Existence and Characterization of Conditional Density Projections *

Ivana Komunjer† Giuseppe Ragusa‡

Abstract

In this paper we propose primitive conditions under which a projection of a conditional density onto a set defined by conditional moment restrictions exists and is unique. Moreover, we provide an analytic expression of the obtained projection. The range of applications where conditional density projections are used is wide. The derived results are potentially useful in a variety of areas including: semiparametric efficient estimation and optimal testing in (conditional) moment models, Bayesian prior determination and inference in semiparametric models, density forecasting, and simulation-based econometric analysis.

Regarding existence, we propose three different combinations of assumptions that are all sufficient to show that the projection exists and is unique. The proposed conditions exhibit a clear trade off between restrictions put on the divergence between the conditional densities and on the moment function which defines the projection set. Depending on the nature of the application, the researcher can pick and choose which set of conditions to use. Our second set of results characterizes the projection. The expression for the projected density is new though not surprising given the previously obtained results for the unconditional case. The projection is characterized by the dual of the original projection problem. In establishing the strong duality, however, we work with a constraint qualification condition that is weaker than that used by Borwein and Lewis (1991a, 1992a, 1993) in their seminal work concerning the unconditional case.

*We would like to thank the Co-Editor, Yuichi Kitamura, and two anonymous referees for their excellent comments and suggestions. We also thank the seminar participants at Rice University, Iowa State University, University of British Columbia, University of Pennsylvania, USC, University of Texas Austin, Rochester, and Joint Montréal Econometrics seminar for their feedback. Previous versions of this paper were circulated under the title "Existence and Uniqueness of Semiparametric Projections."

†Department of Economics, University of California, San Diego; 9500 Gilman Drive MC 0508; La Jolla, CA 92093-0508; komunjer@ucsd.edu.

‡Department of Economics and Finance, Luiss University; Viale Romania 32; 00197 Rome, Italy; gragusa@luiss.it.

1 Introduction

Consider the problem of inferring a function g from a prior guess f, both elements of a space $\mathcal{P}$, when the only available information is that g belongs to some subset $\mathcal{Q}$ of $\mathcal{P}$. This problem is central in applications in statistics, probability theory, information theory, machine learning, physical chemistry, and other scientific fields. A familiar example is when f and g are probability densities in $\mathcal{P}$, while $\mathcal{Q}$ is some known convex subset of that space. A general approach to the inference problem for g is to search for an element g^0 in $\mathcal{Q}$ which minimizes a (pseudo-) distance $\mathcal{D}$ to f, i.e. an element such that

$$\mathcal{D}(g^0, f) = \inf_{g \in \mathcal{Q}} \mathcal{D}(g, f). \tag{1}$$

When well defined and unique, the solution g^0 is called the *projection* of f onto the set $\mathcal{Q}$.

This paper is concerned with the problems of existence, uniqueness and characterization of g^0 when $\mathcal{P}$ is the space of all *conditional* probability densities, and the subset of interest $\mathcal{Q}$ is defined by a set of *conditional* moment restrictions. While the unconditional problem is today well understood (see, e.g., Liese and Vajda, 1987; Borwein and Lewis, 1993; Csiszár, 1995), no results have been derived that would cover the conditional case. The goal of this paper is to fill this gap, and by the same token provide new insights in the mechanics of the underlying theory, and the assumptions it needs to work.

The conditional density projections are of particular interest in many domains of application. In semiparametric models, projections are a constructive way of obtaining the least favorable parametric submodels introduced by Stein (1956). In the context of efficient estimation, Komunjer and Vuong (2009) show that the least favorable distributions naturally lead to the semiparametric efficiency bounds based on the conditional moment restrictions. Understanding under what conditions the least favorable family can be constructed by projection is useful. There is an increasing interest in specification testing in misspecified models that are defined by moment restrictions (Sawa, 1978; White, 1982, 1994; Vuong, 1989; Chor-Yiu and White, 1996; Otsu et al., 2008; Shi, 2014). The problem defining the projection is a natural metric to "measure" the distance of the moment implied measures (the projection) from the true yet unknown distribution. The conditions given in this paper may shed light on over which space this metric is defined and, as a consequence, which divergence is better suited for this kind of tests. Conditional density projections are also the cornerstone of simulation based econometric analysis. They can be used to design the so-called importance function used in the construction of the method of simulated moments estimator (see,

e.g., Gourieroux and Monfort, 1997, for an overview), or else to integrate out the latent variables from the moment restrictions (Schennach, 2014).

Extending the projection existence results to more realistic settings is also the first step in a "least informative" likelihood estimation of complex stochastic models. For instance, macroeconomic models such as the Dynamic Stochastic General Equilibrium Models (DSGE) impose restrictions on the conditional moments of macroeconomic quantities. An approach could be to derive the density that is consistent with these restrictions and use it as basis for inference in either the frequentist or the Bayesian framework. The theory developed in this paper may serve as the foundation of such an approach. Other areas where our results may be useful are optimal testing (Kitamura, 2001), methods of Bayesian prior determinations (Bernardo, 1979, 2005), Bayesian inference in semiparametric models (Zellner, 1996, 2002, 2003; Zellner and Tobias, 2001; Kim, 2002), and density forecasting (Giacomini and Ragusa, 2013). For extensive reviews of applications in econometrics and related fields to which our results may apply see Buck and Macaulay (1991) and Ullah (1996).

The paper derives two sets of results. In the first, we propose alternative combinations of conditions that ensure that the projection g^0 in (1) exists. In infinite dimensional problems such as ours, this is not a trivial problem. There is a rich literature that has investigated the existence of projections in the unconditional case, i.e. in the case where the minimization is over probability densities (or distributions) that satisfy a set of unconditional moment restrictions. A classical reference for the Kullback-Leibler pseudo-distance is Csiszár (1975). For general distances indexed by convex functions see Liese (1975), Borwein and Lewis (1991a, 1993), Csiszár (1995), and citations therein. However, none of this literature explicitly considers the conditional case.

The second set of results derived in this paper focuses on characterizing the form of the solution. As with the existence of g^0, its characterization has up to now remained unknown. Interestingly, available results in the unconditional case only focus on the form of g^0 obtained in L_1 (see, in particular, Borwein and Lewis, 1991a, 1993). Similarly, this is the case considered in Rockafellar (1971). The work by Csiszár (1995), which uses the Orlicz spaces, only addresses the issue of existence and leaves the form of g^0 unknown. In principle, there is no fundamental difficulty in characterizing the projection in the conditional case, though the treatment of the problem requires going back to the basics (for example, Rockafellar, 1974). In particular, one cannot simply extend Borwein and Lewis (1991a, 1993) to the conditional case. The reason has to do with the way the minimization problem (1) is solved. To be able to use the dual of the minimization problem in (1), Borwein and Lewis (1992a) work with a particular form of the constraint qualification condition which, as shown by both Gowda and Teboulle (1990) and Zălinescu (1999), becomes very restrictive

in infinite dimension. Our solution is to work with a weaker condition, which to the best of our knowledge, has not previously been applied in projection problems.

The remainder of the paper is organized as follows. In Section 2 we describe our setup, and introduce the concept of conditional density projection. In Section 3, we derive several results which show that the projection exists under alternative set of assumptions on the moment function and on the form of the divergence. Section 4 details the characterization of the projection. Section 5 introduces the modified the divergence and concludes. A summary of some well known concepts of convex analysis used in the paper is provided in Appendix A. The proofs of the results stated in the main text are relegated to Appendix B.

2 Setup

2.1 Preliminaries

We start with a quick overview of preliminary notions needed to set up our problem; a more thorough treatment of those is given in Appendix A.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and consider an $\mathcal{F}$-measurable random element $X : \Omega \to \mathbb{E}$ where $(\mathbb{E}, \mathcal{E})$ is a measurable space. We shall be interested in the conditional density of X given $\mathcal{G}$ where $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$; this density will be denoted by $f(\omega, x)$. The density $f : \Omega \times \mathbb{E} \to \mathbb{R}_+$ is understood to be with respect to a σ-finite dominating measure ν on $\mathbb{E}$. For instance, if ν is the Lebesgue measure, then the conditional distribution of X is continuous; if on the other hand, ν is the counting measure, then the conditional distribution of X is discrete.

The conditional density f belongs to $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ which is the space of (equivalence classes of) functions $g : \Omega \times \mathbb{E} \to \mathbb{R}$ that are $(\mathcal{G} \otimes \mathcal{E}, \mathcal{B}(\mathbb{R}))$-measurable and such that $|g|$ is $P \times \nu$-integrable.[1] In particular, since f is nonnegative valued, it belongs to the positive cone $\mathcal{P}$ of $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$. Elements of the space $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ (not necessarily nonnegative) shall be generally denoted by g. We pay particular attention to those elements $g \in L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ whose supports are included in that of f, property which we denote by $g \prec\!\!\prec f$. More formally,

$$g \prec\!\!\prec f \quad \text{if for } P\text{-a.e. } \omega \text{ and } \nu\text{-a.e. } x,\ f(\omega, x) = 0 \text{ implies } g(\omega, x) = 0.$$

This property is equivalent to the property of absolute continuity between the corresponding mea-

[1] We use the superscript ν in L_1^ν to emphasize that integrability needs to hold with respect to $P \times \nu$. Later on, we shall introduce spaces with integrability conditions with respect to $P \times \mu$ where μ is the conditional measure corresponding to the conditional density f. Such spaces will bear a superscript μ.

sures (see Appendix A for details).

2.2 Divergence

A divergence $\mathcal{D}$ in $\mathcal{P}$ is any nonnegative extended-real valued function defined on $\mathcal{P} \times \mathcal{P}$ such that $\mathcal{D}(f_1, f_2) = 0$ if and only if $f_1 = f_2$ with probability one.[2] In this paper, we further restrict the class of divergences $\mathcal{D}$ and focus on the ϕ-divergences $\mathcal{D}_\phi$ which are parameterized by a function $\phi : (0, +\infty) \to [0, +\infty)$ that satisfies the following properties.

Assumption A1. *(i) ϕ is twice continuously differentiable on $(0, +\infty)$; (ii) ϕ is strictly convex on $(0, +\infty)$; (iii) $\phi(1) = \phi'(1) = 0$; (iv) $\lim_{u \to 0^+} \phi'(u) < 0$; (v) $\lim_{u \to +\infty} \phi'(u) > 0$.*

It will be convenient to view ϕ as an extended-real valued function defined on $\mathbb{R}$ and taking values in $[0, +\infty]$ (see, e.g. p. 23 in Rockafellar, 1970). This means that the convex function ϕ being defined a priori on $(0, +\infty)$ we can extend it outside its domain by setting $\phi(u) = +\infty$ for all $u \in (-\infty, 0)$. As for the boundary value of zero, we let $\phi(0) = \lim_{u \to 0^+} \phi(u)$, knowing that this limit is possibly $+\infty$.[3] This ensures that the extension of ϕ is lower-semicontinuous on $\mathbb{R}$ (or "closed" in the terminology of Rockafellar (1970)). Note that since by Assumption A1(ii) ϕ is convex on $(0, +\infty)$, its extension is convex on $\mathbb{R}$. Further, to deal with zero and infinity we adopt the understanding that $\phi(+\infty) = \lim_{u \to +\infty} \phi(u)$, $\phi'(0) = \lim_{u \to 0^+} \phi'(u)$, $\phi'(+\infty) = \lim_{u \to +\infty} \phi'(u)$, and $0 \cdot \phi\left(\frac{0}{0}\right) = 0$.

The conjugate of the extended-real valued function ϕ will play an important role in our analysis. The conjugate will be denoted by $\phi^* : \mathbb{R} \to (-\infty, +\infty]$ where

$$\phi^*(v) \equiv \sup_{u \in \mathbb{R}} \{uv - \phi(u)\}.$$

Then ϕ^* is increasing on $\mathbb{R}$ and it is itself a convex lower semi-continuous function. Since ϕ is by Assumption A1 differentiable, we can relate ϕ^* to the Lagendre-Fenchel transform. More details on the properties of ϕ^* are given in Appendix A.

In order to formally define the ϕ-divergences between probability distributions, we first introduce an integral functional which operates on elements of $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ (not necessarily non-negative). The restriction of the latter to $\mathcal{P}$ will then be called a ϕ-divergence. The class of ϕ-divergences among probability distributions was first introduced by Ali and Silvey (1966) and Csiszár (1967).

[2] Divergences are also referred to as directed divergences, generalized entropies, relative entropy functionals, or pseudo-distances in the literature (see, e.g., Ullah, 1996).

[3] Note that the non-negativity of ϕ is also ensured by the strict convexity of ϕ on $(0, +\infty)$, and the requirements in Assumption A1(iii,iv).

Theorem 1 (ϕ-divergence)**.** *Given a function ϕ that satisfies Assumption A1 and a conditional density $f \in \mathcal{P}$, for any $g \in L_1^\nu(\mathcal{G} \otimes \mathcal{E})$, let*

$$\mathcal{D}_\phi(g, f) \equiv \begin{cases} \int f\phi\left(\frac{g}{f}\right) d(P \times \nu), & \text{if } g \ll f, \\ +\infty, & \text{otherwise.} \end{cases}$$

*Then $\mathcal{D}_\phi(\cdot, f)$ is a well defined convex function on $L_1(\mathcal{G}\otimes\mathcal{E})$ which takes values in $[0, +\infty]$. Moreover, $\mathcal{D}_\phi(\cdot, f)$ is strictly convex on its effective domain $\operatorname{dom}\mathcal{D}_\phi(\cdot, f) = \{g \in L_1^\nu(\mathcal{G} \otimes \mathcal{E}) : \mathcal{D}_\phi(g, f) < +\infty\}$, and $\mathcal{D}_\phi(g, f) = 0$ if and only if $g = f$ a.s. We call the restriction of $\mathcal{D}_\phi(\cdot, f)$ to $\mathcal{P}$ a **ϕ-divergence** on $\mathcal{P}$.*

The above theorem defines the class of ϕ-divergences. Doing so formally requires several results which need proof. First is the result that $\mathcal{D}_\phi(g, f)$ is well defined. In particular, the measurability of $f(\omega, x)\phi\big(g(\omega, x)/f(\omega, x)\big)$ needs to be established without resorting to the Carathéodory condition. The latter cannot be imposed on the extended-real valued ϕ, which we have allowed to be infinitely valued. Instead, the measurability will follow from the convexity and lower semi-continuity properties of ϕ on $\mathbb{R}$. Second is the result that $\mathcal{D}_\phi(\cdot, f)$ is a convex function. This result is quite intuitive: since $\mathcal{D}_\phi(\cdot, f)$ is an integral of a convex function ϕ, it is likely to inherit its convexity properties. The last result says that $\mathcal{D}_\phi(\cdot, f)$ has the necessary divergence properties: it is non-negative valued, and equal to zero only at f. This result is simply a consequence of the Jensen's inequality.

Since ϕ was extended to the entire real line, the ϕ-divergence $\mathcal{D}_\phi(g, f)$ is defined on the entire $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ space, even for those elements g that are possibly negative valued. The reason why we define $\mathcal{D}_\phi(\cdot, f)$ on $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ (rather than just on the positive cone $\mathcal{P}$) is that many of our arguments to follow will be stated in terms of the duals, which are easier to analyze when the domain is the entire $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ space. Note however that $\mathcal{D}_\phi(g, f) < +\infty$ only if $g \in \mathcal{P}$; in other words, the ϕ-divergence between any $g \in L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ that takes negative values and the conditional density f is always $+\infty$. Mostly however we shall be interested in ϕ-divergences among conditional densities. Then, the ϕ-divergence between say f_1 and f_2 in $\mathcal{P}$ can also be expressed in terms of the corresponding conditional measures, $\mu_1(\omega, B) = \int_B f_1(\omega, x)d\nu(x)$ and $\mu_2(\omega, B) = \int_B f_2(\omega, x)d\nu(x)$, by defining $\mathcal{D}_\phi(\mu_1, \mu_2) \equiv \int \phi(d\mu_1/d\mu_2)d(P \times \mu_2)$ if $\mu_1 \ll \mu_2$ and $\mathcal{D}_\phi(\mu_1, \mu_2) \equiv +\infty$ otherwise. The two definitions are equivalent and $\mathcal{D}_\phi(\mu_1, \mu_2) = \mathcal{D}_\phi(f_1, f_2)$. This formulation is used by Kitamura and Stutzer (1997) and Kitamura (2001), for example.[4]

[4]Note, however, that our definition of ϕ-divergence (and thus that of Kitamura and Stutzer (1997) and Kitamura (2001)) slightly differs from that of Ali and Silvey (1966) or Csiszár (1995) who allow the divergence to possibly

The class of ϕ-divergences $\mathcal{D}_\phi$ generally includes many distances used in econometrics and statistics. Of particular interest are:

(i) Kullback-Leibler distance (I-divergence) (see, e.g., Kullback and Khairat, 1966; Csiszár, 1975):

$$\phi(u) = \begin{cases} u \ln u - u + 1, & u > 0 \\ 1, & u = 0 \\ +\infty, & u < 0 \end{cases}, \quad \phi^*(v) = \exp v - 1;$$

(ii) reverse I-divergence (Burg entropy) (see, e.g., Burg, 1967):

$$\phi(u) = \begin{cases} -\ln u + u - 1, & u > 0 \\ +\infty, & u = 0 \\ +\infty, & u < 0 \end{cases}, \quad \phi^*(v) = \begin{cases} -\ln(1-v), & v < 1 \\ +\infty, & v \geqslant 1 \end{cases};$$

(iii) (squared) Hellinger distance:

$$\phi(u) = \begin{cases} -4u^{1/2} + 2u + 2, & u > 0 \\ 2, & u = 0 \\ +\infty, & u < 0 \end{cases}, \quad \phi^*(v) = \begin{cases} 2(1-v/2)^{-1} - 2, & v < 2 \\ +\infty, & v \geqslant 2 \end{cases};$$

(iv) χ^2 distance:[5]

$$\phi(u) = \begin{cases} u^2/2 - u + 1/2, & u > 0 \\ 1/2, & u = 0 \\ +\infty, & u < 0 \end{cases}, \quad \phi^*(v) = \begin{cases} v^2/2 + v, & v \geqslant -1 \\ -1/2, & v < -1 \end{cases};$$

Of particular interest is the Cressie-Read family of divergences introduced by Cressie and Read (1984) and parameterized by a real parameter a (the definition of ϕ_a is found in Equation (22) in

remain finite even if μ_1 is not dominated by μ_2. In this case, there is an indeterminacy problem caused by the singular component of μ_1 whenever $\lim_{u\to\infty} \phi(u)/u < +\infty$. We avoid this problem by defining the divergence between g and f to be infinite, whenever the support of g is not included in that of f. It is worth emphasizing that our definition does not require the support of f to be included in that of g so that we may well have regions where $g(\omega, x) = 0$ and yet $f(\omega, x) > 0$. In this case, the behavior of $\mathcal{D}_\phi$ will depend on the value of the function ϕ at zero. In particular, if $\phi(0) = +\infty$, the divergence between such g and f will become infinite.

[5]Note that our extended-real valued ϕ used in the definition of χ^2 distance differs from the definition used for example in Borwein and Lewis (1991a). Since ϕ is in this particular case well defined on $\mathbb{R}$, one possibility would be to simply extend it from $(0, +\infty)$ to $\mathbb{R}$ by using the same formula. This is the approach taken in Borwein and Lewis (1991a). We shall see later on, however, that this definition of ϕ could lead to negative densities, problem which as we shall demonstrate does not occur with our extension of ϕ.

Appendix A). The Cressie-Read family contains the χ^2 distance ($a = 1$), the I-divergence ($a \to 0$), the reverse I-divergence ($a \to -1$), and the Hellinger distance ($a = -1/2$). In the econometric literature, applications of the Cressie-Read distances can be found in Kitamura and Stutzer (1997)'s Exponential Tilting estimator, Kitamura et al. (2009)'s Minimum Hellinger Distance Estimator, and in Newey and Smith (2004)'s Generalized Empirical Likelihood. See Ragusa (2011) for example of other divergences and their applications in econometrics.

Notice that for the reverse I-divergence, for the Hellinger distance, and, in general, for members of the Cressie-Read family with $\alpha < 0$ Assumption A1(v) holds with $\lim_{u\to+\infty} \phi'(u) < +\infty$.

2.3 Projection

We are now ready to formally define conditional density projections. For this, fix $f \in \mathcal{P}$. With the ϕ-divergence as given in Theorem 1, the $\mathcal{D}_\phi$-projection of f onto a subset $\mathcal{Q}$ of $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ is defined as follows:

Definition 1. *The $\mathcal{D}_\phi$-projection of f onto a set $\mathcal{Q} \subseteq L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ is (when it exists) a $g^0 \in \mathcal{Q}$ satisfying: $\mathcal{D}_\phi(g^0, f) = \inf_{g\in\mathcal{Q}} \mathcal{D}_\phi(g, f)$.*

In most statistical and econometric applications the projection set $\mathcal{Q}$ is defined by a set of either unconditional or conditional moment restrictions. The unconditional problem is obtained when $\mathcal{G}$ is the trivial σ-field, i.e. $\mathcal{G} \equiv \{\emptyset, \Omega\}$. When $\mathcal{G}$ is any other sub-σ-field of $\mathcal{F}$ then the problem is conditional. While the unconditional problem has been extensively studied in the literature, little is known about the conditional one. Our setup accommodates both cases.

The choice of the space over which the projection is defined, $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$, merits some discussion. In principle, one could define the projection over other spaces, such as for example $L_p^\nu(\mathcal{G} \otimes \mathcal{E})$ with $p > 1$. Since it holds that $L_p^\nu(\mathcal{G} \otimes \mathcal{E}) \subset L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ for any $1 < p < +\infty$, it is natural to define the projection over the largest space.

Now, consider some known moment function $a : \Omega \times \mathbb{R}^n \to \mathbb{R}^m$ ($m \in \mathbb{N}$, $m < \infty$) that is $(\mathcal{G} \otimes \mathcal{E}, \mathcal{B}(\mathbb{R}^m))$-measurable. Note that the number of components m can be greater than one. We focus on $\mathcal{D}_\phi$-projecting f onto a set of conditional densities that satisfy the conditional moment restrictions $E[a(X)|\mathcal{G}] = 0$ (with probability one). The projection set $\mathcal{Q}$ is then characterized as follows:

$$\mathcal{Q} \equiv \left\{ g \in L_1^\nu(\mathcal{G} \otimes \mathcal{E}) : \int_{\mathbb{E}} a(\omega, x) g(\omega, x) d\nu(x) = 0 \text{ and } \int_{\mathbb{E}} g(\omega, x) d\nu(x) = 1, \text{for a.e. } \omega \right\}. \qquad (2)$$

Put in words, the set $\mathcal{Q}$ is a subset of elements in $L_1^\nu(\mathcal{G}\otimes\mathcal{E})$ that for a.e. ω integrate to one, and satisfy the moment condition $E_g[a(X)|\mathcal{G}]=0$ a.s. (the expectation being taken under g). Note that though we require that $g(\omega,\cdot)$ integrates to one for a.e. ω, we purposefully choose not to impose any non-negativity constraints on g. Those will be automatically satisfied if $\mathcal{D}_\phi(g,f)<+\infty$. In semiparametric applications, the moment function a may further be parameterized by some finite dimensional parameter θ. In that case, the projection set $\mathcal{Q}$ also depends on θ.

3 Existence

3.1 Statement of the Problem

We can now re-formulate our projection problem $\inf_{g\in\mathcal{Q}}\mathcal{D}_\phi(g,f)$ as a constrained optimization problem in $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$. Recall (c.f. footnote 1) that integrability in $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$ holds with respect to $P\times\mu$ where μ is the conditional measure corresponding to the conditional density f.

We start by assuming that there is at least one element g_0 in $\mathcal{Q}$ such that $\mathcal{D}_\phi(g_0,f)<+\infty$, i.e. that the optimization problem is feasible. Since the ϕ-divergence between g_0 and f is finite, we necessarily have $g_0\prec\!\!\prec f$. Now, for any $g\in L_1^\nu(\mathcal{G}\otimes\mathcal{E})$ such that $g\prec\!\!\prec f$, we have $\mathcal{D}_\phi(g,f)=I_\phi\left(\frac{g}{f}\right)$ where[6]

$$I_\phi(\pi)\equiv\int\phi(\pi)d(P\times\mu),\qquad \pi\in L_1^\mu(\mathcal{G}\otimes\mathcal{E}). \tag{3}$$

In addition, note that the constraints that define the projection set $\mathcal{Q}$ in (2) are linear in g; therefore by a simple change of variable $g=\pi f$ we have that $g\in\mathcal{Q}$ if and only if $\pi\in\mathcal{C}$ where we have let

$$\mathcal{C}\equiv\left\{\pi\in L_1^\mu(\mathcal{G}\otimes\mathcal{E}):\int_{\mathbb{E}}a(\omega,x)\pi(\omega,x)d\mu(x)=0\text{ and }\int_{\mathbb{E}}\pi(\omega,x)d\mu(x)=1,\text{for a.e. }\omega\right\}. \tag{4}$$

Note that $\mathcal{C}$ is a convex subset of $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$.

Our projection problem is then equivalent to a constrained optimization problem:

$$\inf_{\pi\in\mathcal{C}}I_\phi(\pi). \tag{5}$$

If a solution π^0 to the problem (5) exists, then the projection g^0 is simply obtained by setting $g^0=\pi^0 f$. Before attempting to characterize a solution to (5), we need to establish if and under what conditions a solution to the above problem exists.

[6]For a detailed derivation of this equality see the proof of Theorem 1 and Equation (23) in Appendix.

General proofs of existence are established using a *minimizing sequence* of (5), i.e. a sequence $\{\pi_n\}$ of elements of $\mathcal{C}$ such that

$$\lim_{n\to\infty} I_\phi(\pi_n) = \inf_{\pi\in\,\mathcal{C}} I_\phi(\pi) = d.$$

If there is at least one element $\pi_0 \in \mathcal{C}$ such that $I_\phi(\pi_0) < +\infty$, i.e. if the optimization problem (5) is feasible, then we know that $d \in [0, +\infty)$. The problem however is that the minimizing sequence $\{\pi_n\}$ need not have a limit $\pi^0 \in \mathcal{C}$ such that $I_\phi(\pi^0) = d$. In the classical finite dimensional case, the objective function is continuous on the projection set which is closed and bounded in say $\mathbb{R}^n$. Then, one can extract a converging subsequence from the minimizing sequence which converges to a limit in $\mathcal{C}$. That the limit of the subsequence is a solution to the problem follows by the continuity of the objective function. It is possible to extend this line of reasoning to infinite dimensional spaces. The idea is to:

(e1) establish the existence of a subsequence $\{\pi_{n_i}\}$ that converges in some sense to a limit π^0;

(e2) establish that the limit π^0 is in $\mathcal{C}$;

(e3) show that the limit is a solution by appealing to the lower-semicontinuity of I_ϕ on $\mathcal{C}$ for an appropriate topology.

A typical setup is the one in which the space under consideration is reflexive. The topology considered is the weak topology, and the notion of convergence that of a weak convergence (see, e.g., Chapter 2 in Ekeland and Témam, 1987). In a reflexive space, it is sufficient to show that the minimizing sequence $\{\pi_n\}$ is bounded in order to establish the property (e1). The problem, however, is that the spaces $L_p^\mu(\mathcal{G}\otimes\mathcal{E})$ are reflexive only if $1 < p < +\infty$. In particular $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$ is not reflexive which means that not every bounded sequence in $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$ has a weakly convergent subsequence.

3.2 Existence in L_1

We first discuss the conditions for projection existence when the space under consideration is $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$ equipped with the weak topology. As already pointed out, bounded subsets of $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$ are not necessarily weakly sequentially compact (i.e. $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$ is not reflexive). The idea is to instead use the weak compactness of the level sets of I_ϕ seen as subsets of $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$ (see, e.g., Borwein and Lewis, 1991b). This requires an additional assumption on the ϕ function that defines the divergence.

Assumption A2. $\lim_{u\to+\infty} \frac{\phi(u)}{u} = +\infty$.

Assumption A2 requires ϕ to tend to infinity faster than a linear function which is not an

innocuous restriction. In particular, it excludes several popular choices for ϕ such as the reverse I-divergence and the (squared) Hellinger distance. This assumption is however instrumental in establishing weak compactness of levels sets of I_ϕ. In a later section we will present a class of modified divergences whose behavior at $+\infty$ satisfies Assumption A2.

The proof of the following lemma adapts the arguments used by Borwein and Lewis (1991b).

Lemma 1. *Let Assumption A1 hold. Then Assumption A2 is sufficient for the level sets of I_ϕ of the form $\left\{\pi \in L_1^\mu(\mathcal{G}\otimes\mathcal{E}) : I_\phi(\pi) \leqslant l\right\}$ with $l > 0$ to be weakly sequentially compact in $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$. Moreover, if μ is not purely atomic, then Assumption A2 is also necessary.*

The result of Lemma 1 can be used to extract a weakly converging sequence from a minimizing sequence, i.e. establish the property (e1) discussed above. As shown in the lemma, Assumption A2 is not only sufficient but also necessary for the result to hold, provided the conditional measure μ is not purely atomic.[7] The key insight behind this assumption is discussed in Borwein and Lewis (1991b) upon which the proof of Lemma 1 is based. Using the results from Rockafellar (1968), Borwein and Lewis (1991b) show that weak compactness of the level sets of integral functionals such as I_ϕ is equivalent to the finiteness of the conjugate ϕ^* on $\mathbb{R}$. It follows immediately that ϕ^* is everywhere finite if and only if its effective domain equals $\mathbb{R}$ which by Lemma 4.2 in Borwein and Lewis (1991a) stated in Appendix A is equivalent to A2. Thus, the sufficiency as well as necessity of this assumption for the weak compactness of the level sets of I_ϕ follows.

We now turn to the second property, i.e. property (e2) above. By definition, when all weakly converging sequences in $\mathcal{C}$ weakly converge to the limits in $\mathcal{C}$, we say that $\mathcal{C}$ is weakly closed. Since the set $\mathcal{C}$ in (4) is convex, the property of being weakly closed is equivalent to being closed. So a simple sufficient condition for the property (e2) is to require that the set $\mathcal{C}$ be closed in the norm topology of $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$. For this, we shall work with the following assumption.

Assumption A3. *The moment function a is essentially bounded, i.e. there exists a positive constant $M < \infty$ such that for a.e. $x \in \mathbb{E}$ and a.e. ω, $|a(\omega, x)| \leqslant M$.*

According to Assumption A3, the moment function a is in $L_\infty^\mu(\mathcal{G}\otimes\mathcal{E})$, i.e. $\|a\|_\infty^\mu < +\infty$. The following lemma formally establishes the needed result.

Lemma 2. *Let Assumption A3 hold. Then the projection set $\mathcal{C}$ is closed in the norm topology of $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$.*

[7] Note that this excludes the case where μ is the empirical measure, for example.

Having established sufficient conditions for (e1) and (e2), it only remains to show (e3) that I_ϕ is lower semi-continuous in the weak topology of $L_1^\mu(\mathcal{G} \otimes \mathcal{E})$. The latter follows from Lemma 1 which implies that the level sets $\left\{\pi \in L_1^\mu(\mathcal{G} \otimes \mathcal{E}) : I_\phi(\pi) \leqslant l\right\}$ with $l > 0$ are weakly closed. Hence, we are now in position to establish the existence of projections g^0 in $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$.

Theorem 2. *Let Assumptions A1, A2, and A3 hold. Assume in addition that the problem is feasible, i.e. there exists at least one $g_0 \in \mathcal{Q}$ such that $\mathcal{D}_\phi(g_0, f) < +\infty$. Then a $\mathcal{D}_\phi$-projection of f onto $\mathcal{Q}$ exists and is unique.*

Theorem 2 shows that under Assumptions A1 through A3, feasibility of the minimization problem in (5) is sufficient to establish existence. Once existence is established, uniqueness follows by the strict convexity of the divergence $\mathcal{D}_\phi(g, f)$ on its effective domain (i.e. on the set of g's for which $\mathcal{D}_\phi(g, f)$ is finite). Combining the statements in Assumptions A2 and A3, the condition under which Theorem 2 holds can be restated as

$$\|a\|_\infty^\mu < \lim_{u \to \infty} \frac{\phi(u)}{u},$$

with the right-hand side being infinite. Interestingly, in the context of projection problems involving unconditional densities and unconditional moment restrictions, Borwein and Lewis (1991a) show that the above condition (together with Assumption A1 and a stronger feasibility condition) is sufficient to guarantee existence. Their result suggests a deep connection between the constraints put on the moment function a (Assumption A3) and those put on the function ϕ defining the divergence (Assumption A2); there is a clear trade-off between restrictions on the growth rate of ϕ and the boundedness of the moment function. Rather than attempting to generalize the result of Borwein and Lewis (1991a) to a conditional case, we proceed with existence results that obtain even if the moment function is unbounded.

3.3 Existence in Orlicz Spaces

As already pointed out, the key difficulty in establishing existence of the projection g^0 in $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ is that we are working in a space that is not reflexive. One immediate solution to this problem is to change the space under consideration. For example, instead of working in $L_1^\mu(\mathcal{G} \otimes \mathcal{E})$ one could work in $L_p^\mu(\mathcal{G} \otimes \mathcal{E})$, $1 < p < +\infty$, which is reflexive. This, however, would amount to putting stronger conditions on the conditional densities under consideration, such as for instance square integrability if one works with $p = 2$. An alternative approach is to work in a space whose structure is deeply

connected to the form of the ϕ-divergence under consideration. Such spaces are called Orlicz spaces and we first give a brief overview of key definitions and results; for details, see, e.g. the book by Krasnosel'skii and Rutickii (1961).

3.3.1 Overview of Orlicz Spaces

Let $\rho : (0, +\infty) \to \mathbb{R}$ be a continuously differentiable convex function that satisfies $\lim_{u\to 0} \rho(u)/u = 0$ and $\lim_{u\to\infty} \rho(u)/u = +\infty$. To each such function ρ we can associate the *Orlicz space* $L_\rho^\mu(\mathcal{G} \otimes \mathcal{E})$ which is the space of (equivalence classes of) functions $h : \Omega \times \mathbb{R}^n \to \mathbb{R}$ that are $(\mathcal{G} \otimes \mathcal{E})$-measurable and such that $\int \rho(\alpha_0|h|)\, d(P \times \mu) < +\infty$ for some $\alpha_0 > 0$, i.e.

$$L_\rho^\mu(\mathcal{G} \otimes \mathcal{E}) \equiv \left\{ h : \int \rho(\alpha_0|h|)\, d(P \times \mu) < +\infty, \text{ for some } \alpha_0 > 0 \right\}.$$

The space $L_\rho^\mu(\mathcal{G} \otimes \mathcal{E})$ equipped with the (Luxemburg) norm $\|\cdot\|_\rho^\mu$ given by

$$\|h\|_\rho^\mu \equiv \inf\{\beta > 0 : \int \rho\left(\frac{|h|}{\beta}\right) d(P \times \mu) \leqslant 1\},$$

is a Banach space. Note that in order to get the usual $L_1^\mu(\mathcal{G} \otimes \mathcal{E})$ space, one would need to set $\rho(u) = u$. This choice of function ρ, however, does not satisfy the limit requirement $\lim_{u\to\infty} \rho(u)/u = +\infty$. Still $L_1^\mu(\mathcal{G} \otimes \mathcal{E})$ can be viewed as the union of all the Orlicz spaces (see the discussion on p.61 of Krasnosel'skii and Rutickii (1961), and the result of Lemma 4 to follow).

Before proceeding let us show why Orlicz spaces are a natural choice of space in our problem.

Lemma 3. *Let $\{\pi_n\} \in \mathcal{C}$ be a minimizing sequence of the problem* (5)*. If the problem is feasible, i.e. if $\inf_{\pi\in\mathcal{C}} I_\phi(\pi) = d < +\infty$, then $\{\pi_n\} \in L_\rho^\mu(\mathcal{G} \otimes \mathcal{E})$ with $\rho(u) = \phi(1 + u)$.*

Lemma 3 shows that when the minimization problem (5) is feasible, any minimizing sequence is (eventually) in the Orlicz space $L_\rho^\mu(\mathcal{G} \otimes \mathcal{E})$ with ρ being chosen as $\rho(u) = \phi(1 + u)$. In this sense, working with an appropriately chosen Orlicz space does not impose any additional assumptions (as working with e.g. L_2 would) but rather fully exploits the existing feasibility condition.

Since most of our results rely on duality theory, it will prove useful to carefully define the *paired spaces* in which the conjugates of various convex functions are computed.[8] For this, consider the

[8]The notion of paired spaces has been introduced by Rockafellar (1974).

following subspace $E^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$ of $L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$,

$$E^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E}) \equiv \left\{ h : \int \rho\left(\alpha|h|\right) d(P\times\mu) < +\infty, \text{ for every } \alpha > 0 \right\}.$$

Like $L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$, its subspace $E^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$ equipped with the (Luxemburg) norm $\|\cdot\|^{\mu}_{\rho}$ is a Banach space. The pairing which we consider is

$$\langle u, v\rangle \equiv \int uvd(P\times\mu), \quad u \in L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E}),\ v \in E^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E}), \tag{6}$$

where ρ^* is the conjugate of ρ. In particular, if $\rho(u) = \phi(1+u)$, then $\rho^*(v) = \phi^*(v) - v$ where ϕ^* is the conjugate of ϕ. The quantity $\langle u, v\rangle$ in (6) is well defined in view of the Hölder inequality in Orlicz spaces:

$$\int |uv| d(P\times\mu) \leqslant 2\|u\|^{\mu}_{\rho}\|v\|^{\mu}_{\rho^*}. \tag{7}$$

In particular, $uv \in L^{\mu}_{1}(\mathcal{G}\otimes\mathcal{E})$.

The pairing $\langle u, v\rangle$ in (6) behaves much like an inner product except that the u argument is restricted to $L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$ and v to $E^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$. We now equip the two spaces with topologies such that the linear functionals $\langle\cdot, v\rangle$ on $L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$ are all continuous and every continuous linear function on $L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$ can be represented in the form $\langle\cdot, v\rangle$ for some $v \in E^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$, with an analogous result for the space $E^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$.[9]

On $L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$, we consider the E-weak topology, which is the weakest topology on $L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$ that makes all the functionals $\langle\cdot, v\rangle$, with $v \in E^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$, continuous. A functional $l : L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E}) \to \mathbb{R}$ is *E-weakly continuous* if for any E-weakly convergent sequence $\{u_n\}$, $u_n \xrightarrow{E} u_0$ implies $\lim_{n\to\infty} l(u_n) = l(u_0)$. A sequence $\{u_n\} \in L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$ is said to *E-weakly converge* to $u^0 \in L_{\rho}(\mathcal{G}\otimes\mathcal{E})$, property which we denote by $u_n \xrightarrow{E} u^0$, if $\lim_{n\to\infty}\int u_n v d(P\times\mu) = \int u^0 v d(P\times\mu)$ for every $v \in E^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$.[10] The main advantage of working with this topology on $L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$ is that every E-weakly continuous linear functional l on $L^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$ is of the form $l(u) = \langle u, v\rangle$ with $v \in E^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$ (see, e.g., Theorem 14.7 in Krasnosel'skii and Rutickii, 1961).

If we reverse the roles played by ρ and its conjugate ρ^*, then the above result also says that the functionals $l^* : L^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E}) \to \mathbb{R}$ defined by $l^*(v) = \langle u, v\rangle$ (where $u \in E^{\mu}_{\rho}(\mathcal{G}\otimes\mathcal{E})$) are precisely all the E-weakly continuous linear functionals on $L^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$. The problem is however that we are

[9] In the familiar case in which $u \in L^{\mu}_{1}(\mathcal{G}\otimes\mathcal{E})$ and $v \in L^{\mu}_{\infty}(\mathcal{G}\otimes\mathcal{E})$, the weak topologies on the two spaces satisfy this requirement.

[10] In the terminology of Krasnosel'skii and Rutickii (1961), this type of convergence is called E_N-weak convergence (see Chapter 2 §14 in Krasnosel'skii and Rutickii, 1961).

working with the space $L^\mu_\rho(\mathcal{G}\otimes\mathcal{E})$ which in general is strictly larger than $E^\mu_\rho(\mathcal{G}\otimes\mathcal{E})$, so we need to be able to characterize functionals of the form $v\mapsto\langle u,v\rangle$ where $u\in L^\mu_\rho(\mathcal{G}\otimes\mathcal{E})\supset E^\mu_\rho(\mathcal{G}\otimes\mathcal{E})$. One easy solution to this problem is to impose conditions on ρ that will ensure that the two spaces coincide. A necessary and sufficient condition for this is the so called Δ_2-condition on the function ρ (see, e.g., Chapter II §10 in Krasnosel'skii and Rutickii, 1961):

$$\lim_{u\to\infty}\frac{u\rho'(u)}{\rho(u)}<+\infty. \tag{8}$$

Condition (8) effectively restricts the growth of ρ to be slower than that of an exponential. Under this condition, $L^\mu_\rho(\mathcal{G}\otimes\mathcal{E})=E^\mu_\rho(\mathcal{G}\otimes\mathcal{E})$. In terms of the function ϕ entering the divergence, the Δ_2-condition in (8) is equivalent to the following.

Assumption A4. $\lim_{u\to\infty}\frac{u\phi'(u)}{\phi(u)}<+\infty$.

As already pointed out, Assumption A4 restricts the growth of ϕ to be slower than that of an exponential. This restriction is easily satisfied by all the members of the Cressie-Read family, including the I-divergence. Under this restriction, the notion of E-weak convergence is equivalent to the usual weak convergence, and we can consider the usual weak topology on $L_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$. Since we are interested in linear functionals $\langle u,\cdot\rangle$ on $E^\mu_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$ which is a subset of $L_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$, it suffices to consider the induced topology on $E^\mu_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$.

Now that we have equipped the spaces $L^\mu_\rho(\mathcal{G}\otimes\mathcal{E})$ and $E^\mu_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$ with compatible topologies and have defined the notions of convergence and continuity, we can proceed with establishing sufficient conditions for the requirements (e1)-(e3) in the general proof of existence.

3.3.2 Existence using strong Cramér condition

We now go back to our general proof strategy and establish existence of the solution π^0 to the problem (5) in the Orlicz space $L^\mu_\rho(\mathcal{G}\otimes\mathcal{E})$ with $\rho(u)=\phi(1+u)$. For this, recall that we first need to: (e1) extract a subsequence $\{\pi_{n_i}\}$ from a minimizing sequence $\{\pi_n\}$ that converges in some sense to a limit π^0. Here, the notion of convergence is that of E-weak convergence, which is particularly useful because every Orlicz space $L^\mu_\rho(\mathcal{G}\otimes\mathcal{E})$ is E-weakly compact (see, e.g., Theorem 14.4 in Krasnosel'skii and Rutickii, 1961), i.e. every bounded sequence contains an E-weakly converging subsequence. This property will allow us to establish (e1) provided we can show that the minimizing sequence $\{\pi_n\}$ is bounded. This in turn will follow from the assumed feasibility of the problem and Lemma 3.

Next, we need to ensure that (e2) the E-weak limit of the subsequence obeys the moment condition. Recall that the subsequence is E-weakly converging to $\pi^0 \in L^\mu_\rho(\mathcal{G}\otimes\mathcal{E})$ means that $\lim_{i\to\infty}\int T\pi_{n_i}d(P\times\mu) = \int T\pi^0 d(P\times\mu)$ for every $T \in E^\mu_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$. In particular, take $T = 1$. Then, $\int \rho^*(\alpha|T|)d(P\times\mu) = \rho^*(\alpha) < +\infty$ for every $\alpha > 0$, where the last inequality follows because ρ^* is finite. (This is a direct consequence of the fact that ρ is a real convex function satisfying $\lim_{u\to\infty}\rho(u)/u = +\infty$, and Lemma 4.2 in Borwein and Lewis (1991a) stated in Appendix A.) Thus, $T = 1$ is an element of $E^\mu_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$, and the limit π^0 satisfies: $\int \pi^0 d(P\times\mu) = \lim_{i\to\infty}\int \pi_{n_i}d(P\times\mu) = 1$. Now, recall that in order for π^0 to satisfy the constraints that define the projection set $\mathcal{C}$ in (4), we need to have $\int_{\mathbb{E}}\pi^0(\omega,x)d\mu(x) = 1$ for a.e. ω. In other words, it is not sufficient that only the unconditional restriction $\int_\Omega\int_{\mathbb{E}}\pi^0(\omega,x)d\mu(x)dP(\omega) = 1$ hold; rather, for a.e. ω the conditional restriction $\int_{\mathbb{E}}\pi^0(\omega,x)d\mu(x) = 1$ needs to be satisfied. Of course, the conditional restriction being stronger than the unconditional one, we shall need to work a set of unconditional restrictions which when taken together are equivalent to the conditional one.

This problem is akin to the problem of transforming a conditional moment restriction to an equivalent set of unconditional moment restrictions. Elegant solutions to this problem have been proposed in the literature on specification testing, and we shall in particular follow here the approach of Stinchcombe and White (1998). The key idea is simple: an element $m \in L_1(\mathcal{G})$ is equal to zero, $m = 0$ a.s., if and only if for every $l \in L_\infty(\mathcal{G})$, $\int_\Omega m(\omega)l(\omega)dP(\omega) = 0$. Here, $m(\omega) \equiv \int_{\mathbb{E}}[\pi^0(\omega,x) - 1]d\mu(x)$ which is in $L_1(\mathcal{G})$ in view of the Hölder inequality (7). So if every $l \in L_\infty(\mathcal{G})$ is in $E^\mu_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$ then $m = 0$ a.s.[11] Now recall that ρ^* is finite, so $|\int \rho^*(|l|)\,d(P\times\mu)| \leqslant \max_{v\in[0,L]}|\rho^*(v)| < +\infty$ where $L \equiv \|l\|_\infty$. So any $l \in L_\infty(\mathcal{G})$ is also in $E^\mu_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$. This establishes that the limit π^0 satisfies $\int_{\mathbb{E}}\pi^0(\omega,x)d\mu(x) = 1$ for a.e. ω.

In order to ensure that π^0 satisfies the conditional moment constraints as defined by the moment function a in (4), similar reasoning applies. It is sufficient to require that $a_i \in E^\mu_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$ for every component i of the moment function a. This leads to the following assumption.

[11] The reasoning is as follows. Take any $l \in L_\infty(\mathcal{G})$. Then,

$$\lim_{i\to\infty}\int_\Omega\int_{\mathbb{E}} l(\omega)\pi_{n_i}(\omega,x)d\mu(x)dP(\omega) = \lim_{i\to\infty}\int_\Omega l(\omega)\left[\int_{\mathbb{E}}\pi_{n_i}(\omega,x)d\mu(x)\right]dP(\omega) = \int_\Omega l(\omega)dP(\omega)$$
$$= \int_\Omega\int_{\mathbb{E}} l(\omega)\pi^0(\omega,x)d\mu(x)dP(\omega),$$

where the last equality follows if $l \in E^\mu_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$. So letting $m(\omega) = \int_{\mathbb{E}}[\pi^0(\omega,x) - 1]d\mu(x)$, we obtain $\int_\Omega m(\omega)l(\omega) = 0$.

Assumption A5. *Each coordinate a_i $(1 \leqslant i \leqslant m)$ of the moment function a satisfies:*

$$\int \phi^* \left(\tau |a_i|\right) d(P \times \mu) < +\infty \quad \textit{for every } \tau > 0.$$

Assumption A5 is best interpreted in the special case of I-divergence, for which $\phi^*(v) = \exp(v) - 1$. In this case, the above requirement becomes $E[\exp(\tau|a_i(X)|)] < +\infty$ for all $\tau > 0$, where the expectation is being taken under $P \times \mu$. Thus, Assumption A5 says that the moment generating function of the moment function a_i has to be finite for all $\tau > 0$ and is akin to the Cramér condition. Although strong, this requirement is substantially weaker than requiring a to be bounded as was done in Assumption A3. Indeed, in most statistical and econometric applications, Assumption A3 is too strong and it is often ruled out by the nature of the model itself. For instance, consider a model with a conditional mean restriction in which $a_i(\omega, x) = x$ for a.e. ω, and let X be conditionally normally distributed. This simple setup violates Assumption A3; Assumption A5 on the other hand remains satisfied for all divergences of the Cressie-Read family with $a \geq 0$. In particular, this includes the I-divergence, obtained when $a = 0$. In the context of the unconditional projection problem, similar condition to Assumption A5 can be found Csiszár (1995).

Finally, we need to ensure (e3) that I_ϕ remains lower semi-continuous on $\mathcal{C}$ in the E-weak topology. Here, the importance of working with the Orlicz space $L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$ in which the Δ_2-condition (8) (or equivalently Assumption A4) holds comes to full light: in those spaces the notions of weak convergence and E-weak convergence coincide. So it will be sufficient to ensure that I_ϕ is lower semi-continuous on $\mathcal{C}$ in the weak topology of $L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$. As already pointed out, a direct implication of Lemma 1 is that I_ϕ is lower semi-continuous in the weak topology of $L^\mu_1(\mathcal{G} \otimes \mathcal{E})$. Since I_ϕ is convex, this is equivalent to lower semi-continuity in the norm topology of $L^\mu_1(\mathcal{G} \otimes \mathcal{E})$ (see, Corollary I.2.2 in Ekeland and Témam, 1987). In order to show that lower semi-continuity remains when we change the space to $L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$, the following result will be crucial.

Lemma 4. *Let Assumptions A1 and A2 hold and take $\rho(u) = \phi(1+u)$. Then, $L^\mu_\rho(\mathcal{G} \otimes \mathcal{E}) \subseteq L^\mu_1(\mathcal{G} \otimes \mathcal{E})$, and there exists $q > 0$ such that $q\|\cdot\|^\mu_1 \leqslant \|\cdot\|^\mu_\rho$.*

Put in words, Lemma 4 says that for any sequence in the Orlicz space $L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$, convergence in Luxemburg norm $\|\cdot\|^\mu_\rho$ implies convergence in $\|\cdot\|^\mu_1$. Now, for any minimizing sequence $\{\pi_n\}$, letting $\rho(u) = \phi(1+u)$ we know that $\{\pi_n\} \in L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$. Thus, if $\|\pi_n - \pi^0\|^\mu_\rho \to 0$ then by the above lemma $\|\pi_n - \pi^0\|^\mu_1 \to 0$; by the lower semi-continuity of I_ϕ in $L^\mu_1(\mathcal{G} \otimes \mathcal{E})$, we then have $\liminf_n I_\phi(\pi_n) \geqslant I_\phi(\pi^0)$. That is, I_ϕ is lower semi-continuous in the norm topology of $L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$,

and by convexity the result remains true if we change topology to the weak topology (see again, Corollary I.2.2 in Ekeland and Témam, 1987). We then have the following result.

Theorem 3. *Let Assumptions A1, A2, A4 and A5 hold. Assume in addition that the problem is feasible, i.e. there exists at least one $g_0 \in \mathcal{Q}$ such that $\mathcal{D}_\phi(g_0, f) < +\infty$. Then a $\mathcal{D}_\phi$-projection of f onto $\mathcal{Q}$ exists and is unique.*

As already discussed, the key feature of the existence approach based on Orlicz spaces is that it allows us to relax the boundedness assumption on the moment condition. This however comes at a cost: Theorem 3 imposes strong moment conditions on the function a. These conditions can be relaxed, albeit under stronger conditions on the function ϕ that defines the divergence.

3.4 Existence using weak Cramér condition

Recall that the role of Assumption A5 was to ensure that all the components a_i of the moment function a that defines the projection set $\mathcal{C}$ belong to the space $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$. In general, the space $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$ is a strict subspace of the Orlicz space $L^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$. One immediate way to weaken the restrictions imposed on a is to ensure that the two spaces coincide. This is accomplished by imposing the following.

Assumption A6. $\lim_{u\to\infty} \frac{u\phi'(u)}{\phi(u)} > 1$.

Recall that under Assumption A2, the ratio $\phi(u)/u$ increases to infinity when u gets large. This implies in particular that $\lim_{u\to\infty} \phi'(u) = +\infty$. So the statement in Assumption A6 can be interpreted as saying that $\phi'(u)$ increases faster than $\phi(u)/u$. It is straightforward to check that Assumption A6 holds for all the members of the Cressie-Reed family with $a > 0$. It does not hold, however, for the I-divergence. Indeed, the limit condition in Assumption A6 can be interpreted as saying that ϕ increases at infinity strictly faster than $u \ln u$.

When the function ϕ satisfies the growth requirements in Assumption A6, the subspace $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$ is actually equal to the entire space $L^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$, i.e. $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E}) = L^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$ (see, e.g., Krasnosel'skii and Rutickii, 1961). When on the other hand, Assumption A6 fails, $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$ is a strict subset of $L^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$. Since under Assumption A6, $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E}) = L^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$, we can now work with the following condition instead of the moment requirement in Assumption A5.

Assumption A7. *Each coordinate a_i $(1 \leqslant i \leqslant m)$ of the moment function a satisfies:*

$$\int \phi^* \left(\tau_i |a_i|\right) d(P \times \mu) < +\infty \quad \textit{for some } \tau_i > 0.$$

The key difference between Assumptions A7 and A5 is that the latter imposes integrability for all values of $\tau > 0$, while the first only requires the result to hold for some $\tau_i > 0$. This difference is significant for functions ϕ which fail to satisfy the requirements in Assumption A6. Indeed, if $\phi(u)$ fails to grow at infinity strictly faster than $u \ln u$, then its conjugate $\phi^*(v)$ grows as an exponential (or faster). In those cases, a moment function a may well pass the integrability requirement in Assumption A7 yet fail the one in Assumption A5. We then obtain the following result.

Theorem 4. *Let Assumptions A1, A2, A4, A6, and A7 hold. Assume in addition that the problem is feasible, i.e. there exists at least one $g_0 \in \mathcal{Q}$ such that $\mathcal{D}_\phi(g_0, f) < +\infty$. Then a $\mathcal{D}_\phi$-projection of f onto $\mathcal{Q}$ exists and is unique.*

To summarize, there are several sets of assumptions that can be used to establish existence (and uniqueness) of the $\mathcal{D}_\phi$-projection of f onto a set of conditional moment restrictions. They all involve a trade-off between restrictions put on the function ϕ used to define the divergence, and the moment function a which defines the projection set. The table below summarizes the key conditions; these are in addition to the feasibility assumption which needs to hold in all cases.

Existence and Uniqueness Results		
Theorem 2	**Theorem 3**	**Theorem 4**
a_i bounded	$\forall \tau > 0,\ E\left[\phi^*(\tau \lvert a_i \rvert)\right] < +\infty$	$\exists \tau_i > 0,\ E\left[\phi^*(\tau_i \lvert a_i \rvert)\right] < +\infty$
$\lim_{u\to\infty} \frac{\phi(u)}{u} = +\infty$	$\lim_{u\to\infty} \frac{\phi(u)}{u} = +\infty$	$\lim_{u\to\infty} \frac{\phi(u)}{u} = +\infty$
	$\lim_{u\to\infty} \frac{u\phi'(u)}{\phi(u)} < +\infty$	$\lim_{u\to\infty} \frac{u\phi'(u)}{\phi(u)} < +\infty$
		$\lim_{u\to\infty} \frac{u\phi'(u)}{\phi(u)} > 1$

In the table above, a_i refers to the component i ($1 \leqslant i \leqslant m$) of the moment function a, the expectation is taken with respect to the product $P \times \mu$, i.e. under the conditional density f.

4 Characterization

Now that we have provided several sets of conditions that guarantee existence and uniqueness of the $\mathcal{D}_\phi$-projection of f onto a set of conditional moment restrictions, we turn our attention to the characterization of the said projection.

4.1 General Problem

To be specific, we first need to set our optimization problem in a particular space. As in the previous section, where we have discussed existence, there are a couple of possible choices for the space in which to work: $L_1^\mu(\mathcal{G} \otimes \mathcal{E})$ is a natural candidate because we are dealing with conditional densities. This space, however, does not take into account the shape of the $\mathcal{D}_\phi$-divergence; as a result, the conditions imposed on the moment function a (that of being bounded) are unnecessarily strong (c.f. Theorem 2). An alternative is to work in the Orlicz space $L_\rho^\mu(\mathcal{G} \otimes \mathcal{E})$ whose geometry is intimately linked to the shape of the $\mathcal{D}_\phi$-divergence, obtained by setting $\rho(u) = \phi(1+u)$. Working in Orlicz spaces allows for weaker assumptions on the moment function a such as those given in Assumptions A5 or A7. This is the approach we shall follow in order to characterize the projection. More specifically, in order to obtain results that apply to a wide range of divergences, including the I-divergence, we shall hereafter work under the assumptions of Theorem 3. Of course, the characterization remains valid if additional divergence restrictions are imposed, as was done in Theorem 4.

As before, we are interested in solving the constrained optimization problem in Equation (5), $\inf_{\pi \in \mathcal{C}} I_\phi(\pi)$, when π is an element of the Orlicz space $L_\rho^\mu(\mathcal{G} \otimes \mathcal{E})$, and the projection set $\mathcal{C}$ is a convex subset of $L_\rho^\mu(\mathcal{G} \otimes \mathcal{E})$ defined by the moment function a, i.e.

$$\mathcal{C} \equiv \left\{ \pi \in L_\rho^\mu(\mathcal{G} \otimes \mathcal{E}) f : \int_{\mathbb{E}} a(\omega, x)\pi(\omega, x)d\mu(x) = 0, \int_{\mathbb{E}} \pi(\omega, x)d\mu(x) = 1, \text{ for a.e. } \omega \right\}.$$

The projection set $\mathcal{C}$ is defined by $m+1$ linear equality constraints with m being the dimension of the moment function codomain, and we can write that $\pi \in \mathcal{C}$ if and only if $(T\pi)(\omega) = c$ for a.e. ω, where $T : L_\rho^\mu(\mathcal{G} \otimes \mathcal{E}) \to L_1^{m+1}(\mathcal{G}) = L_1(\mathcal{G}) \times \cdots \times L_1(\mathcal{G})$ ($m+1$ times) is a linear operator defined by

$$(T\pi)(\omega) \equiv \left(\int_{\mathbb{E}} a(\omega, x)'\pi(\omega, x)d\mu(x), \int_{\mathbb{E}} \pi(\omega, x)d\mu(x) \right)', \tag{9}$$

and

$$c \equiv (0', 1)' \in \mathbb{R}^{m+1}.$$

Our optimization problem can then be written as:

$$\text{minimize } I_\phi(\pi) \text{ subject to } T\pi = c \text{ a.e.}, \tag{10}$$

where $\pi \in L_\rho^\mu(\mathcal{G} \otimes \mathcal{E})$. Note that the problem (10) is a convex optimization problem under linear

equality constraints. There are however several important points to be made:

(i) There are numerous existence and characterization results derived for moment equality constraints (see, e.g., Borwein and Lewis, 1991a,b, 1993). The key difference between those works and our problem (10), is that we work with constraints that are stochastic. Hence, there is an infinite number of linear constraints in our convex optimization problem. This feature makes the conditional problem very different from the unconditional problem in which there is only one (or a finite number) of constraints.[12]

(ii) In principle, the problem (10) can be analyzed using the Lagrange multiplier theorem. However, the application of this result is typically carried out under some differentiability assumption for the objective function $\mathcal{D}_\phi(\cdot, f)$. The problem here is that $\mathcal{D}_\phi(\cdot, f)$ is only finite on the positive cone $\mathcal{P}$. If the true density f is such that it reaches the boundary of the cone $\mathcal{P}$, i.e. if the density f can be arbitrarily close to 0, then establishing differentiability of $\mathcal{D}_\phi(\cdot, f)$ can be a problem. Indeed, if for example the support of f is the entire real line, then we would need to establish the differentiability of $\mathcal{D}_\phi(g, f)$ with respect to g, where g has the same support as f and is thus arbitrarily close to 0. This is a problem since $\mathcal{D}_\phi(g, f)$ becomes infinite as one moves away from g in directions that would lead to a negative density.[13] More formally, say that we work in $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$; then, the gradients of lower semi-continuous convex functions such as $\mathcal{D}_\phi(\cdot, f)$ can only be computed on the interior of their effective domain (see, e.g., p.33 in Rockafellar, 1974). Since the effective domain of $\mathcal{D}_\phi(\cdot, f)$ is a subset of the positive cone $\mathcal{P}$ in $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$, and since the latter has empty interior, it follows that the interior of the effective domain of $\mathcal{D}_\phi(\cdot, f)$ is also empty.

(iii) An elegant solution to the differentiability problem is to work with the dual of the problem in (10). As we shall proceed to show below, working with the dual does not require differentiability assumptions on the primal. Moreover, the dual will be stated in terms of Lagrange multipliers $\eta(\omega)$ and $\lambda(\omega)$ that are functions of the conditioning variable ω alone, unlike the primal which involves optimizing over functions $g(\omega, x)$ of both ω and x.

(iv) Due to the way we extended ϕ on $\mathbb{R}$, a solution g^0 to the problem (10) (when it exists) is automatically non-negative valued, provided there exists at least one $g \in \mathcal{Q}$ such that

[12] The other difference between our approach and that of Borwein and Lewis is that they work in L_1. Thus, as discussed before, they rely on the boundedness of the moment function to establish existence and characterize the solution.

[13] Page 330 in Borwein and Lewis (1991a) contains a simple example illustrating this point.

$\mathcal{D}_\phi(g, f) < +\infty$.

4.2 Dual Problem

Given the paired spaces $L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$ and $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$, and the pairing $\langle \cdot, \cdot \rangle$ in (6), we can proceed with the discussion of the dual to the optimization problem (10). For this, we first transform the latter into an unconstrained optimization problem by modifying the objective function to be minimized. Let $\delta(\cdot|E)$ denote the indicator function of a given set E, i.e.

$$\delta(x|E) \equiv \begin{cases} 0, & x \in E \\ +\infty, & \text{otherwise.} \end{cases}$$

The indicator function δ is convex if and only if the set E is a convex set; this will be the case in our setup. The constrained optimization problem in (10) is then equivalent to:

$$\text{minimize } \big[I_\phi(\pi) + \delta(T\pi|\{c\}) \big], \tag{11}$$

where $\pi \in L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$, the linear operator $T : L^\mu_\rho(\mathcal{G} \otimes \mathcal{E}) \to L^{m+1}_1(\mathcal{G})$ is as previously defined in (9), and $c = (0', 1)' \in \mathbb{R}^{m+1}$ as before.

A careful discussion of the operator T is needed at this point. As discussed in the previous section, for the integrals in (9) to be well defined, we need to put some restrictions on the moment function a. Specifically, if we let Assumption A5 hold, i.e. if we assume every component a_i is in $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$, then

$$\begin{aligned} \int_\Omega |(T\pi)(\omega)| dP(\omega) &= \int_\Omega \left(\sum_{i=1}^m \left| \int_{\mathbb{E}} a_i(\omega, x)\pi(\omega, x) d\mu(x) \right| + \left| \int_{\mathbb{E}} \pi(\omega, x) d\mu(x) \right| \right) dP(\omega) \\ &\leqslant \sum_{i=1}^m \int_\Omega \int_{\mathbb{E}} |a_i(\omega, x)\pi(\omega, x)| d\mu(x) dP(\omega) + \int_\Omega \int_{\mathbb{E}} |\pi(\omega, x)| d\mu(x) dP(\omega) \\ &= \sum_{i=1}^m \int |a_i||\pi| d(P \times \mu) + \int |\pi| d(P \times \mu) < +\infty, \end{aligned}$$

where the last inequality follows by the Hölder inequality in Orlicz spaces (see Equation (7)). Thus, T is well defined. Now for every $\tau \in L^{m+1}_\infty(\mathcal{G})$ consider the linear functional $l : L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$,

$u \mapsto l(u) \equiv \int (Tu)'\tau dP$. More specifically,

$$\begin{aligned} l(u) &= \int_\Omega (Tu)(\omega)'\tau(\omega) dP(\omega) \\ &= \int_\Omega \left(\sum_{i=1}^m \tau_i(\omega) \int_{\mathbb{E}} a_i(\omega, x) u(\omega, x) d\mu(x) + \tau_{m+1}(\omega) \int_{\mathbb{E}} u(\omega, x) d\mu(x) \right) dP(\omega) \\ &= \int_\Omega \int_{\mathbb{E}} \left(\sum_{i=1}^m \tau_i(\omega) a_i(\omega, x) + \tau_{m+1}(\omega) \right) u(\omega, x) d\mu(x) dP(\omega) \\ &= \int_\Omega \int_{\mathbb{E}} v(\omega, x) u(\omega, x) d\mu(x) dP(\omega), \end{aligned}$$

where we have let

$$v(\omega, x) \equiv \sum_{i=1}^m \tau_i(\omega) a_i(\omega, x) + \tau_{m+1}(\omega). \tag{12}$$

Recall that under Assumption A5, each term in the above sum is in $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$ (see the proof of Theorem 3 for details). Since $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$ is a subspace, it then follows that the sum itself is in $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$, i.e. $v \in E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$. Then, using the pairing in (6) we can write that $l(u) = \langle u, v \rangle$ is a continuous linear functional.

The expression for v in Equation (12) allows us to define the conjugate functional $T^* : L^{m+1}_\infty(\mathcal{G}) \to E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$ as $T^*\tau = v$. We are now ready to define the dual of the primal problem in (11):

$$\text{maximize } \left[\langle \tau, c \rangle - I_{\phi^*}\left(T^*\tau\right) \right], \tag{13}$$

with $\tau \in L^{m+1}_\infty(\mathcal{G})$. Recall that $c = (0', 1)' \in \mathbb{R}^{m+1}$. Letting λ denote the first m components of τ, and η denote its last component, i.e. $\lambda \equiv (\tau_1, \ldots, \tau_m)$ and $\eta \equiv \tau_{m+1}$, an explicit expression for the dual is given in the lemma below.

Lemma 5. *Let Assumptions A1, A2, A4 and A5 hold. Then the projection problem:*

$$\min_{g \in \mathcal{Q}} \ \mathcal{D}_\phi(g, f) \tag{P}$$

has a dual:

$$\max_{(\eta, \lambda) \in (L_\infty(\mathcal{G}), L^m_\infty(\mathcal{G}))} \ \left[\int_\Omega \eta(\omega) dP(\omega) - \int_\Omega \int_{\mathbb{E}} \phi^* \left(\eta(\omega) + \lambda(\omega)' a(\omega, x) \right) d\mu(x) dP(\omega) \right]. \tag{D}$$

The key feature of the duality approach is that it transforms an optimization problem in π (or g) which is a function of two variables ω and x, into an optimization problem in η and λ which

are only functions of ω. In particular, in the unconditional version of the problem considered by Borwein and Lewis (1991a, 1993), η and λ are simply constants in $\mathbb{R}$ and $\mathbb{R}^m$, respectively. Though the above result gives the dual of our projection problem, nothing is said about whether the dual is attained. This is the goal to which we turn next.

4.3 Strong Duality

For the dual formulation in Lemma 5 to be useful, one needs to ensure that the strong duality relation between the primal problem (11) and the dual problem (13) holds, i.e. that

$$\min(\mathrm{P}) = \max(\mathrm{D}).$$

The above equality requires a so-called constraint qualification condition. There is a variety of constraint qualification conditions that have been proposed in the literature (see, e.g., Gowda and Teboulle, 1990; Zălinescu, 1999, for reviews). In this paper, we shall work with the following constraint qualification proposed by Zălinescu (1999) (Theorem 8(v)):

$$c \in \mathrm{icr}\left(T\mathrm{dom}I_\phi\right), \tag{14}$$

where $T\mathrm{dom}I_\phi$ is the image by T of the effective domain of I_ϕ, i.e. $T\mathrm{dom}I_\phi \equiv \{u \in L_1(\mathcal{G}) : u = T\pi, I_\phi(\pi) < +\infty\}$, and for any set A, $\mathrm{icr}(A)$ denotes the intrinsic core of A, i.e. $\mathrm{icr}(A) \equiv \{a \in A : \forall b \in \mathrm{aff}\,(A) \setminus \{a\}, \exists x \in (a,b), [a,x] \subset A\}$, $\mathrm{aff}\,(A)$ is the affine hull generated by A, i.e. $\mathrm{aff}\,(A) = \{\sum \alpha_i x_i : \sum \alpha_1 = 1, x_i \in A\}$, and $[x,y]$ (resp. (x,y)) denotes the line segment between x and $y \neq x$, i.e. $[x,y] = \{\alpha x + (1-\alpha)y : 0 \leqslant \alpha \leqslant 1\}$ (resp. $(x,y) = [x,y] \setminus \{x,y\}$) (see, e.g., p.7-8 in Holmes, 1974). The condition (14) is an interiority condition. A simple sufficient condition is given by the following assumption.

Assumption A8. *For every $g_1 \in \mathrm{dom}\mathcal{D}_\phi(\cdot, f) \setminus \mathcal{Q}$ (i.e. such that $\mathcal{D}_\phi(g_1, f) < +\infty$ and $g_1 \notin \mathcal{Q}$) there exist $0 < \alpha < 1$ and $g_2 \in \mathrm{dom}\mathcal{D}_\phi(\cdot, f) \setminus \mathcal{Q}$ satisfying $\alpha g_1 + (1-\alpha)g_2 \in \mathcal{Q}$, i.e. such that:*

$$\alpha \int_{\mathbb{E}} a(\omega, x) g_1(\omega, x) d\nu(x) + (1-\alpha) \int_{\mathbb{E}} a(\omega, x) g_2(\omega, x) d\nu(x) = 0$$
$$\alpha \int_{\mathbb{E}} g_1(\omega, x) d\nu(x) + (1-\alpha) \int_{\mathbb{E}} g_2(\omega, x) d\nu(x) = 1.$$

Put in words, Assumption A8 simply says that for any nonfeasible g_1 in the effective domain of the ϕ-divergence (i.e. such that g_1 does not satisfy the moment restrictions but the ϕ-divergence

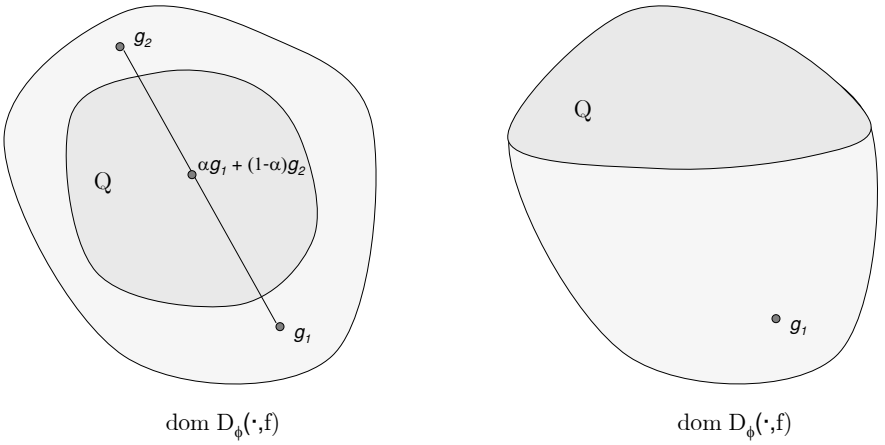


Figure 1: (left) Assumption A8 holds; (right) Assumption A8 is violated.

between g_1 and f is finite), it is possible to find some nonfeasible g_2 also in the effective domain, such that some convex combination of g_1 and g_2 satisfies the moment restrictions in $\mathcal{Q}$. Put differently, for any g_1 not in the projection set $\mathcal{Q}$, it is possible to find some α, $0 < \alpha < 1$, and a g_2 not in $\mathcal{Q}$ such that $\alpha g_1 + (1-\alpha) g_2$ is in $\mathcal{Q}$. In this sense, Assumption A8 can be seen as an interiority condition: indeed if the set $\mathcal{Q}$ was at the boundary of the domain of $\mathcal{D}_\phi(\cdot, f)$, then it would be impossible to always find a line segment (g_1, g_2) that passes through $\mathcal{Q}$. Figure 1 illustrates the point.

Before proceeding, we comment on the constraint qualification condition (14), as compared to the ones previously used in the literature. In the case where the projection set is defined by unconditional moment restrictions, Borwein and Lewis (1993) propose to work with a constraint qualification condition of the form:

$$c \in \text{qri}\left(T \text{dom} I_\phi\right), \tag{15}$$

where $\text{qri}(A)$ denotes a quasi-relative interior of A, $\text{qri}(A) \equiv \left\{a \in A : \overline{\text{cone}(A - a)} \text{ is a subspace}\right\}$, $\text{cone}(B)$ denotes the cone generated by B, i.e. $\text{cone}(B) \equiv \{\lambda b : \lambda \geqslant 0, b \in B\}$, and $\overline{\text{cone}(B)}$ is its closure. This notion is introduced and studied extensively in Borwein and Lewis (1992a,b). What makes this notion particularly useful is the property that for linear mappings T with codomain $\mathbb{R}^{m+1}$, $T\text{qri}(\text{dom} I_\phi) \subset \text{qri}\left(T \text{dom} I_\phi\right)$ (see Proposition 2.7 in Borwein and Lewis, 1992a). Thus, a simple sufficient condition for $c \in \text{qri}\left(T \text{dom} I_\phi\right)$ is that there exists $\pi_0 \in \text{qri}(\text{dom} I_\phi)$ such that $T\pi_0 = c$ a.e. Translated in terms of densities, this gives the constraint qualification condition used

in Borwein and Lewis (1993):[14]

$$\text{there exists } g_0 \in \text{qri}(\text{dom}\mathcal{D}_\phi(\cdot, f)) \text{ such that } g_0 \in \mathcal{Q}. \tag{BL}$$

It is worth emphasizing that Borwein and Lewis's (1993) constraint qualification (15) and its sufficient condition (BL) are only valid for projection sets defined by unconditional moment restrictions. The reason again is that the notion of quasi-relative interior has useful properties when the linear mapping T under consideration has a codomain $\mathbb{R}^{m+1}$. In the case of conditional moment restrictions this is obviously not the case and we are dealing with a linear mapping T that maps to $L_1^{m+1}(\mathcal{G})$.

It would be useful however to be able to compare our constraint qualification condition (14) with that of Borwein and Lewis (1993) stated in (15) (or their sufficient conditions given in Assumption A8 and Equation (BL), respectively). For this, we need a generalized version of Borwein and Lewis's (1993) quasi-relative interior condition, which works for linear mappings that map into general Banach spaces. Gowda and Teboulle (1990) propose one such generalized condition, based on the notion of a strong quasi-relative interior, whereby $\text{sqri}(A) \equiv \{a \in A : \text{cone}(A - a) \text{ is a closed subspace}\}$. Their condition can be written as:

$$c \in \text{sqri}\left(T\text{dom}I_\phi\right) \tag{16}$$

As shown by Gowda and Teboulle (1990) and Zălinescu (1999), the strong quasi-relative interior condition (16) and our condition (14) based on the intrinsic core are related to each other by the following equivalence:

$$c \in \text{sqri}\left(T\text{dom}I_\phi\right) \Longleftrightarrow \begin{cases} c \in \text{icr}\left(T\text{dom}I_\phi\right) \\ \text{aff}\left(c - T\text{dom}I_\phi\right) \text{ is a closed subspace} \end{cases}$$

Thus, our constraint qualification condition (14) is strictly weaker than that of Gowda and Teboulle (1990) in (16) (which we recall again is the generalization of the constraint qualification condition by Borwein and Lewis (1993) that works for problems with conditional moment restrictions). Specifically, our condition (14) does not involve any closedness requirements. This we should point out is one important advantage of working with algebraic interior notions such as the intrinsic core, which unlike the (strong) quasi-relative interior do not depend on the topology. Moreover, without putting strong assumptions on the set $c - T\text{dom}I_\phi$ it is generally very difficult (if not impossible)

[14] See condition (PCQ_2) on p. 255 in Borwein and Lewis (1993)

to ensure that $\text{aff}\,(c - T\text{dom}I_\phi)$ is closed, thus preventing one from using the strong quasi-relative interior condition.

Under the constraint qualification condition (14) we obtain the following strong duality result.

Theorem 5. *Let Assumptions A1, A2, A4, A5, and A8 hold. Assume in addition that the problem is feasible, i.e. there exists* $g_0 \in \mathcal{Q}$ *such that* $\mathcal{D}_\phi(g, f) < +\infty$. *Then* $\min(\text{P}) = \max(\text{D})$, *and there is a unique solution* g^0 *to* P, *and a unique solution* (η^0, λ^0) *to* D.

4.4 Optimality Criterion

At last, we are able to characterize the projection by using the strong duality result of Theorem 5. For this, we first need to introduce the notion of a subgradient. Recall again that we are working in the space $L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$ paired with $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$ with the inner product defined in (6), i.e. for every $(u, v) \in L^\mu_\rho(\mathcal{G} \otimes \mathcal{E}) \times E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$, $\langle u, v\rangle = \int uvd(P \times \mu)$. For the function I_ϕ defined in (3), we say that $v \in E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$ is a subgradient of I_ϕ at $u \in L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$ if

$$I_\phi(u') \geqslant I_\phi(u) + \langle u' - u, v\rangle \quad \text{for all } u' \in L^\mu_\rho(\mathcal{G} \otimes \mathcal{E}), \tag{17}$$

which due to the property of the conjugate I_{ϕ^*} of I_ϕ, $I_{\phi^*}(v) = \sup_u [\langle u, v\rangle - I_\phi(u)]$, is then equivalent to

$$I_{\phi^*}(v) = \langle u, v\rangle - I_\phi(u).$$

The set of all subgradients of I_ϕ at u is denoted by $\partial I_\phi(u)$. The subgradient set $\partial I_\phi(u)$ may be empty; when nonempty, it is always closed and convex in $E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$.

Identical equations to those above will hold for other paired spaces, e.g. $L_1^{m+1}(\mathcal{G})$ and $L_\infty^{m+1}(\mathcal{G})$. In particular, we can use them to characterize the subgradient of $\tau \mapsto \langle \tau, c\rangle$, $\tau \in L_\infty^{m+1}(\mathcal{G}), c = (0', 1)' \in \mathbb{R}^{m+1}$, which is no other than the support function of the set $\{c\}$ (the latter being the conjugate of the indicator function $\delta(\cdot|\{c\})$ introduced before). We simply have that for all $\tau \in L_\infty^{m+1}(\mathcal{G})$, $\partial\langle \tau, c\rangle = c$.

We can now use the above notion of subgradients to derive the optimality condition for the ϕ-projection g^0 of f onto $\mathcal{Q}$, or equivalently for π^0 that solves (11). From Theorem 15 in Rockafellar (1974) (see also his Example 11' on p.50), we know that π^0 solves (P), τ^0 solves (D) and $\min(P) = \max(D)$, if the pair (π^0, τ^0) solves the Kuhn-Tucker condition

$$T\pi^0 = \partial\langle \tau^0, c\rangle = c \quad \text{and} \quad T^*\tau^0 \in \partial I_\phi(\pi^0).$$

As shown in Rockafellar (1971), the second property is equivalent to $T^*\tau^0 \in \partial\phi(\pi^0) = \{\phi'(\pi^0)\}$ since ϕ is differentiable by Assumption A1(i). Now, recall that for any $v \in \mathbb{R}$ we have $(\phi')^{-1}(v) = (\phi^*)'(v)$ (Lemma 6(v) and Assumption A2), so the last equality can be written as $\pi^0 = (\phi^*)'\left(T^*\tau^0\right)$. Translating this result in terms of the projection $g^0 = \pi^0 f$, we have thus shown the following.

Corollary 6. *Under the conditions of Theorem 5, the solution* g^0 *to* P *is given by*

$$g^0(\omega, x) = (\phi^*)'\left(\eta^0(\omega) + \lambda^0(\omega)' a(\omega, x)\right) f(\omega, x),$$

where (η^0, λ^0) *is the unique solution to* D.

5 Discussion and Conclusion

5.1 Modified divergences

In proving the existence of the projection g^0 one faces the trade-off between the restrictions on the growth rate of ϕ and the boundedness of the moment function. Assumption A2 could be relaxed only if one is willing to consider a bounded moment function (or a bounded space Ω). Even in this case, however, Borwein and Lewis (1993) have shown that Assumption A2 is necessary in order to guarantee that the solution does not posses singular components. For general case of a possibly unbounded moment function, the proof of existence of the $\mathcal{D}_\phi$-projection based on weak compactness in Orlicz spaces relies on Assumption A6 which imposes an additional constraint on the the growth rate of ϕ.

These constraints on the rate of growth of ϕ are satisfied only by a small subset of the functions ϕ introduced in Section 2.3. Assumption A2 holds for the Kullback-Leibler distance and for members of the Cressie-Read class with $a > 0$; on the other hand, Assumption A6 holds only for members of the Cressie-Read with $a > 0$. It is, however, possible to modify a function ϕ in order to make it compatible with the rate of growth prescribed by the aforementioned assumptions. The idea is to modify the behavior of the divergence only in the "tail" while leaving it unchanged otherwise.

Consider a divergence function ϕ which satisfies Assumption A1, but not Assumption A2, that is, $d \equiv \lim_{u\to+\infty} \phi'(u) < +\infty$. Note that $d < +\infty$ implies that ϕ does not satisfy Assumption A6.

For some $\vartheta > 0$, let $u_\vartheta \equiv 1 + \vartheta$. The modified divergence ϕ_ϑ is defined as

$$\phi_\vartheta(u) \equiv \begin{cases} \phi(u_\vartheta) + \phi'(u_\vartheta)(u - u_\vartheta) + \frac{1}{2}\phi''(u_\vartheta)(u - u_\vartheta)^2, & u \geqslant u_\vartheta \\ \phi(u), & u \in (0, u_\vartheta) \\ \lim_{u \to 0^+} \phi(u), & u = 0 \\ +\infty, & u < 0 \end{cases}.$$

Put in words, the modified divergence ϕ_ϑ mimics the behavior of the original divergence ϕ up to a cut-off u_ϑ that is strictly greater than one; beyond that cut-off, the original divergence is replaced by a quadratic that is a "smooth" continuation of ϕ, i.e. whose level and slope match that of ϕ at u_ϑ.

It is immediate to verify that the modified divergence satisfies all the requirements of Assumption A1. Furthermore, it holds that

$$\lim_{u \to \infty} \frac{\phi_\vartheta(u)}{u} = +\infty, \qquad \text{and} \qquad \lim_{u \to \infty} \frac{u\phi'_\vartheta(u)}{\phi_\vartheta(u)} = 2,$$

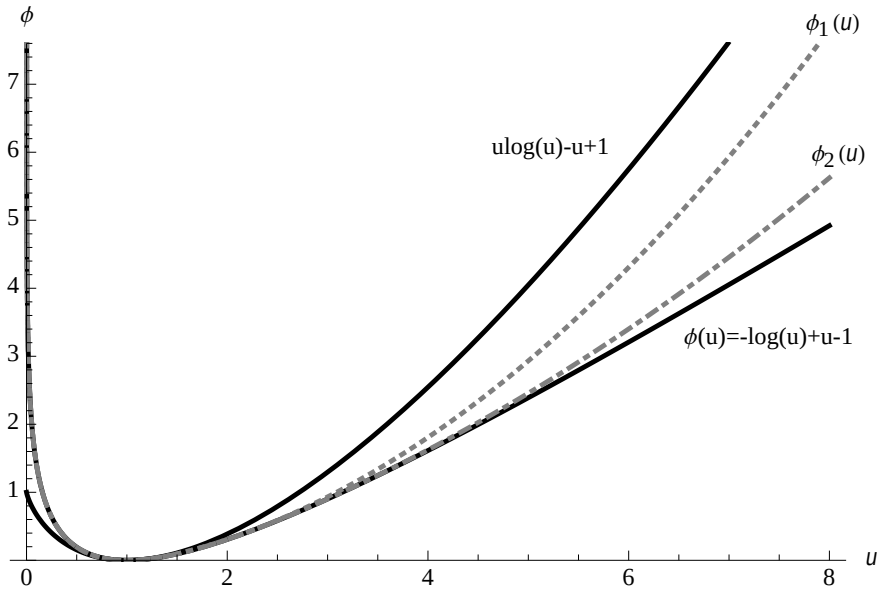
which implies that: (i) the rate of growth of ϕ_ϑ is consistent both with Assumption A2 and Assumption A6, and (ii) the image of ϕ'_ϑ is the real line and thus $\overline{\operatorname{dom} \phi^*_\vartheta} = (-\infty, +\infty)$. The expression for the conjugate is obtained by applying the Legendre-Fenchel transform to obtain

$$\phi^*_\vartheta(\upsilon) = \begin{cases} a_\vartheta \upsilon^2 + b_\vartheta \upsilon + c_\vartheta, & \upsilon > \phi'(u_\vartheta), \\ \phi^*(\upsilon), & \upsilon \leqslant \phi'(u_\vartheta) \end{cases},$$

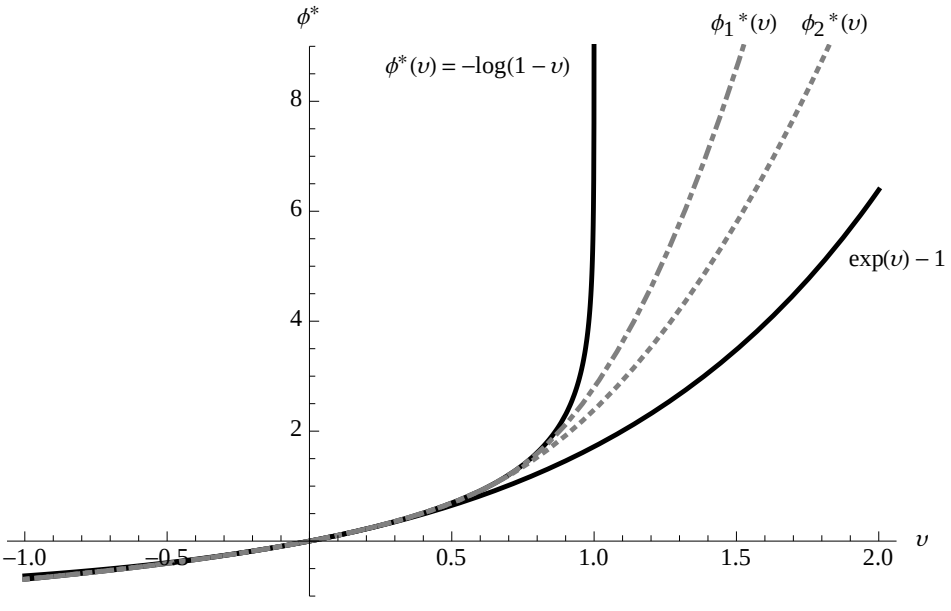
where $a_\vartheta = 1/(2\phi''(u_\vartheta))$, $b_\vartheta = u_\vartheta - 2a_\vartheta\phi'(u_\vartheta)$, and $c_\vartheta = -\phi(u_\vartheta) + a_\vartheta\phi'(u_\vartheta) - u_\vartheta^2/a_\vartheta$. Importantly, the conjugate $\phi^*_\vartheta(u)$ has a closed form expression whenever the original divergence function ϕ does so. The example below illustrates the computation and the properties of the modified divergence.

Example. *(Reverse I-divergence)* The reverse I-divergence, $\phi(u) = -\ln u + u - 1$, does not satisfy either Assumption A2 or Assumption A6. The modified reverse I-divergence is given by

$$\phi_\vartheta(u) = \begin{cases} -\ln(u_\vartheta) + (1 - \frac{1}{u_\vartheta})u + \frac{1}{2u_\vartheta^2}(u - u_\vartheta)^2, & u > u_\vartheta \\ -\ln u + u - 1, & 0 < u \leqslant u_\vartheta \\ +\infty, & u \leqslant 0. \end{cases} \cdot \tag{18}$$



(a) Divergences



(b) Conjugates

Figure 2: The reverse I-divergence and its modifications.

The conjugate of ϕ_θ is given by

$$\phi^*_\vartheta(\upsilon) = \begin{cases} a_\vartheta \upsilon^2 + b_\vartheta \upsilon + c_\vartheta, & \upsilon > 1 - \frac{1}{u_\vartheta} \\ -\ln(1-\upsilon), & \upsilon \leqslant 1 - \frac{1}{u_\vartheta}, \end{cases} \tag{19}$$

where $a_\vartheta = u_\vartheta^2/2$, $b_\vartheta = u_\vartheta(2-u_\vartheta)$, and $c_\vartheta = \ln(u_\vartheta) - u_\vartheta - 1 + u_\vartheta(u_\vartheta - 1)/2$.

Figure 2 draws the I-divergence, the modified versions for $\vartheta = 1$ and $\vartheta = 2$, and the corresponding conjugate functions. The modified divergences' rate of growth is faster than the original one, but slower than the I-divergence.

5.2 An example

We consider now a simple example that highlights the extent to which the discussed conditions for existence can differ in applications. We focus on an unconditional example; its extension to the conditional case is immediate.

Consider projecting the normal $\mathcal{N}(\mu, \sigma^2)$ probability density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\}$$

onto the set $\mathcal{Q}$ of elements g that satisfy the mean zero and unit variance conditions:

$$\int x g(x) dx = 0, \text{ and } \int (x^2 - 1) g(x) dx = 0.$$

The two components of the moment function a that defines the projection set are thus: $a_1(x) = x$ and $a_2(x) = x^2 - 1$.

Consider first the case in which the divergence used to project f onto $\mathcal{Q}$ is the I-divergence. Recall that the I-divergence satisfies the growth condition in Assumption A4. Under Theorem 3, the I-divergence projection exists and is unique if Assumption A5 holds, that is, if

$$\int \exp\left[\tau_1 |x|\right] f(x) dx < +\infty, \quad \int \exp\left[\tau_2 |x^2 - 1|\right] f(x) dx < +\infty.$$

for every $\tau_1 > 0$ and every $\tau_2 > 0$, respectively. The first integral is finite for every $\tau_1 \in \mathbb{R}$ in view of the fact that it is the moment generating function a folded normal distribution which is everywhere

finite. The second integral fails to be finite for every $\tau_2 > 0$, since its minorant

$$\int \exp\left[\tau_2(x^2-1)\right] f(x)dx = \frac{\exp\left[\tau_2\left(\frac{\mu^2}{1-2\sigma^2\tau_2}-1\right)\right]}{\sqrt{1-2\sigma^2\tau_2}},$$

is finite only for $\tau_2 < 1/(2\sigma^2)$.

Although Assumption A5 is not satisfied, the projection of f onto $\mathcal{Q}$ with the I-divergence exists and it is given by a normal distribution with mean zero and unit variance. This follows from the fact that

$$g^0(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\tau_1 a_1(x) + \tau_2 a_2(x) + \tau_3\right]\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

is the density of standard normal distribution—and thus $g^0 \in \mathcal{Q}$—for $\tau_1 = -\mu/\sigma^2$, $\tau_2 = (1-\sigma^2)/(2\sigma^2)$, and $\tau_3 = (1+\mu^2+\sigma^2)/(2\sigma^2)$. That g^0 is the projection follows from Theorem 3 of Csiszar (1975).

It is straightforward to show that for the reverse I-divergence and, in general, for all members of the Cressie-Read family of divergence with $a < 0$ Assumption A5 fails because for these divergences $\phi^*(\tau_1|x|) = +\infty$ for every $x \in (-\frac{1}{a\tau_1}, \frac{1}{a\tau_1})$ and all $\tau_1 > 0$. The projection of f onto $\mathcal{Q}$ could still exist in this case, but since A5 does not hold we are not able to establish it rigorously.

When f is projected onto $\mathcal{Q}$ using one of the modified divergences defined in Section 5.1 checking whether the conditions for existence and characterization are satisfied is simpler. Define $B_\tau = \{x \in \mathbb{R} : \phi^*\left(\tau_1|a_1(x)| + \tau_2|a_2(x)| + \tau_3\right) < +\infty\}$ and let B_τ^c denote its complement. For the modified divergences Assumption A5 becomes

$$\int_{B_\tau} \phi^*\left(\tau_i|a_i(x)|\right) f(x)dx + a_\vartheta \int_{B_\tau^c} \left(\tau_i|a_i(x)|\right)^2 f(x)dx + b_\vartheta \int_{B_\tau^c} \tau_i|a_i(x)| f(x)dx < +\infty,$$

for every $\tau_i > 0$, $i = 1, 2$. By the continuity of a_1 and a_2 the set B_τ is μ-measurable; furthermore, $\int_{B_\tau} f(x)dx < \infty$. As a consequence, Assumption A5 holds in this case if the two rightmost terms of the previous display are finite. A sufficient condition for the two terms to be finite is that the fourth non-centered moment of X is finite, which is the case here because under f all moments of X exist.

The example of this section highlights several important points that hold more generally. First, it is important to have a portfolio of conditions for the projection existence that can be used in different situations. Second, with the modified divergences Assumption A5 reduces to the existence of higher moments of the functions characterizing the set $\mathcal{Q}$. This is a useful feature especially

in econometric applications because conditions on the existence of higher moments of the moment function a are often required for establishing the limit behavior of estimators based on sample projections. Last, but not least, the conditions proposed here are all sufficient but not necessary for the projection to exist.

5.3 Conclusions

In this paper we give sufficient conditions under which the projection of a conditional density onto a set defined by conditional moment restrictions exists, and can be characterized in terms of the dual of the original projection problem. The primitive conditions relate to the properties of the function ϕ defining the divergence, and to the existence of certain higher moments of the moment function a. Both sets of conditions can be relatively easily checked in specific applications. It is worth mentioning that, unlike most of the literature, our setup allows for unbounded moment functions. Our results can be thought as extensions of the results available for the unconditional case, e.g., Borwein and Lewis (1991a, 1993) and Csiszar (1995). The extension is, however, not trivial as each conditional moment constraint can be thought as an infinity of unconditional moment constraints.

The extension is of particular interest in econometric applications where the moments are usually unbounded. Many are the practical cases where our results are relevant. When the moment conditions depend on a parametric component the conditional projections correspond to the population counterpart of the objective function of semiparametric efficient estimator (Kitamura et al., 2004). Having sufficient conditions for the existence of the limit of the objective is the building step to study the properties of these estimators under misspecification. Other immediate applications of the results here are to cases in which one wants to recover a likelihood from a set from conditional moment restrictions. Typical examples are DSGE models, which in their nonlinear form can be cast in terms of a set of restrictions on the conditional expectations of the control variables. The projection studied here could provide a way to conduct likelihood-based inference in this class of models. More generally, the conditional projection here could be useful in all those cases in which additional information is available in terms of conditional moment conditions: the projection of an initial conditional distribution onto the set defined by these conditions could form the basis for inference. We have pointed in the introduction to several works already use these ideas, but additional research assessing the viability of these approaches could be a welcoming addition to the literature. Regardless of the specific application, finding the projection density is a formidable computational task. Devising efficient algorithms for computing the projections would also be an interesting area

of research. Equally interesting would also be extending the existence and characterization results to the more general situation in which the projection set is defined by inequality restrictions.

A Preliminaries

A.1 Conditional measures and densities

Let $(\Omega, \mathcal{F}, P)$ be a probability space and suppose that $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$. When $\mathcal{G}$ is the trivial σ-field, i.e. $\mathcal{G} \equiv \{\emptyset, \Omega\}$, then we deal with an unconditional case; otherwise, the problem is conditional. Here, we do not put any restrictions on $\mathcal{G}$ other than $\mathcal{G} \subset \mathcal{F}$, so our setup accommodates both the conditional and the unconditional problem. Further, let $(\mathbb{E}, \mathcal{E})$ be a measurable space in which $\mathbb{E}$ is a complete separable metric space and $\mathcal{E}$ is the σ-algebra of Borel sets. Then, given an $\mathcal{F}$-measurable random element $X : \Omega \to \mathbb{E}$ we shall be interested in the regular conditional measure of X given $\mathcal{G}$, which we denote by μ. That μ is a regular conditional measure means that $\mu : \Omega \times \mathcal{E} \to \mathbb{R}_+$ satisfies: (i) for each $B \in \mathcal{E}$, $\omega \mapsto \mu(\omega, B)$ is a version of $P(X(\omega) \in B|\mathcal{G})$, and (ii) for a.e. ω, $B \mapsto \mu(\omega, B)$ is a probability measure on $(\mathbb{E}, \mathcal{E})$.

Let ν be a σ-finite measure on $(\mathbb{E}, \mathcal{E})$. For instance, the measurable space could be $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, with ν being the Lebesgue measure; or if the set $\mathbb{E}$ is countable, we could let $\mathcal{E}$ be the set of subsets of $\mathbb{E}$, with ν being the counting measure. We shall assume that for a.e. ω, $\mu(\omega, \cdot)$ is absolutely continuous with respect to ν. Then, there exists $f : \Omega \times \mathbb{E} \to \mathbb{R}_+$ that is $(\mathcal{G} \otimes \mathcal{E}, \mathcal{B}(\mathbb{E}))$-measurable and such that for a.e. ω we have:

$$\mu(\omega, B) = \int_B f(\omega, x) d\nu(x), \tag{20}$$

i.e. f is the conditional density of X given $\mathcal{G}$. Note that since no requirements other than σ-finiteness are put on the dominating measure ν, our setup accommodates continuous as well as discrete random variables. In particular, our setup accommodates the case in which μ is the empirical measure.[15] It follows from joint measurability of f that f is jointly integrable with respect to the product measure $P \times \nu$ (see, e.g., Theorem 11.28 in Aliprantis and Border, 2007), so

$$\int f d(P \times \nu) = \int_{\mathbb{E}} \int_{\Omega} f(\omega, x) dP(\omega) d\nu(x) = \int_{\Omega} \int_{\mathbb{E}} f(\omega, x) d\nu(x) dP(\omega) = 1, \tag{21}$$

where the last equality uses (20).

Now, for $1 \leqslant p < \infty$, let $L_p^\nu(\mathcal{F} \otimes \mathcal{E})$ be the space of (equivalence classes of) functions $g : \Omega \times \mathbb{E} \to \mathbb{R}$ that are $(\mathcal{F} \otimes \mathcal{E}, \mathcal{B}(\mathbb{R}))$-measurable and such that $|g|^p$ is $P \times \nu$-integrable. We use the superscript ν in L_p^ν to emphasize that integrability needs to hold with respect to $P \times \nu$. Later on, we shall introduce

[15] Though in this case the support of the dominating counting measure ν is data dependent.

spaces with integrability conditions with respect to $P \times \mu$ where as before μ is the conditional measure corresponding to the conditional density f. Such spaces will bear a superscript μ. For any $g \in L_p^\nu(\mathcal{F} \otimes \mathcal{E})$, the L_p^ν-norm of g is defined by:

$$\|g\|_p^\nu \equiv \left[\int |g|^p d(P \times \nu) \right]^{1/p}.$$

The L_∞^ν-norm (or the essential sup norm) of g is defined as: $\|g\|_\infty^\nu \equiv \inf \{M > 0 : |g(\omega, x)| \leq M \text{ for } P\text{-a.e. } \omega, \text{ and } \nu\text{-a.e. } x\}$. We say that two elements g_1 and g_2 of $L_p^\nu(\mathcal{F} \otimes \mathcal{E})$ belong to the same equivalence class—property which we denote $g_1 = g_2$ a.s.—if $\|g_1 - g_2\|_p = 0$.

It follows from (21) that $f \in L_1^\nu(\mathcal{F} \otimes \mathcal{E})$. In what follows, we shall work with the $L_1^\nu(\mathcal{F} \otimes \mathcal{E})$ space equipped with the L_1^ν-norm $\|\cdot\|_1$, which is a Banach space. The set of functions $h \in L_1^\nu(\mathcal{F} \otimes \mathcal{E})$ that are $(\mathcal{G} \otimes \mathcal{E}, \mathcal{B}(\mathbb{R}))$-measurable forms a closed subspace of $L_1^\nu(\mathcal{F} \otimes \mathcal{E})$ that we denote by $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$.[16] In particular, we shall be interested in those elements of $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ that are nonnegative valued, so we let $\mathcal{P} \equiv \{g \in L_1^\nu(\mathcal{G} \otimes \mathcal{E}) : g(\Omega \times \mathbb{E}) \subseteq \mathbb{R}_+\}$ be the positive cone in $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$. It follows that $f \in \mathcal{P}$. Hereafter, we shall reserve the letter f to denote conditional densities in $\mathcal{P}$; elements of the space $L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ (not necessarily non-negative) shall be denoted by g.

Finally, we pay particular attention to those elements $g \in L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ whose supports are included in that of f. More formally, if the support of g is included in that of f we shall denote this property by $g \prec\!\!\prec f$; more formally,

$$g \prec\!\!\prec f \quad \text{if for } P\text{-a.e. } \omega \text{ and } \nu\text{-a.e. } x,\ f(\omega, x) = 0 \text{ implies } g(\omega, x) = 0.$$

Note that this property is equivalent to the property of absolute continuity between the corresponding measures. Indeed, if for every $B \in \mathcal{E}$, we let $\mu^g(\omega, B) \equiv \int_B g(\omega, x) d\nu(x)$, then $\mu^g(\omega, \cdot)$ is a finite signed measure on $(\mathbb{E}, \mathcal{E})$. If $A(\omega)$ is a null set of $\mu(\omega, \cdot)$, i.e. if $\int_A f(\omega, x) d\nu(x) = 0$, then for ν-a.e. $x \in A$, $f(\omega, x) = 0$. When $g \prec\!\!\prec f$, the latter implies that for ν-a.e. $x \in A$, $g(\omega, x) = 0$, i.e. $\int_B g(\omega, x) d\nu(x) = 0$ for every $B \subseteq A$, so A is a null set of $\mu^g(\omega, \cdot)$; in other words, $\mu^g(\omega, \cdot)$ is absolutely continuous with respect to $\mu(\omega, \cdot)$, i.e. $\mu^g(\omega, \cdot) \ll \mu(\omega, \cdot)$. By the Radon-Nykodym theorem (see, e.g., Thereom 13.18 in Aliprantis and Border, 2007) there then exists a ($P \times \mu$-almost) unique function π that is $(\mathcal{G} \otimes \mathcal{E}, \mathcal{B}(\mathbb{R}))$-measurable and $P \times \mu$-integrable such that

$$g = \pi f.$$

[16] When the conditioning is done with respect to a sub-σ-field generated by a subvector of X, then the above L_1^ν-norm induces the metric of " integrated L_1^ν-distance" used in Tang and Ghosal (2007).

Moreover, since g is in $L_1^\nu(\mathcal{G}\otimes\mathcal{E})$, it follows that $\int|\pi|d(P\times\mu)=\int|g|d(P\times\nu)<+\infty$ so $\pi\in L_1^\mu(\mathcal{G}\otimes\mathcal{E})$, which is the space of (equivalence classes of) functions $\pi:\Omega\times\mathbb{E}\to\mathbb{R}$ that are $(\mathcal{G}\otimes\mathcal{E},\mathcal{B}(\mathbb{R}))$-measurable and such that $|\pi|$ is $P\times\mu$-integrable. In what follows, we shall often use this fact which allows us to transform our projection problem stated in terms of g into a constrained optimization problem in π.

A.2 Convex Functions and Their Conjugates

Most of our analysis to follow involves real convex functions and their conjugates. To start, we recall some useful concepts from convex analysis; for a detailed discussion, see, e.g., Rockafellar (1970) and Hiriart-Urruty and Lemarechal (1993). We consider non-negative valued real functions $\phi:(0,+\infty)\to[0,+\infty)$ with the following properties:

Assumption A1. *(i) ϕ is twice continuously differentiable on $(0,+\infty)$; (ii) ϕ is strictly convex on $(0,+\infty)$; (iii) $\phi(1)=\phi'(1)=0$; (iv) $\lim_{u\to0^+}\phi'(u)<0$; (v) $\lim_{u\to+\infty}\phi'(u)>0$.*

Assumptions A1(i)-(iii) are fairly standard. In particular, the normalizations $\phi(1)=\phi'(1)=0$ and $\phi'(1)=1$ do not restrict generality, since for any differentiable convex function ϕ there exists another, say $\overline{\phi}$, satisfying $\overline{\phi}(1)=\overline{\phi}'(1)=0$.

It is convenient to view ϕ as an extended-real valued function, defined on $\mathbb{R}$ and taking values in $[0,+\infty]$ (see, e.g. p. 23 in Rockafellar, 1970). This means that the convex function ϕ being defined a priori on $(0,+\infty)$ we can extend it outside its domain by setting $\phi(u)=+\infty$ for all $u\in(-\infty,0)$. As for the boundary value of zero, we let $\phi(0)=\lim_{u\to0^+}\phi(u)$, knowing that this limit is possibly $+\infty$.[17] This ensures that the extension of ϕ is lower-semicontinuous on $\mathbb{R}$ (or "closed" in the terminology of Rockafellar (1970)). Note that since by Assumption A1(ii) ϕ is convex on $(0,+\infty)$, its extension is convex on $\mathbb{R}$. Further, to deal with zero and infinity we adopt the understanding that $\phi(+\infty)=\lim_{u\to+\infty}\phi(u)$, $\phi'(0)=\lim_{u\to0^+}\phi'(u)$, $\phi'(+\infty)=\lim_{u\to+\infty}\phi'(u)$, and $0\cdot\phi\left(\frac{0}{0}\right)=0$.

The conjugate of the convex extended-real valued function ϕ on $\mathbb{R}$, ϕ^*, is itself a convex lower semi-continuous function. Moreover, it follows from the above definition, that ϕ^* is increasing on $\mathbb{R}$. The following result will play an important role in what follows.

Lemma 4.2 in Borwein and Lewis (1991a). *Let $\phi:[0,+\infty)\to[0,+\infty]$ be convex, lower semi-continuous, and such that $\phi(x)<+\infty$ for at least one $x\in[0,+\infty)$. Define $d\equiv\lim_{u\to+\infty}\phi(u)/u$.*

[17] Note that the non-negativity of ϕ is also ensured by the strict convexity of ϕ on $(0,+\infty)$, and the requirements in Assumption A1(iii,iv).

Then $\overline{\operatorname{dom}\phi^*} = (-\infty, d]$ *where* $\operatorname{dom}\phi^* \equiv \{v \in \mathbb{R} : \phi^*(v) < +\infty\}$ *denotes the effective domain of* ϕ^*.

Since ϕ is differentiable on $(0, +\infty)$, we can relate its conjugate to its Legendre transform. For this, we first need to define the image by ϕ of its derivative ϕ'. Let $I_{\phi'}$ denote the image $I_{\phi'} \equiv \phi'((0, +\infty)) = \{v \in \mathbb{R} : v = \phi'(u), u \in (0, +\infty)\}$. Under assumptions A1(i) and (ii), ϕ' is continuous and strictly increasing on $(0, +\infty)$; hence $I_{\phi'} = (\phi'(0), \phi'(+\infty))$. Under assumptions A1(iv) and (v), 0 belongs to $I_{\phi'}$. Note that if Assumptions A1(iv)-(v) hold with $\lim_{u\to 0^+} \phi'(u) = -\infty$ and $\lim_{u\to+\infty} \phi'(u) = +\infty$, then $I_{\phi'} = \mathbb{R}$. The Legendre-Fenchel transform of ϕ is a real mapping $\tilde{\phi} : I_{\phi'} \to \mathbb{R}$ which to every $v \in I_{\phi'}$ associates:

$$\tilde{\phi}(v) \equiv v(\phi')^{-1}(v) - \phi\left((\phi')^{-1}(v)\right).$$

The following lemma establishes several useful properties of $\tilde{\phi}$.

Lemma 6. *Under Assumption A1, we have: (i)* $\tilde{\phi}$ *is twice continuously differentiable on* $I_{\phi'}$*, (ii)* $\tilde{\phi}$ *is strictly convex on* $I_{\phi'}$*, (iii) for any* $v \in I_{\phi'}$*,* $\tilde{\phi}(v) > 0$ *whenever* $v > 0$*, (iv)* $\tilde{\phi}' > 0$ *on* $I_{\phi'}$*, (v)* $\tilde{\phi}'(v) = (\phi')^{-1}(v)$ *for any* $v \in I_{\phi'}$*, (vi)* $\tilde{\phi}''(v) = [\phi''((\phi')^{-1}(v))]^{-1}$ *for any* $v \in I_{\phi'}$.

Proof of Lemma 6. As already noted, assumptions A1(i) and (ii) imply that ϕ' is a homeomorphism, hence an open map. This means in particular that the image of ϕ', $I_{\phi'}$, is open. Now, note that from the expression of the Legendre conjugate, $\tilde{\phi}$ is continuous and differentiable on $I_{\phi'}$. In addition, the derivative of $\tilde{\phi}$ is given by:

$$\tilde{\phi}'(v) = (\phi')^{-1}(v), \text{ for any } v \in I_{\phi'}.$$

Given the strict convexity of ϕ in Assumption A1(ii), ϕ' is continuous and strictly increasing on $(0, +\infty)$; so its inverse $\tilde{\phi}'$ is continuous and strictly increasing on $I_{\phi'}$. Hence, $\tilde{\phi}$ is strictly convex. Since $\lim_{u\to 0} \phi'(0) = \inf I_{\phi'}$ we have $\lim_{v\to\inf I_{\phi'}} (\phi')^{-1}(v) = 0$, i.e. $\lim_{v\to\inf I_{\phi'}} \tilde{\phi}'(v) = 0$. This combined with the fact that $\tilde{\phi}'$ is continuous and strictly increasing on $I_{\phi'}$ then gives $\tilde{\phi}' > 0$ on $I_{\phi'}$. Now, under assumption A1(iii) we have $\tilde{\phi}(0) = 0$. Then, for any $v \in I_{\phi'}$, we have $\tilde{\phi}(v) > 0$ if $v > 0$. Finally, A1(ii) implies $\phi'' > 0$ on $(0, +\infty)$ so $\tilde{\phi}'$ is continuously differentiable on $I_{\phi'}$ with derivative:

$$\tilde{\phi}''(v) = \frac{1}{\phi''((\phi')^{-1}(v))}.$$

This completes the proof of Lemma 6. □

Recall that we can extend ϕ to be a lower semi-continuous convex function on all $\mathbb{R}$ with $(0, +\infty)$ as its effective domain. Thus, we can relate the Legendre conjugate $\tilde{\phi}$ to the ordinary conjugate of the extended ϕ as: $\tilde{\phi} = \phi^*|_{I_{\phi'}}$, i.e. $\tilde{\phi}$ is simply the restriction of ϕ^* to $I_{\phi'}$ (see, e.g., Theorem 26.4 in Rockafellar, 1970).

The following family of functions $\phi \in \mathcal{K}$ introduced by Cressie and Read (1984) is of particular interest in econometrics:

$$\phi_a(u) = \begin{cases} \frac{u^{a+1}-1}{a(a+1)} - \frac{1}{a}u + \frac{1}{a}, & u > 0 \\ \frac{1}{a+1}, & u = 0 \\ +\infty, & u < 0 \end{cases}, \quad a \in A \subset \mathbb{R} \tag{22}$$

with $0^a = 0$ for $a > -1$, $0^a = +\infty$ for $a < -1$, $\phi_{-1}(u) = \lim_{a \to -1} \phi_a(u)$, and $\phi_0(u) = \lim_{a \to 0} \phi_a(u)$, and where the set A contains those $a \in \mathbb{R}$ for which ϕ_ϕ satisfies Assumption A1. Note that the image of ϕ'_a depends on the particular value of a, and we have $I_{\phi'_a} = (-1/a, +\infty)$ if $a > 0$, $I_{\phi'_a} = \mathbb{R}$ if $a = 0$, and $I_{\phi'_a} = (-\infty, -1/a)$ if $a < 0$. The conjugate of ϕ_a is given by

$$\phi^*_\alpha(v) = \begin{cases} \frac{1}{1+a}(1+av)^{(1+a)/a} - \frac{1}{1+a}, & v \in I_{\phi'_a} \\ +\infty, & v \notin I_{\phi'_a} \end{cases}.$$

B Proofs of the results stated in the main text

Proof of Theorem 1. Fix $f \in \mathcal{P}$, and consider any $g \in L_1^\nu(\mathcal{G} \otimes \mathcal{E})$. If the support of g does not contain that of f then $\mathcal{D}_\phi(g, f) = +\infty$ and so is well defined; thus we only need to consider the case in which the support of g contains that of f, i.e. $g \prec\!\!\prec f$. We then have

$$\begin{aligned} \mathcal{D}_\phi(g, f) &= \int_\Omega \int_{\mathbb{E}} f(\omega, x) \phi\left(\frac{g(\omega, x)}{f(\omega, x)}\right) d\nu(x) dP(\omega) \\ &= \int_\Omega \int_{\mathcal{A}(\omega)} f(\omega, x) \phi\left(\frac{g(\omega, x)}{f(\omega, x)}\right) d\nu(x) dP(\omega) \\ &= \int_\Omega \int_{\mathcal{A}(\omega)} \phi\left(\frac{g(\omega, x)}{f(\omega, x)}\right) d\mu(\omega, x) dP(\omega) \\ &= \int_\Omega \int_{\mathbb{E}} \phi\left(\frac{g(\omega, x)}{f(\omega, x)}\right) d\mu(\omega, x) dP(\omega), \end{aligned}$$

where $\mathcal{A}(\omega) \equiv \{x \in \mathbb{E} : f(\omega, x) > 0\}$, the second equality follows by $0 \cdot \phi(0/0) = 0$, the third by change of measure, and the fourth because $\mathcal{A}^c(\omega)$ is of $\mu(\omega, \cdot)$ measure zero.

Let $\pi \equiv g/f$ and note that π is well defined $P \times \mu$ a.e. Moreover, $|\pi|$ is $P \times \mu$-integrable since

$\int |\pi| d(P \times \mu) = \int (|g|/f) d(P \times \mu) = \int |g| d(P \times \nu)$. So consider the following functional defined on $L_1^\mu(\mathcal{G} \otimes \mathcal{E})$:

$$I_\phi(\pi) \equiv \int_\Omega \int_{\mathbb{E}} \phi\big(\pi(\omega, x)\big) d\mu(\omega, x) dP(\omega). \tag{23}$$

Then $I_\phi\left(\frac{g}{f}\right) = \mathcal{D}_\phi(g, f)$.

To show that $I_\phi(\pi)$ is well defined, we use the result of Theorem 1 in Rockafellar (1968). For this, first note that since ϕ is convex and lower semi-continuous on $\mathbb{R}$, it is a normal convex integrand (see, Lemma 1 in Rockafellar, 1968). Next, we need to show that there exists at least one $\pi^0 \in L_\infty^\mu(\mathcal{G} \otimes \mathcal{E})$ such that $I_{\phi^*}(\pi^0) < +\infty$, i.e.

$$\int_\Omega \int_{\mathbb{E}} \phi^*(\pi^0(\omega, x)) d\mu(\omega, x) dP(\omega) < +\infty, \tag{24}$$

where $\|\pi^0\|_\infty^\mu \equiv \inf\{M > 0 : |\pi^0(\omega, x)| \leq M, \text{ for } P\text{-a.e. } \omega, \text{ and } \mu\text{-a.e. } x\} < \infty$. Now, take any $v \in \mathbb{R}$ such that $\phi^*(v) < +\infty$, and let $\pi^0(\omega, x) = v$. Then, $\|\pi^0\|_\infty^\mu = |v| < +\infty$, and

$$\int_\Omega \int_{\mathbb{E}} \phi^*(\pi^0(\omega, x)) d\mu(\omega, x) dP(\omega) = \phi^*(v) \int_\Omega \int_{\mathbb{E}} d\mu(\omega, x) dP(\omega) = \phi^*(v) < +\infty,$$

which shows (24). We can now apply Theorem 1 in Rockafellar (1968) to show that $I_\phi(\pi)$ is a well defined convex function on $L_1^\mu(\mathcal{G} \otimes \mathcal{E})$ with values in $(-\infty, +\infty]$. That $I_\phi(\pi)$ is strictly convex on its effective domain follows by the strict convexity of ϕ on $(0, +\infty)$.

It remains to show that $\mathcal{D}_\phi(g, f) \geqslant 0$ with equality only if $g = f$ with probability one. Take any $g \in L_1^\nu(\mathcal{G} \otimes \mathcal{E})$ such that $\mathcal{D}_\phi(g, f) < +\infty$; then necessarily $g \in \mathcal{P}$. Since $g \in \mathcal{P}$, we have that $\int g d(P \times \nu) = 1$, so from Jensen's inequality, we obtain

$$\int f \phi\left(\frac{g}{f}\right) d(P \times \nu) \geq \phi\left(\int f \frac{g}{f} d(P \times \nu)\right) = \phi\left(\int g d(P \times \nu)\right) = \phi(1) = 0,$$

with equality only if $g = f$ with probability one. □

Proof of Lemma 1. First, note that since $1 \in L_1^\mu(\mathcal{G} \otimes \mathcal{E})$, we have $\inf_{\pi \in L_1^\mu(\mathcal{G} \otimes \mathcal{E})} I_\phi(\pi) = 0$. To show that under Assumption A2, the level sets of I_ϕ are weakly compact, we use Theorem 2.7 in Borwein and Lewis (1991b). Specifically, since the measure $P \times \mu$ is finite Theorem 2.7(B) applies, and the level sets of I_ϕ of the form $\{\pi \in L_1^\mu(\mathcal{G} \otimes \mathcal{E}) : I_\phi(\pi) \leqslant d\}$, $d > 0$, are weakly compact. That Assumption A2 is also necessary when ν is not purely atomic (so that $P \times \mu$ is not purely atomic) follows by Theorem 2.10 in Borwein and Lewis (1991b) by noting that finiteness of ϕ^* is equivalent

to the growth condition in Assumption A2. □

Proof of Lemma 2. We show that under Assumption A3, the projection set $\mathcal{C}$ is closed in the norm topology of $L_1^\mu(\mathcal{G}\otimes\mathcal{E})$. For this, let $\{\pi_i\}$ be any convergent sequence in $\mathcal{C}$, and denote by $\bar{\pi}$ its limit, $\lim_{i\to\infty}\|\pi_i-\bar{\pi}\|_1^\mu=0$. We now show that then $\bar{q}\in\mathcal{C}$, i.e. the set $\mathcal{C}$ is closed. We have:

$$\begin{aligned}\int_\Omega\left|\int_{\mathbb{E}}a(\omega,x)\bar{\pi}(\omega,x)dx\right|dP(\omega) &\leqslant \int_\Omega\left|\int_{\mathbb{E}}a(\omega,x)[\bar{\pi}(\omega,x)-\pi_i(\omega,x)]dx\right|dP(\omega)\\ &\quad+\int_\Omega\left|\int_{\mathbb{E}}a(\omega,x)\pi_i(\omega,x)dx\right|dP(\omega)\\ &=\int_\Omega\left|\int_{\mathbb{E}}a(\omega,x)[\bar{\pi}(\omega,x)-\pi_i(\omega,x)]dx\right|dP(\omega)\\ &\leqslant\int_\Omega\int_{\mathbb{E}}|a(\omega,x)|\cdot|\bar{\pi}(\omega,x)-\pi_i(\omega,x)|dxdP(\omega)\\ &\leqslant M\|\pi_i-\bar{\pi}\|_1^\mu\end{aligned}$$

where the first equality uses $\pi_i\in\mathcal{Q}(\theta)$, and the last inequality follows by Assumption A3. Taking the limit of the above as $i\to\infty$ it then follows that

$$\int_\Omega\left|\int_{\mathbb{E}}a(\omega,x)\bar{\pi}(\omega,x)dx\right|dP(\omega)=0$$

and since the quantity inside the first integral is everywhere non-negative, the above implies that for a.e. ω,

$$\int_{\mathbb{E}}a(\omega,x)\bar{\pi}(\omega,x)dx=0$$

Hence, $\bar{\pi}\in\mathcal{C}$. □

Proof of Theorem 2. Since the problem is assumed feasible, $\inf_{\pi\in\mathcal{C}}I_\phi(\pi)=d<+\infty$. We need to show that there exists $\pi^0\in\mathcal{C}$ such that $I_\phi(\pi^0)=d$. For this, consider

$$\mathcal{C}_d\equiv\{\pi\in\mathcal{C}:I_\phi(\pi)\leqslant 2d\},\tag{25}$$

and let $\{\pi_i\}$ be a sequence in $\mathcal{C}_d$ for which

$$\lim_{i\to\infty}I_\phi(\pi_i)=\inf_{\pi\in\mathcal{C}}I_\phi(\pi)=d.\tag{26}$$

Note that $\mathcal{C}_d=\mathcal{C}\cap\mathcal{L}_{2d}$ where $\mathcal{L}_d\equiv\{\pi\in L_1^\mu(\mathcal{G}\otimes\mathcal{E}):I_\phi(\pi)\leqslant d\}$. Weak sequential compactness of level sets $\mathcal{L}_d$ established in Lemma 1 implies that there exists a subsequence π_{i_k} tending weakly to

some $\pi^0 \in \mathcal{L}_{2d}$. Now, by Lemma 2 $\mathcal{C}$ is closed, and since in addition $\mathcal{C}$ is convex, it is also weakly closed. Thus, the limit π^0 of the subsequence must be in $\mathcal{C}$. It remains to be shown that π^0 is a solution to the problem (5), i.e. that $I_\phi(\pi^0) = d$. This follows by the weak lower semi-continuity of I_ϕ established using Lemma 1, since $I_\phi(\pi^0) \leqslant \liminf_k I_\phi(g_{i_k}) = d$. Uniqueness follows by the strict convexity of I_ϕ on its effective domain. □

Proof of Lemma 3. Consider a minimizing sequence $\{\pi_n\} \in \mathcal{C}$, $I_\phi(\pi_n) \to \inf_{\pi \in \mathcal{C}} I_\phi(\pi) = d$. If $d < +\infty$, then with no loss of generality, we may assume all $I_\phi(\pi_n)$ to be finite. Then necessarily $\pi_n(\omega, x) \geqslant 0$ for P-a.e. ω and μ-a.e. x. Now let $\rho(u) = \phi(1+u)$, and note that under Assumptions A1(ii,iii) and A2, ρ satisfies the needed convexity and limit conditions. Then, it follows from strict convexity of ϕ on $(0, +\infty)$ that for $\alpha_0 = 1/2$,

$$\begin{aligned}\int \phi\left(1 + \frac{1}{2}\pi_n\right) d(P \times \mu) &< \frac{1}{2}\left[\phi(2) + \int \phi\left(\pi_n\right) d(P \times \mu)\right] = \frac{1}{2}\left[\phi(2) + I_\phi(\pi_n)\right] \\ &< +\infty,\end{aligned}$$

i.e. $\{\pi_n\} \in L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$. □

Proof of Theorem 3. The beginning of the proof is similar to that of Theorem 2: consider a sequence $\{\pi_i\}$ in $\mathcal{C}_d$ defined in Equation (25) which satisfies the property in Equation (26). Letting $\rho(u) = \phi(1+u)$ we have that for every i, $\pi_i \in L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$ and

$$\begin{aligned}\|\pi_i\|^\mu_\rho &\leqslant \inf_{k>0} \frac{1}{k}\left(1 + \int \rho\left(k|\pi_i|\right) d(P \times \mu)\right) \leqslant 2\left[1 + \int \phi\left(1 + \frac{1}{2}\pi_i\right) d(P \times \mu)\right] \\ &\leqslant 2\left[1 + \frac{1}{2}\left(\phi(2) + \int \phi\left(\pi_i\right) d(P \times \mu)\right)\right] = 2 + \phi(2) + I_\phi(\pi_i) \leqslant 2 + \phi(2) + 2d,\end{aligned}$$

where the first inequality is a classical inequality between the Luxembourg norm and the so-called Orlicz norm (see, Krasnosel'skii and Rutickii, 1961). Thus, the sequence $\{\pi_i\}$ is bounded. It follows from Theorem 14.4 in Krasnosel'skii and Rutickii (1961) that there exists a subsequence π_{i_k} tending E-weakly to some $\pi^0 \in L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$.

We have already shown (see the discussion preceding the statement of the theorem) that $1 \in E^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$ implies that π^0 satisfies $\int_{\mathbb{E}} \pi^0(\omega, x) d\mu(x) = 1$ for a.e. ω. We now repeat the reasoning with Assumption A5. First, since every component a_j $(1 \leqslant j \leqslant m)$ of the moment function a is in

$E^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$, it follows that

$$\lim_{k\to\infty}\int a_j\pi_{i_k}d(P\times\mu)=\int a_j\pi^0 d(P\times\mu)=0,$$

where the second equality follows because $\pi_{i_k}\in\mathcal{C}$. In particular, it follows from Hölder's inequality (see Equation (7)) that $a_j\pi^0\in L^{\mu}_1(\mathcal{G}\otimes\mathcal{E})$. So letting

$$m(\omega)\equiv\int_{\mathbb{E}}a_j(\omega,x)\pi^0(\omega,x)d\mu(x), \tag{27}$$

we have

$$\int_{\Omega}|m(\omega)|dP(\omega)\leqslant\int_{\Omega}\int_{\mathbb{E}}|a_j(\omega,x)\pi^0(\omega,x)|d\mu(x)dP(\omega)<+\infty,$$

so $m\in L^m_1(\mathcal{G})$. We now show that for every $l\in L^m_\infty(\mathcal{G})$, $\int_\Omega l(\omega)'m(\omega)dP(\omega)=0$, which will then imply that $m=0$ a.s. For this, take any $l\in L^m_\infty(\mathcal{G})$ and any $\tau>0$; letting $L\equiv\|l\|_\infty$, note that for any component j $(1\leqslant j\leqslant m)$,

$$\int\phi^*\left(\tau|a_jl_j|\right)d(P\times\mu)\leqslant\int\phi^*\left(\tau|a_j|L\right)d(P\times\mu)<+\infty,$$

where the first inequality follows because $\phi^*>0$ and $\phi^{*\prime}>0$ on $(0,+\infty)$ (Lemma 6), and the last inequality follows by Assumption A5. Thus $a_jl_j\in E^{\mu}_{\rho^*}(\mathcal{G}\otimes\mathcal{E})$. So for every component j $(1\leqslant j\leqslant m)$,

$$\begin{aligned}\lim_{k\to\infty}\int l_ja_j\pi_{i_k}d(P\times\mu)&=\lim_{k\to\infty}\int_{\Omega}l_j(\omega)\left[\int_{\mathbb{E}}a_j(\omega,x)\pi_{i_k}(\omega,x)d\mu(x)\right]dP(\omega)=0\\&=\int l_ja_j\pi^0d(P\times\mu)=\int_{\Omega}l_j(\omega)m_j(\omega)dP(\omega),\end{aligned}$$

which implies that $\int_\Omega l(\omega)'m(\omega)dP(\omega)=0$. Thus, $m=0$ a.s. and $\pi^0\in\mathcal{C}$. It remains to be shown that π^0 is a solution to the problem (5), i.e. that $I_\phi(\pi^0)=d$. This follows by the weak lower semi-continuity of I_ϕ in $L^{\mu}_\rho(\mathcal{G}\otimes\mathcal{E})$ shown using Lemma 4, since $\liminf_n I_\phi(\pi_n)\geqslant I_\phi(\pi^0)$. As before, uniqueness follows by the strict convexity of I_ϕ on its effective domain. □

Proof of Lemma 4. We first show that $L^{\mu}_\rho(\mathcal{G}\otimes\mathcal{E})\subseteq L^{\mu}_1(\mathcal{G}\otimes\mathcal{E})$. Since ρ is a proper convex function, there exist $s>0$ and $v\geqslant 0$ such that for every $u\in[0,+\infty)$, $\rho(u)\geqslant su-v$. In particular, for any $t\in\mathbb{R}$ and any $\alpha>0$, there exist $a_\alpha>0$ and $b_\alpha\geqslant 0$ such that $|t|\leqslant a_\alpha\rho(\alpha|t|)+b_\alpha$. Take

$h \in L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$ and the corresponding $\alpha_0 > 0$; then,

$$\int |h| d(P \times \mu) \leqslant a_{\alpha_0} \int \rho(\alpha_0 |h|) d(P \times \mu) + b_{\alpha_0} < +\infty,$$

i.e. $h \in L^\mu_1(\mathcal{G} \otimes \mathcal{E})$. Now, the Luxemburg norm $\| \cdot \|^\mu_\rho$ satisfies

$$\int \rho\left(\frac{|h|}{\|h\|^\mu_\rho}\right) d(P \times \mu) = 1$$

(see, e.g., II §9 in Krasnosel'skii and Rutickii, 1961). It is easy to see that under Assumption A1, $\rho(u) = \phi(1+u)$ is positive and strictly increasing on $(0, +\infty)$ so $k_1 \geqslant k_2 > 0$ if and only if $\int \rho(|h|/k_1) d(P \times \mu) \leqslant \int \rho(|h|/k_2)$. Using the same inequality as above, we have

$$\int \rho\left(\frac{|h|}{q\|h\|^\mu_1}\right) d(P \times \mu) \geqslant s \int \frac{|h|}{q\|h\|^\mu_1} d(P \times \mu) - v = \frac{s}{q} - v,$$

so choosing $0 < q \leqslant s/(1+v)$, we get

$$\int \rho\left(\frac{|h|}{q\|h\|^\mu_1}\right) d(P \times \mu) \geqslant 1 = \int \rho\left(\frac{|h|}{\|h\|^\mu_\rho}\right) d(P \times \mu),$$

which implies $q\|h\|^\mu_1 \leqslant \|h\|^\mu_\rho$ as desired. □

Proof of Theorem 4. The beginning of the proof is similar to that of Theorems 2 and 3. We again consider a sequence $\{\pi_i\}$ in $\mathcal{C}_d$ defined in Equation (25) which satisfies the property in Equation (26). As established in the proof of Theorem 3, the sequence $\{\pi_i\}$ is bounded. It then follows from E-weak compactness of $L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$ that there exists a subsequence π_{i_k} tending E-weakly to some $\pi^0 \in L^\mu_\rho(\mathcal{G} \otimes \mathcal{E})$. To show that $\pi^0 \in \mathcal{C}$ the proof is similar to that of Theorem 3. In particular, for any $l \in L^m_\infty(\mathcal{G})$, letting $L \equiv \|l\|_\infty$ and $\tau_L \equiv \tau_j / L$ we have

$$\int \phi^*\left(\tau_L |a_j l_j|\right) d(P \times \mu) \leqslant \int \phi^*\left(\tau_L |a_j| L\right) d(P \times \mu) = \int \phi^*\left(\tau_j |a_j|\right) d(P \times \mu) < +\infty,$$

where the first inequality follows because $\phi^* > 0$ and $\phi^{*\prime} > 0$ on $(0, +\infty)$ (Lemma 6), and the second inequality follows by Assumption A7. Thus $a_j l_j \in L^\mu_{\rho^*}(\mathcal{G} \otimes \mathcal{E})$. By E-weak convergence of

the subsequence π_{i_k} it then follows that

$$\lim_{k\to\infty}\int l_j a_j \pi_{i_k} d(P\times\mu) = \lim_{k\to\infty}\int_\Omega l_j(\omega)\left[\int_{\mathbb{E}} a_j(\omega,x)\pi_{i_k}(\omega,x)d\mu(x)\right]dP(\omega) = 0$$
$$= \int l_j a_j \pi^0 d(P\times\mu) = \int_\Omega l_j(\omega)m_j(\omega)dP(\omega),$$

with m as defined before in Equation (27). Thus, $m = 0$ a.s. and $\pi^0 \in \mathcal{C}$. That π^0 is a unique solution to the problem (5) follows by the same reasoning as in the proof of Theorem 3. □

Proof of Lemma 5. To establish the form of the dual, we apply the results of Rockafellar (1974) (see, e.g., Example 11 on p. 26-27). We only need to formally establish the convex conjugates of the functions $\delta(\cdot|\{c\})$ and I_ϕ. For the first, note that by definition, for any $\tau \in L_\infty^{m+1}$,

$$\begin{aligned}\delta^*(\tau) &= \sup_{x\in \mathrm{dom}\delta(\cdot|\{c\})}[\langle\tau,x\rangle - \delta(x|\{c\})] = \sup_{x=c}[\langle\tau,x\rangle - \delta(x|\{c\})]\\ &= \langle\tau,c\rangle = \int_\Omega \tau(\omega)'c dP(\omega)\\ &= \int_\Omega \tau_{m+1}(\omega)dP(\omega),\end{aligned}$$

since $c = (0,1)' \in \mathbb{R}^{m+1}$. For the second, we apply Theorem 2 in Rockafellar (1968) to the integral functional I_ϕ defined in (23); the result shows that $I_\phi^* = I_{\phi^*}$. For Rockafellar's (1968) result to go through, we need to check that there exists at least one $\pi^0 \in E_{\rho^*}^\mu(\mathcal{G}\otimes\mathcal{E})$ such that $I_{\phi^*}(\pi^0) < +\infty$ (this was established in the proof of Definition 1), and that there exists at least one $\pi \in L_\rho^\mu(\mathcal{G}\otimes\mathcal{E})$ such that $I_\phi(\pi) < +\infty$. For this, it suffices to take any point u such that $\phi(u) < +\infty$, and let $\pi(\omega,x) = u$. Then, $I_\phi(\pi) = \phi(u) < +\infty$. □

Proof of Theorem 5. We use the result of Theorem 8(v) Zălinescu (1999). For this, we need to show that our Assumption A8 implies the constraint qualification condition (14). Notice that the condition (14) can equivalently be written as $0 \in \mathrm{icr}\,(c - T\mathrm{dom}I_\phi)$. Now, since the set $(c - T\mathrm{dom}I_\phi)$ is convex, we can use the following fact (see, e.g., p.8 in Holmes, 1974): when A is convex, $a \in \mathrm{icr}(A)$ is equivalent to the requirement that for all $x \in A\setminus\{a\}$ there exists $y \in A$ such that $a \in (x,y)$. Note that necessarily $x \neq 0$ and $y \neq 0$. So we need to show that for all $u_1 \in (c - T\mathrm{dom}I_\phi)\setminus\{0\}$, there exists $u_2 \in (c - T\mathrm{dom}I_\phi)\setminus\{0\}$ such that $0 \in (u_1,u_2)$. Since $u_1 = c - T\pi_1$ with $\pi_1 \in \mathrm{dom}I_\phi$ and $u_2 = c - T\pi_2$ with $\pi_2 \in \mathrm{dom}I_\phi$, $0 \in (u_1,u_2)$ is equivalent to $c \in (T\pi_1, T\pi_2)$. Now, recall that $u_1 \neq 0$, i.e. $T\pi_1 \neq c$ so $\pi \notin \mathcal{C}$; similarly, $\pi_2 \notin \mathcal{C}$. Thus, a sufficient condition is that for all $\pi_1 \in \mathrm{dom}I_\phi \setminus \mathcal{C}$, there exists $\pi_2 \in \mathrm{dom}I_\phi \setminus \mathcal{C}$ such that $\pi_0 \in (\pi_1,\pi_2)$ with $\pi_0 \in \mathcal{C}$, i.e. $T\pi_0 = c$, which

is what Assumption A8 states. □

References

Ali, S. M. and S. D. Silvey (1966): "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society Ser. B*, 28, 131–142.

Aliprantis, C. D. and K. C. Border (2007): *Infinite Dimensional Analysis*, Berlin: Springer-Verlag, 3rd ed.

Bernardo, J. (1979): "Reference Posterior Distributions for Bayesian Inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 113–147.

——— (2005): "Reference analysis," *Handbook of Statistics*, 25, 17–90.

Borwein, J. M. and A. S. Lewis (1991a): "Duality Relationships for Entropy-Like Minimization Problems," *SIAM Journal on Control and Optimization*, 29, 325–338.

——— (1991b): "On the Convergence of Moment Problems," *Transactions of the American Mathematical Society*, 325, pp. 249–271.

——— (1992a): "Partially finite convex programming, part I: quasi relative interiors and duality theory," *Math. Program.*, 57, 15–48.

——— (1992b): "Partially finite convex programming, part II: explicit lattice models," *Math. Program.*, 57, 49–83.

——— (1993): "Partially-finite Programming in L_1 and the Existence of Maximum Entropy Estimates," *SIAM Journal on Optimization*, 3, 248–267.

Buck, B. and V. Macaulay (1991): *Maximum entropy in action: a collection of expository essays*, Oxford University Press.

Burg, J. (1967): "Maximum Entropy Spectrum Analysis," Paper presented at 37 thAnnual Mtg. of Exploration Geophyslcrists, Oklahoma City.

Chor-Yiu, S. and H. White (1996): "Information criteria for selecting possibly misspecified parametric models," *Journal of Econometrics*, 71, 207–225.

Cressie, N. and T. Read (1984): "Multinomial goodness-of-fit tests," *Journal of the Royal Statistical Society. Series B (Methodological)*, 440–464.

Csiszár, I. (1967): "Information-type measures of the difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, 299–318.

——— (1975): "I-Divergence Geometry of Probability Distributions and Minimization Problems," *The Annals of Probability*, 3, 146–58.

——— (1995): "Generalized projections for non-negative functions," *Acta Mathematica Hungarica*, 68, 161–186.

Ekeland, I. and R. Témam (1987): *Convex Analysis and Variational Problems*, Society for Industrial and Applied Mathematics.

Giacomini, R. and G. Ragusa (2013): "Theory-coherent forecasting," *Journal of Econometrics (in press)*.

Gourieroux, C. and A. Monfort (1997): *Simulation-Based Econometric Methods*, Oxford University Press.

Gowda, M. and M. Teboulle (1990): "A Comparison of Constraint Qualifications in Infinite-Dimensional Convex Programming," *SIAM Journal on Control and Optimization*, 28, 925–935.

Hiriart-Urruty, J.-B. and C. Lemarechal (1993): *Convex Analysis and Minimization Algorithms I: Fundamentals (Grundlehren Der Mathematischen Wissenschaften)*, Springer.

Holmes, R. D. (1974): *Geometric Functional Analysis and Its Applications*, Springer-Verlag.

Kim, J. (2002): "Limited information likelihood and Bayesian analysis," *Journal of Econometrics*, 107, 175–193.

Kitamura, Y. (2001): "Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions," *Econometrica*, 69, 1661–1672.

Kitamura, Y., T. Otsu, and K. Evdokimov (2009): "Robustness, Infinitesimal Neighborhoods, and Moment Restrictions," *Cowles Foundation Discussion Papers*.

Kitamura, Y. and M. Stutzer (1997): "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861–874.

Kitamura, Y., G. Tripathi, and H. Ahn (2004): "Empirical Likelihood-Based Inference in Conditional Moment Restriction Models," *Econometrica*, 72, 1667–1714.

Komunjer, I. and Q. Vuong (2009): "Semiparametric Efficiency Bound in Time-Series Models for Conditional Quantiles," *Econometric Theory*, forthcoming, forthcoming.

Krasnosel'skii, M. and Y. Rutickii (1961): *Convex Functions and Orlicz Spaces*, P.Noordhoff Ltd.

Kullback, S. and M. A. Khairat (1966): "A Note on Minimum Discrimination Information," *The Annals of Mathematical Statistics*, 37, 279–280.

Liese, F. (1975): "On the existence of $f-$projections," *Colloq. Math. Soc. J. Bolyai*, 16, 431–446, budapest.

Liese, F. and I. Vajda (1987): *Convex Statistical distances*, Leipzig: Teubner.

Newey, W. and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–256.

Otsu, T., M. H. Seo, and Y.-J. Whang (2008): "Testing for Non-Nested Conditional Moment Restrictions using Unconditional Empirical Likelihood," Cowles Foundation Discussion Paper No. 1660.

Ragusa, G. (2011): "Minimum divergence, generalized empirical likelihoods, and higher order expansions," *Econometric Reviews*, 30, 406–456.

Rockafellar, R. T. (1968): "Integrals which are convex functionals," *Pacific Journal of Mathematics*, 24, 525–539.

——— (1970): *Convex Analysis*, Princeton, New Jersey: Princeton University Press.

——— (1971): "Integrals which are convex functionals. II." *Pacific Journal of Mathematics*, 39, 439–469.

——— (1974): "Conjugate Duality and Optimization, volume 16 of Regional Conferences Series in Applied Mathematics," *SIAM, Philadelphia*.

Sawa, T. (1978): "Information Criteria for Discriminating Among Alternative Regression Models," *Econometrica*, 46, 1273–1291.

Schennach, S. M. (2014): "Entropic Latent Variable Integration via Simulation," *Econometrica*, 82, 345–385.

Shi, X. (2014): "A Non-Degenerate Vuong Test," Manuscript.

Stein, C. (1956): "Efficient Nonparametric Testing and Estimation," in *Proceedings of the Third Berkeley Symposium in Mathematical Statistics and Probability*, Berkeley: University of California Press, vol. 1, 187–196.

Stinchcombe, M. B. and H. White (1998): "Consistent Specification Testing with Nuisance Parameters Present Only under the Alternative," *Econometric Theory*, 14, pp. 295–325.

Tang, Y. and S. Ghosal (2007): "Posterior Consistency of Dirichlet Mixtures for Estimating a Transition Density," *Journal of Statistical Planning and Inference*, 137, 1711–1726.

Ullah, A. (1996): "Entropy, divergence and distance measures with econometric applications," *Journal of Statistical Planning and Inference*, 49, 137–162.

Vuong, Q. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307–333.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

——— (1994): *Estimation, Inference and Specification Analysis*, Cambridge University Press.

Zellner, A. (1996): "Models, prior information, and Bayesian analysis," *Journal of Econometrics*, 75, 51–68.

——— (2002): "Information processing and Bayesian analysis," *Journal of Econometrics*, 107, 41–50.

——— (2003): "Some Recent Developments in Econometric Inference," *Econometric Reviews*, 22, 203–215.

Zellner, A. and J. Tobias (2001): "Further Results on Bayesian Method of Moments Analysis of the Multiple Regression Model," *International Economic Review*, 42, 121–139.

Zălinescu, C. (1999): "A comparison of constraint qualifications in infinite-dimensional convex programming revisited," *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, 40, 353–378.