sum_{i=1}^{n} \tilde{\Phi}(X_i) \tilde{\Phi}(X_i)^t.
$$

Then we obtain

$$\begin{aligned} T_n(y) - I_\varphi(y) &= \langle \tilde{\xi}_n, \tilde{y} \rangle - \varphi'(1) - \langle \tilde{y}_n, \tilde{\xi}_n - \tilde{\xi}_0 \rangle - \frac{1}{2\varphi''(1)} (\tilde{\xi}_n - \xi)^t \bar{V}_n (\tilde{\xi}_n - \xi_0) \\ & \quad -o_P(1/n) - \langle \tilde{\xi}_0, \tilde{y} \rangle + \int_{\mathcal{X}} \varphi^* \big(\langle \tilde{\xi}_0, \tilde{\Phi}(x) \rangle \big) \mu_0(dx) \\ &= \langle \tilde{\xi}_n - \tilde{\xi}_0, \tilde{y} - \tilde{y}_n \rangle - \frac{1}{2\varphi''(1)} (\tilde{\xi}_n - \xi_0)^t \bar{V}_n (\tilde{\xi}_n - \xi_0) + o_P(1/n), \end{aligned}$$

since $\int_{\mathcal{X}} \varphi^* \big(\langle \tilde{\xi}_0, \tilde{\Phi}(x) \rangle \big) \mu_0(dx) = \varphi'(1)$. From (5.3), we have

$$\sqrt{n}(\tilde{\xi}_n - \tilde{\xi}) = -V_{\xi_0}^{-1}(\tilde{y}_n - \tilde{y_0}) + o_P(1),$$

where the matrix V_{ξ_0} is defined in (5.4). Since $\bar{V}_n \to \mathbb{E}\big[\tilde{\Phi}(X)\tilde{\Phi}(X)^t\big]$ element-wise as $n \to \infty$, and since $I_\varphi(y) = 0$ when $y = y_0$, we obtain

$$T_n(y) = \frac{\varphi''(1)}{2} (\tilde{y}_n - \tilde{y}_0)^t V_{\tilde{\xi}_0}^{-1} (\tilde{y}_n - y_0) + o_P(1/n). \qquad (5.7)$$

Letting $\Sigma = Cov_{\mu_0}\big(\Phi(X)\big)$, we may write

$$\begin{aligned} V_{\tilde{\xi}_0} &= \mathbb{E}\big[\tilde{\Phi}(X)\tilde{\Phi}(X)^t\big] \\ &= \begin{pmatrix} 1 & y_0^t \\ y_0 & \Sigma \end{pmatrix} \end{aligned}$$

Using the following relation for an invertible matrix defined by block (see e.g., Zhang, 2005):

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix},$$

we obtain the expression of the inverse of $V_{\tilde{\xi}_0}$:

$$V_{\tilde{\xi}_0}^{-1} = \begin{pmatrix} 1 + y_0^t \Sigma^{-1} y_0 & -y_0^t \Sigma^{-1} \\ -\Sigma^{-1} y_0 & \Sigma^{-1} \end{pmatrix}. \qquad (5.8)$$

Reporting (5.8) in (5.7), and since $(\tilde{y}_n - \tilde{y}_0) = (0, \hat{y}_n - y_0)$ yields

$$\frac{2n}{\varphi''(1)} T_n(y) = (\hat{y}_n - y_0) \Sigma^{-1} (\hat{y}_n - y_0) + o_P(1),$$

from which the result follows. $\square$.

A Technical Lemma

Lemma A.1 *Suppose that φ satisfies Assumption 1.*

(i) For all $p \in \{0;1;2\}$ and for all $\tilde{\xi} \in \tilde{\Xi}$, the map $f_p : \mathcal{X} \to \mathbb{R}$ defined by

$$f_p(x) = \varphi^{*(p)}\big(\langle \tilde{\xi}, \tilde{\phi}(x) \rangle\big)$$

is μ_0-integrable, where $\varphi^{*(p)}$ denotes the p^{th} derivative of φ^*.

(ii) Furthermore, for $p = 1$ or $p = 2$, f_p is in $L_\infty(\mathcal{X}, \mu_0)$.

Proof Let us start by recalling the properties of φ^*. First, since φ is essentially smooth, φ^* is strictly convex (Rockafellar, 1970), and since $\mathrm{dom}(\varphi) = (0, +\infty)$, φ^* is monotone increasing. Consequently, $\varphi^{*'}$ and $\varphi^{*''}$ are positive, and additionally, $\varphi^{*'}$ is monotone increasing. Combination of these facts entails that $\varphi^{*''}(u) \to 0$ as $u \to -\infty$. At last, $\varphi^*(u)/u \to 0$ as $u \to -\infty$ since $\inf \mathrm{dom}(\varphi) = 0$.

Given $\tilde{\xi} \in \tilde{\Xi}$, let $a = \operatorname{ess\,sup} \langle \tilde{\xi}, \tilde{\Phi}(x) \rangle < \kappa$ by definition of $\tilde{\Xi}$.

For $p = 0$, since $\varphi^*(u)/u \to 0$ as $u \to -\infty$, there exists $\alpha < 0$ such that $|\varphi^*(u)| \le |u|$ whenever $u \le \alpha$. Let

$$A = \big\{ x \in \mathcal{X} \,:\, \langle \tilde{\xi}, \tilde{\Phi}(x) \rangle \le \alpha \big\}.$$

First, for μ_0-a.e. x, we have

$$|f_0(x) \mathbf{1}_{A^c}(x)| \le \sup_{[\alpha, a]} |\varphi^*(u)| < \infty,$$

and second

$$|f_0(x) \mathbf{1}_A(x)| \le \langle |\tilde{\xi}|, |\tilde{\Phi}(x)| \rangle.$$

Since $\tilde{\Phi}$ is μ_0-integrable, and since μ_0 is finite, we conclude that f_0 is μ_0-integrable.

For $p = 1$, since $\varphi^{*'}$ is positive monotone increasing, we have $0 \le f_1(x) \le \varphi^{*'}(a)$ μ_0-a.e., and so f_1 is in $L_\infty(\mathcal{X}, \mu_0)$.

For $p = 2$, since $\varphi^{*''}$ is positive with $\varphi^{*''}(u) \to 0$ as $u \to -\infty$, we have $0 \le f_2(x) \le \sup_{u \in (-\infty, a]} \varphi^{*''}(u)$ μ_0-a.e., so f_2 is in $L_\infty(\mathcal{X}, \mu_0)$. □

Lemma A.2 *For all $p \in \{0;1;2\}$ and for all $\tilde{\xi} \in \tilde{\Xi}$, there exists $\varepsilon > 0$ and a μ_0-integrable function h such that*

$$\sup_{\tilde{v} \in B_\varepsilon(\tilde{\xi})} \left| \varphi^{*(p)}\big(\langle \tilde{v}, \tilde{\Phi}(x)\rangle\big)\right| < h(x),$$

where $B_\varepsilon(\tilde{\xi})$ is the Euclidean ball centered at $\tilde{\xi}$ and of radius ε. Moreover, for $p = 1$ or $p = 2$, h may be taken as a constant function.

Proof Choose ε small enough such that the ball is included in a cube in turn included in $\tilde{\Xi}$, and denote by $\tilde{v}_i$ the vertices of the cube, for $i = 1, \ldots, 2^{k+1}$. For all $\tilde{v} \in \tilde{\Xi}$, let $C(\tilde{v}) = \operatorname{ess\,sup} \langle \tilde{v}, \tilde{\Phi}(x)\rangle$, which is strictly less than κ by construction. Then, for all $\tilde{v} \in B_\varepsilon(\tilde{\xi})$, and for μ_0-almost every x, we have

$$\langle \tilde{\xi}, \tilde{\Phi}(x)\rangle - \varepsilon \|\tilde{\Phi}(x)\| \leq \langle \tilde{v}, \tilde{\Phi}(x)\rangle \leq \max_i C(\tilde{v}_i), \tag{A.1}$$

where $\|\tilde{\Phi}(x)\|$ denotes the Euclidean norm in $\mathbb{R}^{k+1}$, and where the upper inequality follows from the convexity of the cube. Since φ^* is monotone increasing, it follows that

$$\sup_{\tilde{v} \in B_\varepsilon(\tilde{\xi})} \left|\varphi^*\big(\langle \tilde{v}, \tilde{\Phi}(x)\rangle\big)\right| \leq \max\left\{ \left|\varphi^*\big(\max_i C(\tilde{v}_i)\big)\right| ; \left|\varphi^*\big(\langle \tilde{\xi}, \tilde{\Phi}(x)\rangle - \varepsilon\|\tilde{\Phi}(x)\|\big)\right| \right\},$$

for μ_0-a.e. x. Since μ_0 is a finite measure, it is sufficient to prove that the second term in the maximum is μ_0 integrable. As in the proof of Lemma A.1, let $\alpha \leq 0$ be such that $|\varphi^*(u)| \leq |u|$ for all $u \leq \alpha$, and let

$$A = \big\{x \in \mathcal{X} \,:\, \langle \tilde{\xi}, \tilde{\Phi}(x)\rangle - \varepsilon\|\tilde{\Phi}(x)\| \leq \alpha\big\}.$$

We have

$$\left|\varphi^*\big(\langle \tilde{\xi}, \tilde{\Phi}(x)\rangle - \varepsilon\|\tilde{\Phi}(x)\|\big)\right| \mathbf{1}_{A^c}(x) \leq \sup_{[\alpha\,;\,\max_i C(\tilde{v}_i)]} |\varphi^*(u)|,$$

and

$$\left|\varphi^*\big(\langle \tilde{\xi}, \tilde{\Phi}(x)\rangle - \varepsilon\|\tilde{\Phi}(x)\|\big)\right| \mathbf{1}_{A}(x) \leq \left|\langle \tilde{\xi}, \tilde{\Phi}(x)\rangle - \varepsilon\|\tilde{\Phi}(x)\|\right|,$$

and $\int_{\mathcal{X}} \left|\langle \tilde{\xi}, \tilde{\Phi}(x)\rangle - \varepsilon\|\tilde{\Phi}(x)\|\right| \mu_0(dx)$ is finite since the components of $\tilde{\Phi}$ are in $L_2(\mathcal{X}, \mu_0)$ and since $\mu_0(A^c) < \infty$. This proves the result for $p = 0$.

For $p = 1$, since $\varphi^{*'}$ is positive and monotone increasing, the result follows directly from (A.1).

For $p = 2$, the result follows from the fact that $\varphi^{*''}$ is positive with $\varphi^{*''}(u) \to 0$ as $u \to -\infty$ and (A.1). □

Acknowledgements

The author is indebted to an anonymous referee for a very careful reading of the paper and insightful comments.

References

[1] Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*, Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester.

[2] Bertail, P. (2006). Empirical likelihood in some semiparametric models. *Bernoulli*, **Vol. 12**, pp. 299-331.

[3] Blot, J. (1991). On global implicit functions, *Nonlinear Analysis, Theory, Methods and Applications*, **Vol. 17**, pp. 947-959.

[4] Borwein, J.M. and Lewis, A.S. (1991). Duality relationships for entropy-like minimization problems, *SIAM J. Control and Optimization*, **Vol. 29**, pp. 325-338.

[5] Borwein, J.M. and Lewis, A.S. (1993a). Partially-finite programming in L_1 and the existence of maximum entropy estimates, *SIAM J. Optimization*, **Vol. 3**, pp. 248-267.

[6] Borwein, J.M. and Lewis, A.S. (1993b). On the failure of maximum entropy reconstruction for Fredholm equations and other infinite systems. *Mathematical Programming*, **Vol. 63**, pp. 251-261.

[7] Bredon, G.E. (1993). *Topology and Geometry*, Volume 139 of *Graduate Texts in Mathematics*, Springer-Verlag, New York.

[8] Broniatowski, M. (2004). Estimation of the Kullback-Leibler divergence. *Mathematical Methods of Statistics*, **Vol. 12**, pp. 391-409.

[9] Broniatowski, M. and Keziou, A. (2006). Minimization of ϕ-divergences on sets of signed measures. *Studia Scientarum Mathematicarum Hungarica*, **Vol. 43**, pp. 403-442.

[10] Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Math. Hungar.*, **Vol. 68**, pp. 161-186.

[11] Dacunha-Castelle, D. and Gamboa, F. (1988). Maximum d'entropie, fonctions de type négatif et généralisation des familles exponentielles. Prépublications. Université d'Orsay, France.

[12] Dacunha-Castelle, D. and Gamboa, F. (1990). Maximum d'entropie et problèmes des moments. *Annales de l'Institut Henri Poincaré*, Section B, **Vol. 26**, pp. 567-596.

[13] Decarreau, A., Hilhorst, D., Lemaréchal, C. and Navaza, J. (1992). Dual methods in entropy maximization. Application to some problems in crystallography. *SIAM Journal on Optimization*, **Vol. 2**, pp; 173-197.

[14] Dieudonné, J. (1972). *Eléments d'Analyse*, Tome I, Fondements de l'Analyse Moderne. Gauthier-Vilars, Paris.

[15] Gamboa, F. and Gasiat, E. (1997). Bayesian methods and maximum entropy for ill-posed inverse problems. *The Annals of Statistics*, **Vol. 25**, pp. 328-350.

[16] Keziou, A. (2003). Dual representation of ϕ-divergence and applications. *C. R. Math. Acad. Sci. Paris*, **Vol. 336**, pp. 857-862.

[17] Léonard, C. (2003). Minimizers of energy functionals under not very integrable constraints. *Journal of Convex Analysis*, **vol. 10**, pp. 63-88.

[18] Liese, F. and Vajda, I. (1987). *Convex Statistical Distances*. Teubner Texts in Mathematics, **Vol. 95**, Leipzig.

[19] McCullagh, P., and Nelder, J.A. (1983). *Generalized Linear Models*. Monographs on Statistics and Aplied Probability. Chapman & Hall, London.

[20] Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **Vol. 75**, pp. 237-249.

[21] Owen, A.B. (2001). *Empirical Likelihood*. Chapman and Hall.

[22] Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*, Chapman & Hall/CRC.

[23] Pardo, J.A., and Pardo, M.C. (2008). Minimum ϕ-divergence estimator and ϕ-divergence statistics in generalized linear models with binary data. *Methodology and Computing in Applied Probability*, **Vol. 10**, pp. 357-379.

[24] Rockafellar, R.T. (1970). *Convex Analysis*, Princeton University Press, Princeton, New Jersey.

[25] van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

[26] van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[27] Zhang, F. (2005). *The Schur Complement And Its Applications*, Numerical Methods and Algorithms, **4**, Springer-Verlag, New York.