



MÉMOIRE D'HABILITATION À DIRIGER DES RECHERCHES

LABORATOIRE DE MATHÉMATIQUES DE REIMS

Spécialité : Mathématiques

Option : Statistique

Contribution à l'Inférence Statistique et ses Applications

par

Amor KEZIOU

Soutenue le 21 novembre 2016, devant le jury composé de

Jan BEIRLANT (**Rapporteur**, Katholieke Universiteit Leuven)

Patrice BERTAIL (**Rapporteur**, Université Paris Ouest)

Gérard BIAU (Université Pierre et Marie Curie)

Michel BRONIATOWSKI (Université Pierre et Marie Curie)

Raymond BRUMMELHUIS (Université de Reims Champagne-Ardenne)

Laurence CARASSUS (Université de Reims Champagne-Ardenne)

Konstantinos ZOGRAFOS (**Rapporteur**, University of IOANNINA)

Remerciements

En tout premier lieu, je souhaite remercier très chaleureusement Jan Beirlant, Patrice Bertail et Konstantinos Zografos d'avoir accepté de rapporter mon mémoire d'HDR, et qui m'ont fait le grand honneur de participer au jury.

Je souhaite vivement remercier Gérard Biau, Michel Broniatowski, Raymond Brummelhuis et Laurence Carassus pour l'intérêt qu'ils ont accordé à ce travail et pour avoir accepté de participer au jury.

Les travaux présentés dans ce mémoire ont été obtenus, en grande partie, en collaboration avec de nombreuses personnes que je souhaite vivement remercier. Je souhaite tout d'abord remercier Michel Broniatowski pour m'avoir initié à la recherche pendant ma thèse et pour les travaux que nous avons réalisés. Qu'il trouve ici l'expression de ma reconnaissance. Je remercie tous mes autres co-auteurs Salim Bouzebda, Michel Couderchet, Georges Delaunay, Achouak El Arfaoui, Issam Elhattab, Mohammed El Rhabi, Hassan Fenniri, Abdelghani Ghazdali, Emmanuel Guillon, Abdelilah Hakim, Samuela Leoni-Aubin, Tewfik Lounis, Eric Moreau, Mohamed Salem Ould Mohamed, Michaël Paris, Philippe Regnault, Stéphanie Sayen et Tarek Zari.

Je souhaite aussi adresser mes remerciements à tous les collègues du LMR et du département MMI de l'URCA, particulièrement, à Laurent Di Menza et à Michael Pevzner pour leurs soutien et encouragements. Je remercie également Christine Ardilly, Sergine Bristiel, Christelle Marion et tous les autres personnels administratifs.

Mes ultimes remerciements vont à ma femme pour son soutien constant, à mon fils Eyad, à toute ma famille et à tous mes proches.

Publication list

(A) Papers in reviewed journals

- (1) Semiparametric estimation of mutual information and related criteria : Optimal test of independence. *To appear in IEEE Trans. Info. Theory*, (with P. Regnault);
- (2) Blind noisy mixture separation for independent/dependent sources through a regularized criterion on copulas. *To appear in Signal Processing*, (with A. Ghazdali, M. ElRhabi, H. Fenniri, A. Hakim);
- (3) New blind source separation method of independent/dependent sources. *Signal Processing*, no. 104, 319-324, (2014), (with H. Fenniri, A. Ghazdali and E. Moreau);
- (4) New entropy estimator with an application to test of normality. *Comm. Statist. Theory Methods*, 42 (2013), no. 12, 2245-2270, (with S. Bouzebda, I. Elhattab and T. Lounis);
- (5) A robust algorithm for convolutive blind source separation in presence of noise. *Signal Processing*, no. 93(4), 818-827, (2013), (with M. El Rhabi, H. Fenniri and E. Moreau);
- (6) A semiparametric maximum likelihood ratio test for the change point in copula models. *Stat. Methodol.* 14 (2013), 39-61, (with S. Bouzebda);
- (7) Divergences and duality for estimation and test under moment condition models. *J. Statist. Plann. Inference*, 142 (2012), no. 9, 2554-2573, (with M. Broniatowski);
- (8) Is organic matter alone sufficient to predict isoproturon sorption in calcareous soils? *Science of The Total Environment*, 432:251-6, (2012), (with A. El Arfaoui, S. Sayen, M. Paris, M. Couderchet and E. Guillon);
- (9) K-sample problem using strong approximations of empirical copula processes. *Math. Methods Statist.* 20 (2011), no. 1, 14-29, (with S. Bouzebda et T. Zari);
- (10) New estimates and tests of independence in semiparametric copula models. *Kybernetika (Prague)*, 46 (2010), no. 1, 178-201, (with S. Bouzebda);
- (11) A new test procedure of independence in copula models via χ^2 -divergence. *Comm. Statist. Theory Methods*, 39 (2010), no. 1-2, 1-20, (with S. Bouzebda);
- (12) New blind separation criterion for second order cyclostationary sources, *ARIMA Rev. Afr. Rech. Inform. Math. Appl.* 12 (2010), 1-14, (with M. S. Ould Mohamed, H. Fenniri and G. Delaunay);
- (13) Parametric estimation and tests through divergences and the duality technique. *J. Multivariate Anal.* 100 (2009), no. 1, 16-36, (with M. Broniatowski);
- (14) A test of independence in some copula models. *Math. Methods Statist.* 17, no. 2, 123-137, (2008), (with S. Bouzebda);
- (15) Empirical likelihood for semiparametric two-sample density ratio models. *Journal of Statistical Planning and Inference*, 138 (4) : 915-928, (2008), (with S. Leoni-Aubin);
- (16) Minimization of φ -divergences on sets of signed measures. *STUDIA Sci. Math. Hungarica*, 43 (4), 403-442, (2006), (with M. Broniatowski);

(B) Papers published in CRAS

- (1) A semiparametric test of independence in copula models for censored data. *C. R. Math. Acad. Sci. Paris*, 348 (2010), no. 7-8, 449-453, (with S. Bouzebda);
- (2) Estimation and tests of independence in copula models via divergences. *C. R. Math. Acad. Sci. Paris*, 347 (2009), no. 11-12, 667-672, (with S. Bouzebda);

- (3) Test of homogeneity in semiparametric two-sample density ratio models. *C. R. Math. Acad. Sci. Paris*, 340 (12) : 905-910, (2005), (with S. Leoni-Aubin);
- (4) Dual representation of φ -divergences and applications. *C. R. Math. Acad. Sci. Paris*, 336 (10) : 857-862, (2003);
- (5) Sur l'estimation de l'entropie des lois à support dénombrable. *C. R. Math. Acad. Sci. Paris*, 335 (9): 763-766, (2002);

(C) **Book chapters**

- (1) Multivariate Divergences with Application in Multisample Density Ratio Models. *Geometric Science of Information*, 444-453, (2015), Springer;
- (2) Generalized Mutual-Information Based Independence Tests. *Geometric Science of Information*, 454-463, (2015), Springer, (with P. Regnault);

Contents

Introduction générale	3
Part 1 : Divergences between measures	13
1 Divergences between measures	15
1.1 Introduction and notation	15
1.2 Existence of φ -projections on sets of signed measures	22
1.3 Characterization of φ -projections on sets of signed measures	24
1.3.1 on general sets	24
1.3.2 on sets defined by linear constraints	25
1.4 Fenchel duality for φ -divergences	26
1.5 Applications to the minimization of divergences on sets of signed finite measures satisfying linear constraints	28
2 Multivariate divergences	31
2.1 Introduction, notation and motivation	31
2.2 Fenchel duality of convex functions on finite dimensional vector space	33
2.3 Fenchel duality of convex functionals on possibly infinite dimensional vector spaces	34
2.4 Dual representations of the multivariate ψ -divergences	35
Part 2 : New methods in statistical inference	37
3 Parametric Estimation and Tests	39
3.1 Introduction and notation	39
3.2 Parametric estimation and tests	41
3.2.1 The asymptotic properties of $\widehat{\alpha}_\varphi(\theta)$ and $\widehat{D}_\varphi(\theta, \theta_T)$ for a given θ in Θ	43
3.2.2 The asymptotic behavior of the MD φ DE's	46
3.2.3 Composite tests by minimum φ -divergence	47
3.3 Non regular models. A simple solution for mixture models	49
3.3.1 A simple solution to the problem of testing the number of components in a mixture	51
3.3.2 Confidence regions for the mixture parameters	51
3.4 Concluding remarks and possible developments	52
4 Estimation and test under moment condition models	53
4.1 Introduction and notation	53
4.2 Statistical divergences	55
4.3 Minimum divergence estimates	55
4.4 Dual representation of φ -divergences under constraints	57
4.5 Asymptotic properties of the estimates of the parameter and the divergences	60
4.5.1 Asymptotic properties under the model	60
4.5.2 Asymptotic properties of the estimates of the divergences for a given value of the parameter	61
4.5.3 Asymptotic properties under misspecification	63

4.6	Concluding remarks and possible developments	64
5	Statistical inference for semiparametric two-sample density ratio models	65
5.1	Introduction and notation	65
5.1.1	Comparison of two populations	65
5.1.2	Logistic model and multiplicative-intercept risk model	66
5.1.3	The profile EL and its irregularity under the null hypothesis $\mathcal{H}_0 : Q = P$	68
5.2	Adjustment of the profile empirical likelihood	69
5.3	Asymptotic behavior of the estimate and test statistic under the null and the alternative hypotheses, and approximation of the power function	71
5.4	Discussion and directions for future research	72
6	Statistical inference for semiparametric multi-sample density ratio models	75
6.1	Introduction, notation and motivation	75
6.2	Inference with multiple samples in MDRM through conjugate representation of ψ -divergences	76
6.3	EL and ψ -divergences under the exponential MDRM	78
6.4	Asymptotic behavior of the estimates and test statistics	80
6.5	Discussion and directions for future research	82
7	Semiparametric estimation of mutual information and related criteria	83
7.1	Introduction and notations	83
7.2	φ -mutual informations, dual representations and estimation strategy	85
7.2.1	Introducing φ -mutual informations	85
7.2.2	Semiparametric modeling of the ratio $d\mathbb{P}/d\mathbb{P}^\otimes$	86
7.2.3	Dual representation and dual estimation of φ -MI	88
7.2.4	A model selection procedure for the ratio $d\mathbb{P}/d\mathbb{P}^\otimes$ through φ -MI criterion	91
7.3	Asymptotic properties of the estimates	91
7.3.1	Consistency	91
7.3.2	The limiting distribution of the estimate $\hat{I}_{\varphi_1}$ of KL-MI	92
7.3.3	Bootstrap calibration	93
7.4	Large deviations principle and Bahadur asymptotic efficiency	93
7.5	Concluding remarks and directions for future research	95
8	A semiparametric likelihood-type test of independence in copula models	97
8.1	Introduction and motivations	97
8.2	The maximum pseudo-likelihood procedure	99
8.2.1	Estimation	99
8.2.2	Tests of Independence	102
8.3	Concluding remarks and possible developments	104
9	A test procedure of independence in copula models via χ^2-divergence	105
9.1	Introduction and motivations	105
9.2	A semiparametric estimation procedure through χ^2 -divergence	108
9.3	A test based on “ χ^2 -divergence”	110
9.4	Conclusion and possible developments	112
10	A semiparametric likelihood ratio test to the change point in copula models	113
10.1	Introduction and motivations	113
10.2	Main results	115
10.2.1	Gaussian approximation	117
10.3	The asymptotic distribution of the SLR statistic under the alternative hypothesis	118
10.3.1	Approximation of the power function of the test 2	120
10.4	Concluding remarks and directions for future research	120

11 Other contribution : ICA and BSS using divergences and copulas	121
11.1 Introduction	122
11.2 The BSS model	123
11.2.1 BSS model in the noise-free case	123
11.2.2 BSS model for noisy mixtures	123
11.3 Brief recall on copula	124
11.4 A combined TV-Copula approach for BSS	125
11.4.1 Setp 1 : Denoising the observed signals	125
11.4.2 Step 2 : BSS via minimizing TV-regularized KL-divergence between copulas	125
11.5 Discretization and Statistical estimation	127
11.5.1 Denoising the discrete observed signal	128
11.5.2 BSS by minimizing TV-regularized KL-divergence between copulas	129
11.6 Conclusion and directions for future research	130
Bibliographie	131

Introduction générale

Ce mémoire présente une synthèse de mes travaux de recherche. Il est composé de deux parties. Une première partie sur l'étude des divergences entre mesures et l'obtention des représentations duales de celles-ci. La deuxième est consacrée à la présentation de notre contribution en statistique inférentielle grâce notamment à l'utilisation des divergences et de leurs représentations duales. Je décris brièvement, dans ce qui suit, le contenu de chacune des deux parties. Des perspectives de recherche sont présentées à la fin des chapitres.

Partie 1 : Divergences entre mesures

Cette première partie est composée de deux chapitres. Le Chapitre 1 présente les résultats des articles [Keziou \(2003a\)](#) et [Broniatowski and Keziou \(2006\)](#) sur les φ -divergences entre deux mesures (une mesure signée finie et une loi de probabilité). Soit $\varphi : \mathbb{R} \mapsto [0, +\infty]$ une fonction convexe fermée vérifiant $\varphi(1) = 0$ telle que son domaine $\text{dom}_\varphi := \{x \in \mathbb{R} \text{ t.q. } \varphi(x) < \infty\} =: (a_\varphi, b_\varphi)$ soit un intervalle de $\mathbb{R}$, avec $a_\varphi < 1 < b_\varphi$. Soit $(\mathcal{X}, \mathcal{B})$ un espace mesurable. Pour toute loi de probabilité P et toute mesure signée finie (m.s.f.) Q sur $(\mathcal{X}, \mathcal{B})$, la φ -divergence, entre Q et P , est définie par

$$D_\varphi(Q, P) := \int_{\mathcal{X}} \varphi \left(\frac{dQ}{dP}(x) \right) dP(x),$$

si Q est absolument continue (a.c.) par rapport à P , et $D_\varphi(Q, P) = +\infty$, sinon. Les φ -divergences ont été introduites par [Csiszár \(1963, 1967b\)](#) et [Ali and Silvey \(1966\)](#), appelées souvent, en théorie de l'information, f -divergences. Les divergences bien connues de Kullback-Leibler (KL), Kullback-Leibler modifiée (KL_m), χ^2 , χ^2 modifiée (χ_m^2) et Hellinger (H) sont obtenues, respectivement, pour les fonctions $\varphi(x) = x \log x - x + 1$, $\varphi(x) = -\log x + x - 1$, $\varphi(x) = (x-1)^2/2$, $\varphi(x) = (x-1)^2/(2x)$ et $\varphi(x) = 2(\sqrt{x} - 1)^2$. Soient $\mathcal{F}$ une classe quelconque de fonctions mesurables $f : \mathcal{X} \rightarrow \mathbb{R}$, $\mathcal{F}_b$ la classe de toutes les fonctions mesurables bornées $f : \mathcal{X} \rightarrow \mathbb{R}$, $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ l'espace vectoriel (e.v.) engendré par la classe de fonctions $\mathcal{F} \cup \mathcal{F}_b$, et M l'e.v. des mesures signées finies sur $(\mathcal{X}, \mathcal{B})$. Considérons le sous-espace vectoriel de M , $\mathcal{M}_{\mathcal{F}} := \{Q \in M \text{ t.q. } \int_{\mathcal{X}} |f(x)| d|Q|(x) < \infty, \forall f \in \mathcal{F}\}$. On munit $\mathcal{M}_{\mathcal{F}}$ de la topologie $\tau_{\mathcal{F}}$ la moins fine qui rend continues toutes les fonctions $Q \in \mathcal{M}_{\mathcal{F}} \mapsto \int_{\mathcal{X}} f(x) dQ(x)$, $\forall f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$. Nous avons obtenu le résultat suivant.

Proposition 0.1. ([Keziou \(2003a\)](#) Proposition 2.1 ou [Keziou \(2003b\)](#) Chapitre 1). $\mathcal{M}_{\mathcal{F}}$ muni de la topologie $\tau_{\mathcal{F}}$ est un espace vectoriel topologique Hausdorff localement convexe. Son dual topologique est l'espace des applications $\{Q \in \mathcal{M}_{\mathcal{F}} \mapsto \int_{\mathcal{X}} f(x) dQ(x); f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle\}$. De plus, pour toute loi de probabilité P , les fonctions convexes $Q \in \mathcal{M}_{\mathcal{F}} \mapsto D_\varphi(Q, P) \in [0, +\infty]$ sont semi-continues inférieurement par rapport à la topologie $\tau_{\mathcal{F}}$.

D'après la proposition précédente, la transformée de Fenchel-Legendre (ou la conjuguée convexe) $T(\cdot, P)$ de $Q \in \mathcal{M}_{\mathcal{F}} \mapsto D_\varphi(Q, P)$ est définie donc, sur le dual $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$, comme suit

$$\forall f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle, \quad T(f, P) := \sup_{Q \in \mathcal{M}_{\mathcal{F}}} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - D_\varphi(Q, P) \right\}.$$

La proposition précédente montre également que toutes les conditions du théorème général de dualité de Fenchel (c.f. e.g. [Dembo and Zeitouni \(1998\)](#)) sont vérifiées pour les fonctions convexes $Q \mapsto D_\varphi(Q, P)$

sur l'e.v. $\mathcal{M}_{\mathcal{F}}$ muni de la topologie $\tau_{\mathcal{F}}$. Nous avons obtenu ainsi la représentation duale suivante pour les divergences.

Proposition 0.2. (*Keziou (2003a) Proposition 2.2*). Soit P une loi de probabilité. $\forall Q \in \mathcal{M}_{\mathcal{F}}$, on a

$$D_{\varphi}(Q, P) = \sup_{\langle \mathcal{F} \cup \mathcal{F}_b \rangle} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - T(f, P) \right\}. \quad (0.1)$$

La représentation duale précédente reste valable si l'espace vectoriel $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ est remplacé par une classe de fonctions vérifiant certaines conditions, pour laquelle une forme explicite de la conjuguée convexe $T(f, P)$ a été obtenue.

Theorem 0.1. (*Keziou (2003a) Théorème 2.1*). Supposons que φ soit strictement convexe et deux fois continûment dérivable sur l'intérieur de son domaine $]a_{\varphi}, b_{\varphi}[$. Soient P une loi de probabilité et Q une mesure signée finie telles que $D_{\varphi}(Q, P) < \infty$. Supposons que les trois conditions suivantes soient vérifiées : (i) $\int_{\mathcal{X}} |f(x)| d|Q|(x) < \infty, \forall f \in \mathcal{F}$; (ii) $\varphi'(dQ/dP) \in \mathcal{F}$; (iii) $\varphi'^{-1}(f(x))$ est bien définie pour tout $x \in \mathcal{X}$. Alors, la divergence $D_{\varphi}(Q, P)$ admet la représentation duale suivante

$$D_{\varphi}(Q, P) = \sup_{f \in \mathcal{F}} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - \int_{\mathcal{X}} f(x) \varphi'^{-1}(f(x)) - \varphi(\varphi'^{-1}(f(x))) dP(x) \right\}. \quad (0.2)$$

De plus, le supremum est unique P -p.s. et est atteint en $f = \varphi'(dQ/dP)$.

Nous avons ensuite utilisé la représentation (0.2) pour résoudre le problème de définition des estimateurs plug-in des divergences et proposé de nouveaux estimateurs des paramètres (dépendants de la φ -divergence utilisée), et des estimateurs des φ -divergences, convergents, efficaces et adaptés aux modèles considérés, dans un cadre paramétrique (discret ou non) (c.f. Keziou (2003b) Chapitre 1), semi-paramétrique à deux échantillons (Keziou (2003b) Chapitre 2) ou non-paramétrique restreint par des contraintes linéaires (Keziou (2003b) Chapitre 3). Dans l'article Broniatowski and Keziou (2006), nous avons considéré le problème de minimisation des divergences entre une loi de probabilité donnée P et des sous-ensembles Ω de l'e.v. $\mathcal{M}_{\mathcal{F}}$, des mesures signées finies qui intègrent tous les éléments d'une classe donnée $\mathcal{F}$ de fonctions, bornées ou non bornées. L'espace $\mathcal{M}_{\mathcal{F}}$ est muni de la topologie faible $\tau_{\mathcal{F}}$. Nous avons dans cet article traité les problèmes d'existence et de caractérisation des projections de P sur des ensembles de mesures Ω . Nous avons également considéré le problème de l'égalité duale lorsque l'ensemble Ω est défini par des contraintes linéaires. De nouvelles versions des représentations duales (0.1) et (0.2) des divergences ont été obtenues dans cet article sous des conditions plus faibles :

Theorem 0.2. (*Broniatowski and Keziou (2006) Theorem 4.3*). Supposons que φ soit dérivable. Alors, pour toute mesure $Q \in \mathcal{M}_{\mathcal{F}}$ tel que $D_{\varphi}(Q, P) < \infty$ et $\varphi'(dQ/dP) \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$, la divergence $D_{\varphi}(Q, P)$ admet la représentation duale

$$D_{\varphi}(Q, P) = \sup_{f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - \int_{\mathcal{X}} \varphi^*(f(x)) dP(x) \right\}, \quad (0.3)$$

où $\varphi^*(\cdot)$ est la conjuguée convexe de $\varphi(\cdot)$, i.e., la fonction définie par

$$\varphi^*(\cdot) : t \in \mathbb{R} \mapsto \varphi^*(t) := \sup_{x \in \mathbb{R}} \{tx - \varphi(x)\}.$$

Le supremum dans (0.3) est atteint en $\bar{f} = \varphi'(dQ/dP)$. De plus, si φ est essentiellement régulière, alors le supremum $\bar{f} = \varphi'(dQ/dP)$ est unique P -p.s..

Le théorème précédent reste valable si l'espace vectoriel $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ est remplacé par une classe de fonctions quelconque $\mathcal{F}$ vérifiant certaines conditions.

Theorem 0.3. (*Broniatowski and Keziou (2006) Theorem 4.4*). Supposons que φ soit dérivable. Soit $\mathcal{F}$ une classe de fonctions mesurables $f : \mathcal{X} \rightarrow \mathbb{R}$. Alors, pour toute mesure $Q \in \mathcal{M}_{\mathcal{F}}$ telle que $D_{\varphi}(Q, P) < \infty$ et $\varphi'(dQ/dP) \in \mathcal{F}$, la divergence $D_{\varphi}(Q, P)$ admet la représentation duale

$$D_{\varphi}(Q, P) = \sup_{f \in \mathcal{F}} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - \int_{\mathcal{X}} \varphi^*(f(x)) dP(x) \right\}. \quad (0.4)$$

Le supremum dans (0.4) est atteint en $\bar{f} = \varphi'(dQ/dP)$. De plus, si φ est essentiellement régulière, alors le supremum $\bar{f} = \varphi'(dQ/dP)$ est unique P -p.s..

Le Chapitre 2 présente les nouveaux critères des divergences multivariées (appelées ψ -divergences) que nous avons introduites récemment, c.f. [Keziou \(2015\)](#), et qui sont motivées par leur lien direct avec les modèles semi-paramétriques de rapport des densités à plusieurs échantillons. Soit P_0 une loi de probabilité sur $(\mathcal{X}, \mathcal{B})$. Soit

$$\psi(\cdot) : \mathbf{x} := (x_1, \dots, x_K)^{\top} \in \mathbb{R}^K \mapsto \psi(\mathbf{x}) \in [0, +\infty] \quad (0.5)$$

une fonction propre, fermée, convexe, vérifiant $\psi(\mathbf{1}_K) = 0$, où $\mathbf{1}_K := (1, \dots, 1)^{\top} \in \mathbb{R}^K$, $K \geq 1$, telle que $\mathbf{1}_K$ soit un point intérieur du domaine de $\psi(\cdot)$, $\text{dom}_{\psi} := \{\mathbf{x} \in \mathbb{R}^K \text{ t.q. } |\psi(\mathbf{x})| < \infty\}$. On définit la ψ -divergence (multivariée), entre K mesures signées finies et une loi de probabilité P_0 , comme suit

$$D_{\psi}(Q_1, \dots, Q_K; P_0) := \int_{\mathcal{X}} \psi \left(\frac{dQ_1}{dP_0}(x), \dots, \frac{dQ_K}{dP_0}(x) \right) dP_0(x), \quad (0.6)$$

si $Q_1, \dots, Q_K$ sont toutes a.c. par rapport à P_0 ; et $D_{\psi}(Q_1, \dots, Q_K; P_0) = +\infty$, sinon. Evidemment, pour $K = 1$, on retrouve les critères classiques des φ -divergences (entre deux mesures). Il est clair que toute ψ -divergence est positive, et que la fonctionnelle $(Q_1, \dots, Q_K) \in M^K \mapsto D_{\psi}(Q_1, \dots, Q_K; P_0) \in [0, +\infty]$ est convexe sur l'e.v. produit M^K , pour toute probabilité P_0 . De plus, si $\psi(\cdot)$ est strictement convexe sur un voisinage de $\mathbf{1}_K$, alors on a la propriété fondamentale suivante

$$D_{\psi}(Q_1, \dots, Q_K; P_0) \geq 0, \text{ avec } D_{\psi}(Q_1, \dots, Q_K; P_0) = 0 \text{ ssi } Q_1 = \dots = Q_K = P_0. \quad (0.7)$$

Une ancienne définition différente des divergences multivariées (entre plus de deux lois de probabilités) a été introduite par [Györfi and Nemetz \(1978\)](#), et qui nécessite une mesure dominante σ -finie λ de toutes les lois de probabilités. Elle a été utilisée par [Zografos \(1994\)](#), [Zografos \(1998\)](#) et [Morales et al. \(1998\)](#), pour traiter des problèmes d'estimation et de test à plusieurs échantillons, pour des lois discrètes. Cette définition n'est pas appropriée à notre contexte, notamment pour l'application du théorème de dualité et l'obtention des représentations duales pour ces critères. Nous avons considéré la dualité de Fenchel, et nous avons obtenu des représentations duales pour les ψ -divergences, vues comme des fonctionnelles convexes sur des sous-espaces vectoriels bien choisis de l'e.v. produit M^K . Nous avons ensuite utilisé ces représentations pour définir de nouveaux estimateurs et des tests dans les modèles semi-paramétriques des rapports de densités à plusieurs $(1 + K)$ échantillons, généralisant ainsi nos résultats sur la vraisemblance empirique duale obtenus dans le cas des modèles de rapport de densités à deux échantillons, c.f. [Keziou and Leoni-Aubin \(2008\)](#).

Partie 2 : Nouvelles méthodes d'inférence statistique

Cette partie du mémoire présente les estimateurs et tests obtenus, grace notamment à l'utilisation des représentations duales des φ -divergences et ψ -divergences multivariées, que nous avons établies, avec des choix adaptés de la classe des fonctions $\mathcal{F}$ selon le contexte. Elle est composée de neuf chapitres. Dans un cadre paramétrique, le Chapitre 3 présente une procédure d'estimation et de test, basée sur les représentations duales des φ -divergences. La classe d'estimateurs proposés, c.f. [Broniatowski and Keziou \(2009\)](#), contient l'estimateur du maximum de vraisemblance qui correspond au choix particulier de la divergence de Kullback-Leibler modifiée, i.e., lorsque $\varphi(x) = -\log x + x - 1$. Nous avons étudié les propriétés asymptotiques des estimateurs proposés, ainsi que les statistiques des tests simples ou composés considérés, à la fois sous les hypothèses nulles/modèles et sous les alternatives/misspecifications.

L'estimateur du maximum de vraisemblance et la statistique du rapport de vraisemblance ont été étudiés comme cas particuliers. Nous avons obtenu ainsi des approximations des fonctions de puissance pour des tests simples ou composés. Nous avons obtenu également une solution au problème d'irrégularité du test du rapport de vraisemblance pour tester le nombre de composantes dans un mélange. Le problème d'irrégularité du rapport de vraisemblance est dû au fait que, sous l'hypothèse nulle, la valeur du paramètre est un point frontière du domaine; un test basé, sur l'écriture variationnelle de la divergence de χ^2 entre mesures signées, a été proposé et étudié.

Le Chapitre 4 présente les résultats obtenus par [Broniatowski and Keziou \(2012\)](#). Nous avons considéré les problèmes d'estimation et de test dans des modèles semi-paramétriques définis par des contraintes linéaires, en utilisant les φ -divergences et les techniques de dualité de l'analyse convexe. Les modèles considérés sont définis par des équations de la forme $\mathbb{E}[g(X, \theta_T)] = \mathbf{0}$, où $\mathbb{E}[\cdot]$ désigne l'espérance mathématique, et $g := (g_1, \dots, g_\ell)^\top \in \mathbb{R}^\ell$ est une fonction vectorielle spécifiée, définie sur $\mathbb{R}^m \times \mathbb{R}^d$. On trouve, dans la littérature, de nombreux exemples de ces modèles; c.f. e.g. [Qin and Lawless \(1994\)](#), [Haberman \(1984\)](#), [Sheehy \(1987\)](#), [McCullagh and Nelder \(1983\)](#), [Owen \(2001\)](#) et les références ci-incluses. [Bertail \(2006\)](#) et [Hjort et al. \(2009\)](#) étendent ces modèles au cas où les contraintes peuvent dépendre d'un paramètre de nuisance, éventuellement de dimension infinie, et au cas où le nombre de contraintes ℓ peut être grand en comparaison avec la taille de l'échantillon, et donnent de nombreux exemples d'application. En notant P_X la loi de X , l'équation précédente peut s'écrire sous la forme $\int_{\mathbb{R}^m} g(x, \theta_T) dP_X(x) = \mathbf{0}$. Notons M l'ensemble des mesures signées finies sur $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$, et définissons, pour tout $\theta \in \Theta$, les ensembles de mesures $\mathcal{M}_\theta := \{Q \in M; \int_{\mathbb{R}^m} g(x, \theta) dQ(x) = \mathbf{0}\}$. Le modèle considéré s'écrit donc $\mathcal{M} := \bigcup_{\theta \in \Theta} \mathcal{M}_\theta = \bigcup_{\theta \in \Theta} \{Q \in M; \int_{\mathbb{R}^m} g(x, \theta) dQ(x) = \mathbf{0}\}$. Soit $X_1, \dots, X_n$ un échantillon i.i.d. de X . Nous avons considéré les problèmes du test de validité de ces modèles, l'estimation du paramètre θ_T , la construction de régions de confiance, et la construction de tests simples ou composés relatifs au paramètre. Notons que ces problèmes ont été considérés par de nombreux auteurs de la littérature; [Hansen \(1982\)](#) considère la méthode des moments généralisée (MMG), [Hansen et al. \(1996\)](#) introduisent un autre estimateur appelé "the continuous updating estimate" (CUE). L'approche de vraisemblance empirique (VE), développée par [Owen \(1988, 1990\)](#), a été étudiée, dans le contexte du modèle précédent, par [Qin and Lawless \(1994\)](#) et [Imbens \(1997\)](#), en introduisant l'estimateur de vraisemblance empirique. Une large littérature en économétrie a été également consacrée à l'étude de ces modèles; [Smith \(1997\)](#) et [Newey and Smith \(2004\)](#) ont fourni une classe d'estimateurs appelés estimateurs de vraisemblance empirique généralisée (VEG) qui contient les estimateurs de VE et CUE. [Schennach \(2007\)](#) discute la normalité asymptotique de l'estimateur de VE sous l'hypothèse de mauvaise spécification du modèle; l'auteur montre que l'estimateur VE cesse d'être $\sqrt{n}$ -convergent si les fonctions g_j sont non bornées et si P_X est de support non borné. [Newey and Smith \(2004\)](#) ont montré que l'estimateur de VE hérite des propriétés d'optimalité en terme d'efficacité du maximum de vraisemblance paramétrique (c.f. [Pfanzagl and Wefelmeyer \(1978\)](#)) lorsque il est corrigé du biais. [Corcoran \(1998\)](#) et [Baggerly \(1998\)](#) ont prouvé que, parmi une classe de statistiques de test par minimum de divergences empiriques, le rapport de VE est la seule statistique corrigible au sens de Bartlett. Le problème de construction des régions de confiance pour le paramètre θ_T a été traité dans l'article de [Owen \(1990\)](#). Les problèmes du test du modèle, d'estimation du paramètre et de la construction des tests relatifs à celui-ci, ont été traités, en utilisant l'approche de VE, dans les articles [Qin and Lawless \(1994\)](#), [Smith \(1997\)](#) et [Newey and Smith \(2004\)](#); ils obtiennent les lois limites des estimateurs et des statistiques des tests proposés, sous l'hypothèse nulle et sous le modèle. [Imbens \(1997\)](#) discute la loi asymptotique de l'estimateur de VE sous l'hypothèse de mauvaise spécification du modèle, et donne la variance limite de l'estimateur, sans préciser les conditions ni donner une preuve. [Bertail \(2006\)](#) utilise la dualité pour étudier la statistique du rapport de VE et sa Bertlett correctibilité pour des modèles semi-paramétriques plus généraux dépendants d'un paramètre de nuisance éventuellement de dimension infinie. La principale contribution de notre article [Broniatowski and Keziou \(2012\)](#) était la caractérisation précise, en utilisant nos résultats antérieurs [Broniatowski and Keziou \(2006\)](#), du problème dual d'une large classe d'estimateurs et de tests (incluant ceux de VE et de VEG) et son utilisation pour obtenir les lois limites sous les hypothèses alternatives des tests considérés et sous l'hypothèse de mauvaise spécification du modèle. Ces lois limites ont été utilisées pour obtenir

des approximations des fonctions de puissances des tests considérés. Tous les résultats ont été démontrés et systématiquement illustrés par simulations.

Le Chapitre 5 rapporte les articles [Keziou and Leoni-Aubin \(2005\)](#) et [Keziou and Leoni-Aubin \(2008\)](#); nous avons considéré les problèmes d'estimation et de test statistique à deux échantillons dans un cadre semi-paramétrique. On dispose de deux échantillons : $X_{0,1}, \dots, X_{0,n_0}$ et $X_{1,1}, \dots, X_{1,n_1}$, respectivement, de loi P et Q , sur $\mathbb{R}^m$, et on suppose que le rapport dQ/dP est de la forme du modèle d'[Anderson \(1979\)](#)

$$\frac{dQ}{dP}(x) = \exp \left\{ \alpha_T + \beta_T^\top \mathbf{b}(x) \right\} =: r(\theta_T, x), \quad x \in \mathbb{R}^m, \quad (0.8)$$

où $\theta_T^\top := (\alpha_T, \beta_T^\top)$ est la vraie valeur du paramètre, que l'on suppose appartenir à un domaine $\Theta \subset \mathbb{R}^{1+d}$, et $\mathbf{b}(\cdot)$ est une fonction de base spécifiée, définie sur $\mathbb{R}^m$ à valeurs dans $\mathbb{R}^d$. Ces modèles généralisent le modèle classique de regression logistique, et possèdent de nombreuses applications en apprentissage statistique. En effet, ces modèles peuvent être utilisés pour traiter les problèmes suivants : détection des outliers, tests à deux échantillons, test de rupture, estimation de la densité conditionnelle, classification binaire, estimation de l'information mutuelle, test d'indépendance, selection de variables, clustering, analyse en composantes indépendantes, etc.; c.f. e.g. [Sugiyama et al. \(2012a\)](#). Les lois Q et P sont inconnues, seul le rapport est spécifié de manière paramétrique. Nous avons considéré les deux problèmes suivants : estimation du paramètre d'intérêt θ_T et le test de l'hypothèse d'égalité $\mathcal{H}_0 : Q = P$ contre l'hypothèse alternative $\mathcal{H}_1 : Q \neq P$. Lorsque les deux échantillons sont indépendants, [Qin \(1998\)](#) propose une méthode utilisant la vraisemblance empirique pour l'estimation du paramètre θ_T . Nous avons montré par ailleurs que le rapport de vraisemblance empirique, correspondant au problème du test précédent, est mal défini sous l'hypothèse $\mathcal{H}_0$. Nous avons obtenu une version duale du rapport de vraisemblance empirique qui est bien définie sous les deux hypothèses, ainsi qu'un estimateur de vraisemblance empirique duale pour le paramètre θ_T . Nous avons également montré que la statistique du test du rapport de VE duale est égale, à une normalisation près ($2n$), à l'estimateur empirique obtenu via la représentation duale, d'une divergence particulière, de type Kullback-Leibler "pondérée". Il s'agit de la divergence associée à la fonction convexe réelle $\varphi_\rho(x) = \varphi(x) - \varphi(\rho_0 + \rho_1 x)$, avec $\varphi(x) = x \log x - x + 1$, où ρ_0 et ρ_1 sont, respectivement, les valeurs des limites (quand $n := n_0 + n_1 \rightarrow \infty$), de n_0/n et n_1/n . Cela a permis l'étude de l'estimateur et du test proposés à la fois sous l'hypothèse nulle et sous l'alternative ou l'hypothèse de mauvaise spécification du modèle, et la définition d'une classe plus générale d'estimateurs et tests dépendants du choix de la fonction φ_ρ . Les lois asymptotiques de l'estimateur du paramètre et de la statistique du test de VE duale ont été obtenues sous l'hypothèse nulle et sous l'hypothèse alternative. Une approximation de la fonction de puissance a été également obtenue.

Le Chapitre 6 décrit des résultats des articles [Keziou \(2015\)](#) et [Keziou \(2016\)](#). Nous avons introduit les divergences multivariées, entre K , $K \geq 1$, mesures signées finies $(Q_1, \dots, Q_K)$ et une loi de probabilité de référence P , toutes définies sur un même espace mesurable $(\mathcal{X}, \mathcal{B})$ quelconque. Ces divergences généralisent les divergences entre une mesure signée Q et une loi de probabilité P . Nous avons obtenu de nouvelles représentations duales pour ces divergences vues comme des fonctionnelles convexes sur des e.v. produits de mesures signées, en utilisant des arguments de dualité dans un contexte multivarié très technique, c.f. Chapitre 2. Nous avons ensuite utilisé les représentations obtenues pour définir une nouvelle classe d'estimateurs dans des modèles semi-paramétriques multiples de rapports de densités à partir de $1 + K$ échantillons, généralisant ainsi l'approche de vraisemblance empirique duale proposée dans le cas à deux échantillons. Nous avons montré la consistance de nos estimateurs, et obtenu les lois asymptotiques des estimateurs et statistiques de tests proposés sous l'hypothèse nulle, sous l'hypothèse alternative ou l'hypothèse de mauvaise spécification du modèle. Les estimateurs de vraisemblance empirique, c.f. e.g. [Qin and Zhang \(1997\)](#), [Qin \(1998\)](#) et [Keziou and Leoni-Aubin \(2008\)](#), ont été traités comme cas particulier, correspondant à une certaine divergence de type Kullback-Leibler pondéré, avec $K = 1$. L'introduction des divergences multivariées et l'obtention de leurs représentations duales ont permis l'étude des estimateurs et des statistiques de test sous les hypothèses nulles et sous les hypothèses alternatives. D'autre part, ces critères peuvent être utilisés pour définir de nouvelles procédures de selection de modèles dans le contexte des modèles semi-paramétriques de rapports de densités à plusieurs

échantillons, ainsi que de nouvelles méthodes de classification supervisée multi-groupe.

Dans le Chapitre 7, nous présentons nos travaux [Keziou and Regnault \(2015\)](#) et [Keziou and Regnault \(2016\)](#). Nous avons introduit de nouveaux tests d'indépendance dans un cadre semi-paramétrique par utilisation des représentations duales des φ -informations mutuelles (d'une loi de probabilité jointe), définies comme les φ -divergences entre la loi jointe et la loi produit de ces marges. Nous avons obtenu notamment un nouveau estimateur de l'information mutuelle classique possédant des propriétés optimales d'efficacité dans un contexte semi-paramétrique. Soit (X, Y) un vecteur aléatoire, à valeurs dans un espace produit mesurable $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}_X \otimes \mathcal{A}_Y)$, de loi jointe $\mathbb{P}$ et de lois marginales $\mathbb{P}_X$ et $\mathbb{P}_Y$. Soit $\varphi : \mathbb{R} \rightarrow [0, +\infty]$ une fonction convexe fermée telle que $\varphi(1) = 0$. On définit la φ -information mutuelle (φ -IM) de (X, Y) , ou de sa loi $\mathbb{P}$, comme la φ -divergence entre la loi jointe $\mathbb{P}$ et la loi produit $\mathbb{P}^\otimes := \mathbb{P}_X \otimes \mathbb{P}_Y$, des marges $\mathbb{P}_X$ et $\mathbb{P}_Y$; précisément

$$I_\varphi(\mathbb{P}) := D_\varphi(\mathbb{P}, \mathbb{P}^\otimes) = \int_{\mathcal{X} \times \mathcal{Y}} \varphi \left(\frac{d\mathbb{P}}{d\mathbb{P}^\otimes}(x, y) \right) d\mathbb{P}^\otimes(x, y).$$

Notons que $I_\varphi(\mathbb{P}) \geq 0$; de plus, si φ est strictement convexe sur un voisinage de 1, alors on a $I_\varphi(\mathbb{P}) = 0$ ssi $\mathbb{P} = \mathbb{P}^\otimes$, i.e., la φ -IM de (X, Y) est nulle ssi X et Y sont indépendantes. L'information mutuelle classique (IM) est obtenue pour la fonction particulière $\varphi(x) = x \log x - x + 1$, comme la divergence de Kullback-Leibler entre $\mathbb{P}$ et $\mathbb{P}^\otimes$. Notons M^1 l'espace des lois de probabilité sur $\mathcal{X} \times \mathcal{Y}$. Pour tout $P \in M^1$, on note par $P^\otimes := P_1 \otimes P_2$, la loi produit des marges P_1 et P_2 de P . Nous avons considéré le modèle semi-paramétrique suivant. On suppose que la loi $\mathbb{P}$ de (X, Y) appartient au modèle $\mathcal{M}_\Theta := \{P \in M^1 : \frac{dP}{dP^\otimes} = h_\theta(\cdot, \cdot), \theta \in \Theta\}$, où $\Theta \subseteq \mathbb{R}^{1+d}$ est le domaine du paramètre, et $h_\theta(\cdot, \cdot)$ est une fonction spécifiée, définie sur $\mathcal{X} \times \mathcal{Y}$ à valeurs dans $\mathbb{R}$, dépendante d'un paramètre θ de dimension $(1+d)$. On suppose que le modèle précédent vérifie les deux conditions suivantes: (i) $(h_\theta(\cdot, \cdot) = h_{\theta'}(\cdot, \cdot), \mathbb{P}^\otimes - p.s.) \Rightarrow (\theta = \theta')$ (identifiabilité); (ii) Il existe un unique $\theta_0 \in \text{int}(\Theta)$ vérifiant $h_{\theta_0}(x, y) = 1$, $\mathbb{P}^\otimes - p.s.$. Notons θ_T la vraie valeur du paramètre, i.e., la valeur unique vérifiant $\frac{d\mathbb{P}}{d\mathbb{P}^\otimes}(x, y) = h_{\theta_T}(x, y)$, $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$. Plusieurs exemples de ce modèle sont donnés dans [Keziou and Regnault \(2015\)](#) et [Keziou and Regnault \(2016\)](#). Sous les deux conditions: (C.1) la φ -IM est finie, $I_\varphi(\mathbb{P}) < \infty$; (C.2) $\forall \theta \in \Theta$, on a $\int_{\mathcal{X} \times \mathcal{Y}} |\varphi'(h_\theta(x, y))| d\mathbb{P}(x, y) < \infty$, une application directe de la représentation duale des divergences, c.f. [Keziou \(2003a\)](#) et [Broniatowski and Keziou \(2006\)](#), permet d'avoir l'écriture variationnelle suivante de la φ -IM de $\mathbb{P}$

$$I_\varphi(\mathbb{P}) = \sup_{\theta \in \Theta} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} \varphi'(h_\theta(x, y)) d\mathbb{P}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} \varphi^*(\varphi'(h_\theta(x, y))) d\mathbb{P}^\otimes(x, y) \right\},$$

où φ^* est la conjugué convexe de φ . De plus, le supremum est unique, et est atteint en $\theta = \theta_T$. La représentation précédente joue un rôle central dans la construction des estimateurs des φ -IM et des tests d'indépendance. Soient $(X_1, Y_1), \dots, (X_n, Y_n)$ un échantillon de (X, Y) de loi $\mathbb{P}$, et notons $\hat{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, Y_i)}$ la mesure empirique associée. Le but est de tester l'hypothèse d'indépendance $\mathcal{H}_0 : \mathbb{P} = \mathbb{P}^\otimes$ contre l'alternative $\mathcal{H}_1 : \mathbb{P} \neq \mathbb{P}^\otimes$, ce qui est équivalent, dans le contexte des φ -IM, à tester $I_\varphi(\mathbb{P}) = 0$ contre $I_\varphi(\mathbb{P}) > 0$. Au vu de la représentation duale précédente, nous avons proposé les estimateurs suivants, de la φ -IM et du paramètre θ_T ,

$$\begin{aligned} \hat{I}_\varphi &:= \sup_{\theta \in \Theta} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} \varphi'(h_\theta) d\hat{\mathbb{P}} - \int_{\mathcal{X} \times \mathcal{Y}} \varphi^*(\varphi'(h_\theta)) d\hat{\mathbb{P}}^\otimes \right\} \\ &= \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi'(h_\theta(X_i, Y_i)) - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \varphi^*(\varphi'(h_\theta(X_i, Y_j))) \right\} \\ \hat{\theta}_\varphi &:= \arg \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi'(h_\theta(X_i, Y_i)) - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \varphi^*(\varphi'(h_\theta(X_i, Y_j))) \right\}. \end{aligned}$$

Notons que l'estimateur plug-in direct de la φ -IM n'est pas bien défini généralement. D'autre part, si l'espace $\mathcal{X} \times \mathcal{Y}$ est discret fini, nous avons montré que l'estimateur dual précédent coïncide avec l'estimateur plug-in direct avec un choix spécifique du modèle $h_\theta(\cdot, \cdot)$. Nous avons montré la consistance

des estimateurs précédents, obtenu les lois limites des estimateurs $\hat{I}_\varphi$ sous l'hypothèse d'indépendance, et défini une classe de tests d'indépendance basés sur ces estimateurs. On a montré que cette classe recouvre le test d'indépendance classique du χ^2 pour des lois discrètes. Enfin, nous avons établi un principe de grandes déviations, et montré que le test utilisant le nouveau estimateur de l'IM est optimal au sens de Bahadur parmi tous les tests proposés.

Dans le Chapitre 8, nous avons proposé un test d'indépendance, dans des modèles de copules semi-paramétriques, de type rapport de vraisemblance; c.f. [Bouzebda and Keziou \(2008\)](#). Soit $(X_{1,1}, X_{2,1}), \dots, (X_{1,n}, X_{2,n})$ un échantillon d'un vecteur aléatoire bivarié (X_1, X_2) de loi P , de fonction de répartition $F(\cdot, \cdot)$ et de fonctions de répartitions marginales (que l'on suppose continues) $F_1(\cdot)$ et $F_2(\cdot)$. Notons $C(\cdot, \cdot)$ la copule associée à $F(\cdot, \cdot)$, i.e., la fonction définie, pour tout $(u_1, u_2) \in (0, 1)^2$, par $C(u_1, u_2) = P\{F_1(X_1) \leq u_1, F_2(X_2) \leq u_2\}$. On suppose que la copule C appartient à un modèle semi-paramétrique $\{C_\theta; \theta \in \Theta\}$, typiquement indexé par un paramètre $\theta \in \Theta \subset \mathbb{R}^d$; c.f. e.g. [Kimeldorf and Sampson \(1975b\)](#), [Kimeldorf and Sampson \(1975a\)](#) et [Joe \(1997\)](#). Notons θ_T la vraie valeur du paramètre, et θ_0 la valeur particulière du paramètre correspondant à l'hypothèse d'indépendance. Pour la majorité des modèles de la littérature, la valeur d'indépendance θ_0 est un point frontière du domaine Θ . L'approche non paramétrique d'estimation des copules a été initiée par [Deheuvels \(1979b\)](#) qui a introduit le processus empirique de la copule. Dans [Bouzebda and Keziou \(2008\)](#), nous avons considéré des modèles semi-paramétriques de copules avec des marginales inconnues générales. Pour estimer la vraie valeur inconnue θ_T , des procédures d'estimation semi-paramétriques basées sur la maximisation, sur l'espace du paramètre Θ , d'un critère bien choisi de pseudo-vraisemblance, a été étudiée par [Genest et al. \(1995\)](#), [Shih and Louis \(1995\)](#), [Wang and Ding \(2000\)](#) et [Tsukahara \(2005\)](#) parmi d'autres. Dans ces travaux, la normalité asymptotique a été établie pour $\sqrt{n}(\tilde{\theta}_n - \theta_T)$, où $\tilde{\theta}_n$ est un estimateur de θ_T , lorsque celui-ci est un point intérieur de Θ . Nous avons montré dans [Bouzebda and Keziou \(2008\)](#) que la normalité asymptotique de l'estimateur n'est plus valide sous l'hypothèse d'indépendance des marges, i.e., si $\theta_T = \theta_0$, θ_0 étant un point frontière du domaine Θ . D'autre part, la loi limite de la statistique du rapport des pseudo-vraisemblances, du test d'indépendance des marges, n'est plus une loi du χ^2 . Nous avons donc étudié ce problème lorsque θ est univarié, et nous avons obtenu les lois asymptotiques de l'estimateur de pseudo-vraisemblance de θ_T sous l'hypothèse d'indépendance des marges. Nous avons également proposé un test statistique de type rapport des pseudo-vraisemblances, pour lequel nous avons obtenu les lois asymptotiques à la fois sous l'hypothèse d'indépendance et sous l'alternative. Ces résultats ont été également établis pour des données censurées dans [Bouzebda and Keziou \(2010c\)](#).

Les résultats précédents, de l'article [Bouzebda and Keziou \(2008\)](#), ont été établis dans le cas où le paramètre est de dimension un. L'extension de ces résultats au cas multidimensionnel devient immédiatement un problème complexe. Nous avons donc proposé une alternative à l'approche de maximisation de la pseudo-vraisemblance dans [Bouzebda and Keziou \(2010b\)](#), Chapitre 9 de ce mémoire, en utilisant la représentation duale de la divergence du χ^2 , considérée sur un domaine étendue Θ_e du domaine Θ du paramètre. La valeur d'indépendance θ_0 devient ainsi un point intérieur du nouveau domaine Θ_e . Nous avons montré la normalité asymptotique de l'estimateur obtenu qui reste valable sous l'hypothèse d'indépendance, et nous avons proposé un test d'indépendance pour lequel nous avons montré que la loi asymptotique est une loi de χ^2 classique avec un nombre de degrés de liberté égale à la dimension du paramètre. La généralisation aux autres critères des φ -divergences a été réalisée dans nos travaux [Bouzebda and Keziou \(2009\)](#) et [Bouzebda and Keziou \(2010a\)](#).

Le Chapitre 10 présente le test, de non changement des valeurs des paramètres des modèles de copules semi-paramétriques; c.f. [Bouzebda and Keziou \(2013\)](#). Le test proposé est de type rapport de pseudo-vraisemblance de densités de copules, et nous avons montré que sa loi limite est la même que celle du test du rapport de vraisemblance paramétrique classique (c.f. e.g. [Csörgő and Horváth \(1997\)](#)). Sous certaines conditions, nous avons montré que la loi limite de la statistique du test, convenablement centrée et normalisée, est la loi double exponentielle. Nous avons également traité le problème de l'approximation gaussienne de la statistique du test. Enfin, la loi limite de la statistique du test, sous l'alternative, a été

établie également et une approximation de la fonction de puissance a été obtenue, qui fait apparaître une divergence s'écrivant comme mélange convexe de deux divergences de Kullback-Leibler.

Autre contribution : Analyse en composantes indépendantes et Séparation aveugle de sources via divergences et copules

Le dernier Chapitre décrit notre contribution en analyse en composantes indépendantes (ACI) et en séparation aveugle de sources (SAS) par utilisation des critères de divergences et les copules. Certains résultats ont été obtenus dans le cadre de la préparation de deux thèses de doctorat, de Ould Mohamed et de Ghazdali, que j'ai encadrées à 50% chacune; c.f. [Ould Mohamed \(2010\)](#) et [Ghazdali \(2015\)](#). L'ACI, et plus généralement la SAS, est l'un des thématiques les plus attractifs du traitement statistique du signal. Elle a fait l'objet de nombreux travaux de recherche depuis les années 80, à cause de leur richesse en termes d'applications dans des domaines très variés; c.f. e.g. [Choi and Cichocki \(1997\)](#) en reconnaissance de la parole, [De Young and Evans \(2007\)](#) en communication numérique sans fil, [Ferreira et al. \(2009\)](#) et [Congedo et al. \(2008\)](#) en traitement du signal biomédical. L'ACI est l'une des techniques utilisées pour répondre au problème de la SAS sous l'hypothèse d'indépendance des sources. L'ACI peut être considérée aussi comme une méthode qui complète l'ACP, pour transformer les variables, observées sur les individus, en des nouvelles variables indépendantes. Ensuite, ces nouvelles variables (indépendantes) peuvent être utilisées, pour construire des modèles de régression (ou des classifieurs) plus efficaces en terme de qualité de prédiction. L'objectif de la SAS est d'estimer les signaux sources à partir des signaux observés qui sont des mélanges inconnus de celles-ci. Les signaux observés $\mathbf{x}(n) := (x_1(n), \dots, x_d(n))^T \in \mathbb{R}^d$, $n = 1, \dots, N$, sont mélanges de signaux "sources" $\mathbf{s}(n) \in \mathbb{R}^m$, $n = 1, \dots, N$,

$$\mathbf{x}(n) := \mathbf{A}[\mathbf{s}(n)] + \mathbf{b}(n), n = 1, \dots, N, \quad (0.9)$$

où $\mathbf{A}[\cdot]$ est l'opérateur de mélange (inconnu); $\mathbf{s}(n) := (s_1(n), \dots, s_m(n))^T$, $n = 1, \dots, N$, sont les signaux sources (inconnues); et $\mathbf{b}(n) \in \mathbb{R}^d$, $n = 1, \dots, N$, est le bruit (inconnu). La présence du bruit rend difficile la reconstitution des sources. Pour simplifier le problème, considérons le cas non bruité des mélanges linéaires instantanés, où le nombre de signaux observés est égal au nombre de signaux sources, i.e., $d = m$. Le modèle (0.9) s'écrit dans ce cas sous la forme

$$\mathbf{x}(n) := \mathbf{A} \mathbf{s}(n), n = 1, \dots, N, \quad (0.10)$$

où $\mathbf{A}$ est la matrice de mélange, de dimension $m \times m$. Les solutions pour ce dernier sont donc de la forme

$$\mathbf{y}(n) := \mathbf{B} \mathbf{x}(n), n = 1, \dots, N, \quad (0.11)$$

où $\mathbf{B}$ est une matrice $m \times m$. Le problème est donc d'obtenir un "estimateur" $\hat{\mathbf{B}}$ (fonction seulement des signaux observés $\mathbf{x}(n)$, $n = 1, \dots, N$) qui soit le plus proche possible de la solution idéale $\mathbf{A}^{-1}$, ce qui conduit à l'estimation suivante des signaux sources $\mathbf{s}(n)$

$$\hat{\mathbf{s}}(n) := \hat{\mathbf{B}} \mathbf{x}(n), n = 1, \dots, N. \quad (0.12)$$

Il est à noter que le modèle de la SAS comporte deux indéterminations intrinsèques au problème : une première de permutation et une deuxième du facteur d'échelle, en effet, nous avons les deux égalités $\mathbf{x}(n) := \mathbf{A} \mathbf{s}(n) = \mathbf{A} \mathbf{P} \mathbf{P}^T \mathbf{s}(n)$, $n = 1, \dots, N$ et $\mathbf{x}(n) := \mathbf{A} \mathbf{s}(n) = \mathbf{A} \mathbf{\Lambda}^{-1} \mathbf{\Lambda} \mathbf{s}(n)$, $n = 1, \dots, N$, pour toute matrice de permutation $\mathbf{P}$ et pour toute matrice diagonale inversible $\mathbf{\Lambda}$. Donc la solution idéale $\mathbf{A}^{-1}$ ne peut être estimée qu'à une permutation $\mathbf{P}$ et un facteur d'échelle $\mathbf{\Lambda}$ près. Nous avons considéré ce problème en adoptant un point de vue utilisant la modélisation aléatoire suivante. Les signaux sources $\mathbf{s}(n) \in \mathbb{R}^m$, $n = 1, \dots, N$, sont considérés comme une suite de N réalisations indépendantes d'un vecteur aléatoire $\mathbf{S} \in \mathbb{R}^m$. Par conséquent, les signaux observés $\mathbf{x}(n) := \mathbf{A} \mathbf{s}(n)$, $n = 1, \dots, N$, peuvent être vus comme des réalisations i.i.d. du vecteur aléatoire $\mathbf{X} := \mathbf{A} \mathbf{S} \in \mathbb{R}^m$, et les solutions $\mathbf{y}(n) := \mathbf{B} \mathbf{x}(n)$, $n = 1, \dots, N$, peuvent être à leur tour considérées également comme des réalisations i.i.d. du vecteur aléatoire

$\mathbf{Y} := \mathbf{B}\mathbf{X} = \mathbf{B}\mathbf{A}\mathbf{S} =: \mathbf{M}\mathbf{S} \in \mathbb{R}^m$. Le principe de séparation aveugle de sources repose sur le résultat suivant. On montre que si les trois conditions suivantes sont vérifiées : (i) les composantes $S_1, \dots, S_m$ de $\mathbf{S}$ sont indépendantes; (ii) au plus une composante S_i est gaussienne et (iii) $\forall i = 1, \dots, m$, la loi de la composante S_i n'est pas une mesure de Dirac, alors le problème est identifiable (à une permutation et un facteur d'échelle près), et on a $Y_1, \dots, Y_m$ sont indépendantes ssi $\mathbf{M} = \mathbf{\Lambda}\mathbf{P}$, où $\mathbf{\Lambda}$ est une matrice diagonale (d'échelle), et $\mathbf{P}$ une matrice de permutation. Autrement dit, pour retrouver le vecteur source $\mathbf{S}$ il faut et il suffit de trouver la matrice $\mathbf{B}$ qui rend indépendantes les composantes de $\mathbf{Y}$. La démonstration de ce résultat, c.f. e.g. Comon (1994), repose sur les théorèmes de Darmois-Skitovich et Marcinkiewicz-Dugué; c.f. Darmois (1953) et Dugué (1951). Les conditions (ii) et (iii) précédentes sont faibles et sont généralement vérifiées dans la pratique. Par contre la condition (i) d'indépendance des sources est une condition forte. La plupart des méthodes de la SAS de la littérature utilisent néanmoins cette hypothèse et reposent sur le principe précédent. On trouve dans la littérature des procédures (dites de décorrélation) utilisant les statistiques d'ordre deux ou d'ordre supérieur (c.f. e.g. Pesquet and Moreau (2001), Castella et al. (2007)), d'autres consistent à optimiser (sur l'espace des matrices séparantes $\mathbf{B}$) un estimateur d'une mesure de dépendance des composantes du vecteur $\mathbf{Y}$. Comme mesures de dépendance qui ont été utilisées, on trouve le critère de l'information mutuelle (c.f. e.g. Comon (1994), Pham (2002)), et les critères de α -divergences, β -divergences et divergences de Renyi (c.f. e.g. Cichocki and Amari (2010) et Vrins et al. (2007) parmi d'autres). Dans nos articles Keziou et al. (2008) et Ould Mohamed et al. (2010), nous avons proposé un nouveau critère de SAS pour des sources cyclo-stationnaires d'ordre deux avec des fréquences cycliques connues et différentes. La convergence de l'algorithme proposé a été démontrée, et les résultats de simulations montrent la performance du critère proposé en comparaison avec ceux bien connus de Sobi, CyscloSobi et Jade. Nous avons proposé dans l'article El Rhabi et al. (2013) une nouvelle méthode de SAS pour des mélanges linéaires (instantanés ou convolutifs) de sources en présence du bruit. Notre approche procède en deux étapes. Une première étape de débruitage des signaux observés, suivie d'une étape de séparation par minimisation des divergences (entre la loi jointe et le produit des marges) pénalisée par la variation totale des signaux. L'efficacité et la robustesse de cette procédure ont été illustrées pour des signaux simulés et pour des signaux images réelles. Comme la plupart des méthodes de SAS, les méthodes précédentes supposent l'indépendance des signaux sources. Nous avons par ailleurs montré que la séparation des sources dépendantes est possible et proposé récemment, dans l'article Keziou et al. (2014), une nouvelle méthode prometteuse qui permet de séparer des mélanges linéaires instantanés de sources indépendantes/dépendantes, via la minimisation de la divergence de Kullback-Leibler entre densités de copules. En se basant sur les travaux El Rhabi et al. (2013) et Keziou et al. (2014), nous avons proposé une méthode pour séparer des mélanges linéaires instantanés, en présence du bruit, de sources indépendantes/dépendantes. Elle consiste à minimiser un estimateur de la divergence de Kullback-Leibler entre copules, pénalisée par un terme de type variation totale; c.f. Ghazdali et al. (2017). Dans ce dernier Chapitre 11, nous présentons les résultats de notre article Ghazdali et al. (2017), qui traite un cas plus général et utilise nos travaux antérieurs El Rhabi et al. (2013) et Keziou et al. (2014).

Part 1 : Divergences between measures

Chapter 1

Divergences between measures

In this Chapter, we present the results obtained in [Keziou \(2003a\)](#) and [Broniatowski and Keziou \(2006\)](#). We consider the minimization problem of φ -divergences between a given probability measure P and subsets Ω of the vector space $\mathcal{M}_{\mathcal{F}}$ of all signed finite measures which integrate a given class $\mathcal{F}$ of bounded or unbounded measurable functions. The vector space $\mathcal{M}_{\mathcal{F}}$ is endowed with the weak topology induced by the class $\mathcal{F} \cup \mathcal{F}_b$ where $\mathcal{F}_b$ is the class of all bounded measurable functions. We treat the problems of existence and characterization of the φ -projections of P on Ω . We consider also the dual equality and the dual attainment problems when Ω is defined by linear constraints. Using duality theory of convex analysis in infinite dimension, we obtain new dual representations for φ -divergences, viewed as convex functionals on well-chosen vector spaces of signed finite measures.

1.1 Introduction and notation

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and P be a given reference probability measure (p.m.) on $(\mathcal{X}, \mathcal{B})$. Denote M the real vector space of all signed finite measures on $(\mathcal{X}, \mathcal{B})$ and $M(P)$ the vector subspace of all signed finite measures absolutely continuous (a.c) with respect to (w.r.t.) P . Denote also M^1 the set of all p.m.'s on $(\mathcal{X}, \mathcal{B})$ and $M^1(P)$ the subset of all p.m.'s a.c w.r.t. P . Let φ be a proper¹ closed² convex function from $\mathbb{R}$ to $[0, +\infty]$ with $\varphi(1) = 0$ and such that its domain $\text{dom}_{\varphi} := \{x \in \mathbb{R} \text{ such that } \varphi(x) < \infty\}$ is an interval with endpoints $a_{\varphi} < 1 < b_{\varphi}$ (which may be finite or infinite). For any signed finite measure Q in $M(P)$, the φ -divergence between Q and P , is defined by

$$D_{\varphi}(Q, P) := \int_{\mathcal{X}} \varphi \left(\frac{dQ}{dP}(x) \right) dP(x). \quad (1.1)$$

When Q is not a.c. w.r.t. P , we set $D_{\varphi}(Q, P) = +\infty$. The φ -divergences between p.m.'s were introduced by [Csiszár \(1963, 1967b\)](#) as f -divergences. See also [Ali and Silvey \(1966\)](#). The definition of φ -divergences of [Csiszár \(1963, 1967b\)](#) between p.m.'s requires a common dominating σ -finite measure, denote it λ , for Q and P . Note that the two definitions of φ -divergences coincide on the set of all p.m.'s a.c w.r.t. P and dominated by λ . The φ -divergences between any signed finite measure Q and a p.m. P were introduced by [Csiszár et al. \(1999\)](#); they gave the following definition

$$D_{\varphi}(Q, P) := \int_{\mathcal{X}} \varphi(q(x)) dP(x) + b\sigma_Q^+(\mathcal{X}) - a\sigma_Q^-(\mathcal{X}), \quad (1.2)$$

where $a := \lim_{x \rightarrow -\infty} \varphi(x)/x$, $b := \lim_{x \rightarrow +\infty} \varphi(x)/x$, and $Q = qP + \sigma_Q$ and $\sigma_Q = \sigma_Q^+ - \sigma_Q^-$ are, respectively, the Lebesgue decomposition of Q and the Jordan decomposition of the singular part σ_Q . The definitions (1.1) and (1.2) coincide when Q is a.c. w.r.t. P or when $a = -\infty$ or $b = +\infty$. Since we will consider

¹A function φ is called proper if its domain is non void and $\varphi(x) > -\infty$, for all x .

²The closedness of φ (on $\mathbb{R}$) means that if a_{φ} (or b_{φ}) are finite numbers then $\varphi(x)$ tends to $\varphi(a_{\varphi})$ when $x \downarrow a_{\varphi}$ (or $\varphi(x)$ tends to $\varphi(b_{\varphi})$ when $x \uparrow b_{\varphi}$).

optimization of $Q \mapsto D_\varphi(Q, P)$ on sets of signed finite measures all a.c. w.r.t. P , it is more adequate for our sake to use the definition (1.1). For all p.m. P , the mappings $Q \in M \mapsto D_\varphi(Q, P)$ are convex and take nonnegative values. When $Q = P$ then $D_\varphi(Q, P) = 0$. Furthermore, if the function $x \mapsto \varphi(x)$ is strictly convex on a neighborhood of $x = 1$, then the following fundamental property holds

$$D_\varphi(Q, P) = 0 \text{ if and only if } Q = P. \quad (1.3)$$

All these properties are presented in [Csiszár \(1963, 1967c,b,a\)](#) and [Liese and Vajda \(1987\)](#) Chapter 1, for φ -divergences defined on the set of all p.m.'s M^1 . When the φ -divergences are defined on M , then the same properties hold. When defined on M^1 , the Kullback-Leibler (KL), modified Kullback-Leibler (KL_m), χ^2 , modified χ^2 (χ_m^2), Hellinger (H), and L_1 divergences are, respectively, obtained for the convex functions $\varphi(x) = x \log x - x + 1$, $\varphi(x) = -\log x + x - 1$, $\varphi(x) = \frac{1}{2}(x - 1)^2$, $\varphi(x) = \frac{1}{2}(x - 1)^2/x$, $\varphi(x) = 2(\sqrt{x} - 1)^2$ and $\varphi(x) = |x - 1|$. All those divergences except the L_1 one, belong to the class of the so-called power divergences introduced in [Cressie and Read \(1984\)](#) (see also [Liese and Vajda \(1987\)](#) Chapter 2). They are defined through the class of convex functions

$$x \in]0, +\infty[\mapsto \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)} \quad (1.4)$$

if $\gamma \in \mathbb{R} \setminus \{0, 1\}$, $\varphi_0(x) := -\log x + x - 1$ and $\varphi_1(x) := x \log x - x + 1$. (For all $\gamma \in \mathbb{R}$, we define $\varphi_\gamma(0) := \lim_{x \downarrow 0} \varphi_\gamma(x)$). So, the KL -divergence is associated to φ_1 , the KL_m to φ_0 , the χ^2 to φ_2 , the χ_m^2 to φ_{-1} and the Hellinger distance to $\varphi_{1/2}$. The Kullback-Leibler divergence (KL -divergence) is sometimes called Boltzmann-Shannon relative entropy. It appears in the domain of large deviations and it is frequently used in the domain of the reconstruction of laws, and in particular in the classical moment problem (see e.g. [Csiszár et al. \(1999\)](#) and the references therein). The modified Kullback-Leibler divergence (KL_m -divergence) is sometimes called Burg relative entropy. It is frequently used in Statistics and it leads to efficient methods in statistical estimation and tests problems; in fact, the celebrate “maximum likelihood” method can be seen as an optimization problem of the KL_m -divergence between the discrete or continuous parametric model and the empirical measure associated to the data; see e.g. [Keziou \(2003a\)](#) and [Broniatowski and Keziou \(2009\)](#). On the other hand, the more recent “empirical likelihood” method can also be seen as an optimization problem of the KL_m -divergence between some set of measures satisfying some linear constraints and the empirical measure associated to the data; see [Owen \(2001\)](#) and the references therein, [Bertail \(2004, 2006\)](#), and [Broniatowski and Keziou \(2012\)](#). The Hellinger divergence is also used in Statistics, it leads to robust statistical methods in parametric and semi-parametric models; see e.g. [Beran \(1977\)](#), [Lindsay \(1994\)](#) and [Jiménez and Shao \(2001\)](#). We extend the definition of the power divergences functions $Q \in M^1 \mapsto D_{\varphi_\gamma}(Q, P)$ onto the whole vector space of signed finite measures M via the extension of the definition of the convex functions φ_γ : For all $\gamma \in \mathbb{R}$ such that the function $x \mapsto \varphi_\gamma(x)$ is not defined on $] -\infty, 0[$ or defined but not convex on whole $\mathbb{R}$, we extend its definition as follows

$$x \in] -\infty, +\infty[\mapsto \begin{cases} \varphi_\gamma(x) & \text{if } x \in [0, +\infty[, \\ +\infty & \text{if } x \in] -\infty, 0[. \end{cases} \quad (1.5)$$

Note that for the χ^2 -divergence for instance, $\varphi(x) = \varphi_2(x) := \frac{1}{2}(x - 1)^2$ is defined and convex on whole $\mathbb{R}$.

The convex conjugate (or Fenchel-Legendre transform) of φ will be denoted φ^* , i.e.,

$$\varphi^*(\cdot) : t \in \mathbb{R} \mapsto \varphi^*(t) := \sup_{x \in \mathbb{R}} \{tx - \varphi(x)\}, \quad (1.6)$$

and the endpoints of dom_{φ^*} (the domain of φ^*) will be denoted a_{φ^*} and b_{φ^*} with $a_{\varphi^*} \leq b_{\varphi^*}$. Note that $\varphi^*(\cdot)$ is in turn proper closed convex function. In particular, $a_{\varphi^*} < 0 < b_{\varphi^*}$, $\varphi^*(0) = 0$ and

$$a_{\varphi^*} = \lim_{y \rightarrow -\infty} \frac{\varphi(y)}{y}, \quad b_{\varphi^*} = \lim_{y \rightarrow +\infty} \frac{\varphi(y)}{y}. \quad (1.7)$$

By the closedness of φ , the convex conjugate φ^{**} of φ^* coincides with φ , i.e.,

$$\varphi^{**}(t) := \sup_{x \in \mathbb{R}} \{tx - \varphi^*(x)\} = \varphi(t), \quad \text{for all } t \in \mathbb{R}. \quad (1.8)$$

Note that, the lower semi-continuity³ and the closedness properties are equivalent, for the proper convex functions defined on $\mathbb{R}$ (endowed with the usual topology). We say that φ (resp. φ^*) is differentiable if it is differentiable on $]a_\varphi, b_\varphi[$ (resp. $]a_{\varphi^*}, b_{\varphi^*}[$), the interior of its domain. We say also that φ (resp. φ^*) is strictly convex if it is strictly convex on $]a_\varphi, b_\varphi[$ (resp. $]a_{\varphi^*}, b_{\varphi^*}[$). The strict convexity of φ is equivalent to the condition that its conjugate φ^* is “essentially smooth”, i.e., differentiable with

$$\begin{aligned} \lim_{t \downarrow a_{\varphi^*}} \varphi^{*'}(t) &= -\infty & \text{if } a_{\varphi^*} > -\infty, \\ \lim_{t \uparrow b_{\varphi^*}} \varphi^{*'}(t) &= +\infty & \text{if } b_{\varphi^*} < +\infty. \end{aligned} \quad (1.9)$$

Conversely, φ is essentially smooth if and only if φ^* is strictly convex; see e.g. [Rockafellar \(1970\)](#) Section 26 for the proofs of these properties. If φ is differentiable, we denote φ' the derivative function of φ , and we define $\varphi'(a_\varphi)$ and $\varphi'(b_\varphi)$ to be the limits (which may be finite or infinite) $\lim_{x \downarrow a_\varphi} \varphi'(x)$ and $\lim_{x \uparrow b_\varphi} \varphi'(x)$, respectively. We denote $\text{Im}\varphi'$ the set of all values of the function φ' , i.e., $\text{Im}\varphi' := \{\varphi'(x) \text{ s.t. } x \in [a_\varphi, b_\varphi]\}$. If additionally the function φ is strictly convex, then φ' is increasing on $[a_\varphi, b_\varphi]$. Hence, it is one-to-one function from $[a_\varphi, b_\varphi]$ to $\text{Im}\varphi'$, we denote in this case φ'^{-1} the inverse function of φ' from $\text{Im}\varphi'$ to $[a_\varphi, b_\varphi]$. Note that if φ is differentiable, then for all $x \in]a_\varphi, b_\varphi[$, we have

$$\varphi^*(\varphi'(x)) = x\varphi'(x) - \varphi(x). \quad (1.10)$$

If additionally φ is strictly convex, then for all $t \in \text{Im}\varphi'$, we have

$$\varphi^*(t) = t\varphi'^{-1}(t) - \varphi(\varphi'^{-1}(t)) \quad \text{and} \quad \varphi^{*'}(t) = \varphi'^{-1}(t). \quad (1.11)$$

On the other hand, if φ is essentially smooth, then the interior of the domain of φ^* coincides with that of $\text{Im}\varphi'$, i.e., $(a_{\varphi^*}, b_{\varphi^*}) = (\varphi'(a_\varphi), \varphi'(b_\varphi))$. If additionally φ is twice continuously differentiable, then φ^* is twice continuously differentiable (on $]a_{\varphi^*}, b_{\varphi^*}[$) with $\varphi^{*'}(t) = \varphi'^{-1}(t)$ and $\varphi^{*''}(t) = \frac{1}{\varphi''(\varphi'^{-1}(t))}$. In particular, $\varphi^{*'}(0) = 1$ and $\varphi^{*''}(0) = 1$. Obviously, since φ is assumed to be closed, we have $\varphi(a_\varphi) = \lim_{x \downarrow a_\varphi} \varphi(x)$ and $\varphi(b_\varphi) = \lim_{x \uparrow b_\varphi} \varphi(x)$, which may be finite or infinite. Hence, by closedness of φ^* , we have also $\varphi^*(a_{\varphi^*}) = \lim_{t \downarrow a_{\varphi^*}} \varphi^*(t)$ and $\varphi^*(b_{\varphi^*}) = \lim_{t \uparrow b_{\varphi^*}} \varphi^*(t)$. Finally, the first and second derivatives of φ in a and b are defined to be the limits of $\varphi'(x)$ and $\varphi''(x)$ when $x \downarrow a$ and when $x \uparrow b$. The first and second derivatives of φ^* in a^* and b^* are defined in a similar way. Using the above properties, we give in Table 1.1 below, the convex conjugates φ^* of some functions φ , associated to standard divergences. We determine also their domains, (a_φ, b_φ) and $(a_{\varphi^*}, b_{\varphi^*})$.

Table 1.1: Convex conjugates for some standard divergences.

D_φ	φ	dom_φ	dom_{φ^*}	φ^*
KL_m	$\varphi(x) := -\log x + x - 1$	$]0, +\infty[$	$] -\infty, 1[$	$\varphi^*(t) = -\log(1 - t)$
KL	$\varphi(x) := x \log x - x + 1$	$[0, +\infty[$	$\mathbb{R}$	$\varphi^*(t) = e^t - 1$
$D_{\chi_m^2}$	$\varphi(x) := \frac{1}{2} \frac{(x-1)^2}{x}$	$]0, +\infty[$	$] -\infty, \frac{1}{2}[$	$\varphi^*(t) = 1 - \sqrt{1 - 2t}$
χ^2	$\varphi(x) := \frac{1}{2} (x - 1)^2$	$\mathbb{R}$	$\mathbb{R}$	$\varphi^*(t) = \frac{1}{2} t^2 + t$
H	$\varphi(x) := 2(\sqrt{x} - 1)^2$	$[0, +\infty[$	$] -\infty, 2[$	$\varphi^*(t) = \frac{2t}{2-t}$

The domain of the φ -divergence will be denoted $\text{dom}D_\varphi$, i.e.,

$$\text{dom}D_\varphi := \{Q \in M \text{ s.t. } D_\varphi(Q, P) < \infty\}. \quad (1.12)$$

³We say a function φ is lower semi-continuous if the level sets $\{x \text{ s.t. } \varphi(x) \leq \alpha\}$, $\forall \alpha \in \mathbb{R}$, are closed.

Definition 1.1. Let Ω be some subset in M . The φ -divergence between the set Ω and a p.m. P , noted $D_\varphi(\Omega, P)$, is

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

Definition 1.2. Assume that $D_\varphi(\Omega, P)$ is finite. A measure $Q^* \in \Omega$ such that

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

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

If φ is a strictly convex, then the function $Q \in M(P) \mapsto D_\varphi(Q, P)$ is strictly convex, and then the φ -projection of P on some convex set Ω , is uniquely defined, whenever it exists.

Let $g_i : \mathcal{X} \mapsto \mathbb{R}$, $i = 1, \dots, l$, be measurable real valued functions on $\mathcal{X}$. Denote $g := (g_0, g_1, \dots, g_l)^\top$ with $g_0 := \mathbf{1}_\mathcal{X}$. We assume that the functions $g_0, g_1, \dots, g_l$ are linearly independent in the sense that : $P(\{x \in \mathcal{X} \text{ s.t. } \lambda^\top g(x) \neq 0\}) > 0$ for any $\lambda \in \mathbb{R}^{1+l}$ with $\lambda \neq 0$. For all $\lambda \in \mathbb{R}^{1+l}$, we denote by $\lambda_0, \lambda_1, \dots, \lambda_l$ the $(1+l)$ coordinates of λ .

Denote M_g the set of all signed finite measures with total mass one, a.c. w.r.t. P , which integrate the functions g_i and satisfy a finite number of linear constraints, i.e.,

$$M_g := \left\{ Q \in M(P) \text{ s.t. } Q(\mathcal{X}) = 1 \text{ and } \int_{\mathcal{X}} g_i(x) dQ(x) = 0 \text{ for all } i = 1, \dots, l \right\}. \quad (1.13)$$

We consider the optimization problem

$$\inf_{Q \in M_g} D_\varphi(Q, P). \quad (1.14)$$

The Lagrangian “dual” problem associated with (1.14) is

$$\sup_{\lambda \in \mathbb{R}^{1+l}} \left\{ \lambda_0 - \int_{\mathcal{X}} \varphi^* \left(\lambda^\top g(x) \right) dP(x) \right\}. \quad (1.15)$$

We will consider the problem of the “dual” equality $\inf(1.14) = \sup(1.15)$, the existence of optimal solutions in (1.15), and in particular the problems of the existence and the characterization of the optimal solutions in (1.14), i.e., the φ -projections of P on the set M_g .

These problems intervene in the domain of the reconstruction of laws, in particular, the classical moment problem. Also they appear frequently in Statistics; in fact, the “empirical likelihood” method, which is the nonparametric version of the celebrated maximum likelihood method, can be seen as an optimization problem of the KL_m -divergence between some set of measures defined as in (1.13) and the empirical measure associated to the data; see e.g. [Broniatowski and Keziou \(2012\)](#).

In the vocabulary of the duality theory, a measure Q in M_g which realizes the infimum in (1.14) (i.e., a φ -projection of P on M_g in the vocabulary of φ -divergences theory) is called “a primal optimal solution” or simply “an optimal solution”, while a point λ in $\mathbb{R}^{1+l}$ which realizes the supremum in (1.15) is called “a dual optimal solution”.

For the optimization problem of convex function $f : \mathbb{R}^n \mapsto]-\infty, +\infty]$ on convex sets C in $\mathbb{R}^n$ subject to linear constraints $Ax = b \in \mathbb{R}^m$ where A is some $m \times n$ -matrix, a sufficient condition, in order that the equality

$$\inf_{\{x \in C; Ax=b\}} f(x) = \sup_{t \in \mathbb{R}^m} \left\{ b^\top t - f^* \left(A^\top t \right) \right\} \quad (1.16)$$

holds with dual attainment, is that there exists a point $\hat{x}$ in the relative interior⁴ of the convex set $C \cap \text{dom} f$ such that $A\hat{x} = b$; see e.g. [Rockafellar \(1970\)](#) for the proofs of these results.

⁴i.e., the interior in the real affine subspace $\langle C \cap \text{dom} f \rangle$ of $\mathbb{R}^n$ endowed with the relative topology of the usual topology on $\mathbb{R}^n$.

In order to make the set M_g closed and the linear functions $Q \in M \mapsto \int_{\mathcal{X}} g_i(x) dQ(x)$ continuous, which is required for applying the duality theory and solving the problem of existence of φ -projections of P on the set M_g , we endow the vector space M by the weak topology, denote it $\tau_{\mathcal{F}}$, induced by $\mathcal{F} \cup \mathcal{F}_b$, where $\mathcal{F} := \{g_0, g_1, \dots, g_l\}$ and $\mathcal{F}_b$ is the set of all bounded $\mathcal{B}$ -measurable real valued functions on $\mathcal{X}$; see Section 1.2 below for precise definition of the $\tau_{\mathcal{F}}$ -topology.

Note that the relative interior of the convex set M_g is generally empty in the weak topology $\tau_{\mathcal{F}}$. Borwein and Lewis (1992b) have extended the idea of the relative interior (r.i.) of convex sets in $\mathbb{R}^n$ to a new notion which have called “the quasi relative interior” (q.r.i.) of convex subsets of an arbitrary Hausdorff topological vector space X (finite or infinite dimension), that they have used to construct a powerful duality theory for the optimization problem of convex function $f : X \mapsto (-\infty, +\infty]$ on convex sets $C \subseteq X$ subject to linear constraints. In particular, when X is locally convex, they obtain similar results as in (1.16) when the relative interior is replaced by the quasi relative interior; see Borwein and Lewis (1992b) Corollary 4.8. The main advantage of the quasi relative interior of convex subset C of infinite dimension vector space X is that it is frequently nonempty even when the relative interior of C is empty.

If $\int_{\mathcal{X}} |g_i(x)| dP(x)$ is finite, for all $i = 1, \dots, l$, then the convex conjugate of the convex function $Q \mapsto D_{\varphi}(Q, P)$ (on the vector space $\mathcal{M}_{\mathcal{F}}(P)$ of all signed finite measures Q a.c. w.r.t. P and which integrate all the elements of $\mathcal{F}$, i.e., all the functions g_i) can be written as

$$D_{\varphi}^*(f) := \sup_{Q \in \mathcal{M}_{\mathcal{F}}(P)} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - D_{\varphi}(Q, P) \right\} = \int_{\mathcal{X}} \varphi^*(f(x)) dP(x), \text{ for all } f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle;$$

see Section 1.4 below for details. So, in this case, as in Borwein and Lewis (1991), we can apply Corollary 4.8 in Borwein and Lewis (1992a) to obtain the dual equality $\inf(1.14) = \sup(1.15)$ with dual attainment, whenever there exists a measure Q_0 in M_g which belongs to the quasi relative interior of $M_g \cap \text{dom} D_{\varphi}$. This condition is called “constraint qualification”. We can prove also from Borwein and Lewis (1992a) that a measure Q_0 is in the q.r.i. of $M_g \cap \text{dom} D_{\varphi}$ iff $a_{\varphi} < \frac{dQ_0}{dP} \leq \frac{d\hat{Q}}{dP} < b_{\varphi}$, P -almost everywhere (P -a.e.). We summarize these results and some other results about the problems of the existence and the characterization of the primal optimal solution (i.e., the φ -projection of P on M_g) in the following two Theorems and two Corollaries. For proofs, we refer to Borwein and Lewis (1992a) Theorem 3.10, Borwein and Lewis (1991) Corollary 2.6 and Theorem 4.8, and Csizsár *et al.* (1999) Theorem II.2.

Theorem 1.1. *If $\int_{\mathcal{X}} |g_i(x)| dP(x)$ is finite, for all $i = 1, \dots, l$, and if the following constraint qualification⁵:*

$$\text{there exists } \hat{Q} \in M_g \cap \text{dom} D_{\varphi} \text{ s.t. } a_{\varphi} < \frac{d\hat{Q}}{dP} \leq \frac{d\hat{Q}}{dP} < b_{\varphi} \quad (P - \text{a.e.}) \quad (1.17)$$

holds, then $\inf(1.14) = \sup(1.15)$ and there is attainment in (1.15). Suppose additionally that φ^ is essentially smooth (which is equivalent to the strict convexity of φ), and that there exists a dual optimal solution $\bar{\lambda}$ which is an interior point of the set*

$$\text{dom} D_{\varphi}^* := \left\{ \lambda \in \mathbb{R}^{1+l} \text{ s.t. } \int_{\mathcal{X}} \varphi^* \left(\lambda^{\top} g(x) \right) dP(x) \text{ is finite} \right\}, \quad (1.18)$$

then the unique optimal solution of (1.14) (i.e., the φ -projection of P on M_g), which we denote Q^ , exists and it is given by*

$$\frac{dQ^*}{dP}(x) = \varphi^{*'} \left(\bar{\lambda}^{\top} g(x) \right). \quad (1.19)$$

Remark 1.1. *In (1.18), for brevity, the definition of $\text{dom} D_{\varphi}^*$, which usually is the set of functions f such that $D_{\varphi}^*(f) < \infty$, is modified here.*

⁵The strict inequalities in (1.17) mean that $P \left\{ \frac{d\hat{Q}}{dP} \leq a_{\varphi} \right\} = P \left\{ \frac{d\hat{Q}}{dP} \geq b_{\varphi} \right\} = 0$.

Remark 1.2. If all functions g_i belong to $L_\infty(\mathcal{X}, P)$, and if for a dual optimal solution $\bar{\lambda} \in \mathbb{R}^{1+l}$ the following condition

$$a_{\varphi^*} < \text{essinf } \bar{\lambda}^\top g(\cdot) \leq \text{esssup } \bar{\lambda}^\top g(\cdot) < b_{\varphi^*} \quad (1.20)$$

holds, then $\bar{\lambda}$ is an interior point of $\text{dom} D_\varphi^*$. Hence, under assumption (1.20), all results in the above Theorem hold whenever the constraint qualification (1.17) is met.

If all functions g_i belong to $L_\infty(\mathcal{X}, P)$, and if additionally the convex function φ^* is everywhere finite (i.e., $a_{\varphi^*} = -\infty$ and $b_{\varphi^*} = +\infty$), then obviously condition (1.20) holds since $\text{dom} D_\varphi^* = \mathbb{R}^{1+l}$ in this case. Hence, under the constraint qualification (1.17), all results in the above Theorem hold. We state this result in the following Corollary.

Corollary 1.1. Suppose that all functions g_i belong to $L_\infty(\mathcal{X}, P)$ and that φ^* is everywhere finite (i.e., $a_{\varphi^*} = -\infty$ and $b_{\varphi^*} = +\infty$). If the constraint qualification (1.17) holds, then $\inf(1.14) = \sup(1.15)$ and there is attainment in (1.15). Suppose additionally that φ^* is everywhere differentiable (which is equivalent to the strict convexity of φ), then the unique optimal solution Q^* of (1.14) (i.e., the φ -projection of P on M_g) exists and it is given by

$$\frac{dQ^*}{dP}(x) = \varphi^{*'} \left(\bar{\lambda}^\top g(x) \right), \quad (1.21)$$

where $\bar{\lambda} \in \mathbb{R}^{1+l}$ is any dual optimal solution.

In the following Theorem and Corollary, we give sufficient conditions for the uniqueness of the dual optimal solution (see Borwein and Lewis (1991) Theorem 4.5 for the proof). Note that the strict convexity of φ^* is equivalent to the condition that its conjugate φ is essentially smooth.

Theorem 1.2. Suppose that all assumptions of Theorem 1.1 are met. Suppose furthermore that the function φ is essentially smooth. Then the dual optimal solution $\bar{\lambda}$ is unique. Moreover, the unique optimal solution Q^* of (1.14) exists and it is given by

$$\frac{dQ^*}{dP}(x) = \varphi^{*'} \left(\bar{\lambda}^\top g(x) \right) = \varphi'^{-1} \left(\bar{\lambda}^\top g(x) \right). \quad (1.22)$$

Corollary 1.2. Suppose that all assumptions of Corollary 1.1 are met. Suppose additionally that the function φ is essentially smooth. Then the dual optimal solution $\bar{\lambda}$ is unique. Moreover, the unique optimal solution Q^* of (1.14) exists and it is given by

$$\frac{dQ^*}{dP}(x) = \varphi^{*'} \left(\bar{\lambda}^\top g(x) \right) = \varphi'^{-1} \left(\bar{\lambda}^\top g(x) \right). \quad (1.23)$$

The important Corollary 1.1, which essentially requires that the constraint qualification (1.17) holds, applies in the KL -divergence case since the corresponding conjugate φ^* is everywhere finite (see also Borwein and Lewis (1993) for other examples), but it fails in the two important cases of Burg relative entropy (KL_m -divergence in the context of divergences) and Hellinger divergence without additional conditions since the corresponding conjugates φ^* are infinite on the intervals $[1, +\infty)$ and $[2, +\infty)$, respectively. Léonard (2001b) consider the optimization problem (1.14) when the set M_g is replaced by the subset

$$M_{os} := \left\{ Q \in M(P) \text{ s.t. } q := \frac{dQ}{dP} \in L_{\varphi_m^{**}} \text{ and } \int_{\mathcal{X}} g(x) dQ = (1, 0, \dots, 0)^\top \right\}, \quad (1.24)$$

where $L_{\varphi_m^{**}}$ is the Orlicz space defined as follows:

$$L_{\varphi_m^{**}} := \{ q : \mathcal{X} \rightarrow \mathbb{R}; \text{ measurable such that } \|q\|_{\varphi_m^{**}} < \infty \} \quad (1.25)$$

$$\text{with } \|q\|_{\varphi_m^{**}} := \inf \left\{ a > 0; \int_{\mathcal{X}} \varphi_m^{**} \left(\frac{|q(x)|}{a} \right) dP(x) \leq 1 \right\},$$

and φ_m^{**} is the convex conjugate of the convex function φ_m^* defined by $\varphi_m^*(t) := \max(\varphi^*(t), \varphi^*(-t))$ for all $t \in \mathbb{R}$. Without the constraint qualification (1.17), under the following integrability condition

$$\text{for any } \lambda \in \mathbb{R}^{1+l}, \quad \int_{\mathcal{X}} \varphi^* \left(\lambda^\top g(x) \right) dP(x) < \infty, \quad (1.26)$$

applying the duality theory on Orlicz spaces, Léonard (2001b) obtained the dual equality

$$\inf_{Q \in M_{os}} \varphi(Q, P) = \sup_{\lambda \in \mathbb{R}^{1+l}} \left\{ \lambda_0 - \int_{\mathcal{X}} \varphi^* \left(\lambda^\top g(x) \right) dP(x) \right\}. \quad (1.27)$$

Moreover, if the value is finite, then there exists at least one φ -projection of P on M_{os} ; see Léonard (2001b) Theorem 3.4 for details in more general context. A characterization of the φ -projections of P on the set M_{os} (with finite or infinite number of linear constraints) is stated by Léonard (2001c) under condition (1.26); see Léonard (2001c) Theorem 4.4, 4.5 and 4.6. Note that the integrability condition (1.26) implies that φ^* is everywhere finite, and these results apply in the important KL -divergence case with finite or infinite number of linear constraints. However, the condition (1.26) does not hold in the KL_m -divergence and Hellinger divergence cases since the domains of the corresponding φ^* functions are proper subsets of $\mathbb{R}$, and the important result (1.27) does not apply in these two important cases. Under the weaker integrability assumption

for any $\lambda \in \mathbb{R}^{1+l}$, there exists $\alpha > 0$ s.t.

$$\int_{\mathcal{X}} \varphi^* \left(\alpha \lambda^\top g(x) \right) dP(x) + \int_{\mathcal{X}} \varphi^* \left(-\alpha \lambda^\top g(x) \right) dP(x) < \infty, \quad (1.28)$$

the dual equality (1.27) may fail; see Léonard (2001b) Theorem 3.3.

The goal of the present Chapter is to present results of existence and characterization of the φ -projections of a given p.m. P on some subsets Ω of $\mathcal{M}_{\mathcal{F}}$, the space of all signed finite measures which integrate a given class $\mathcal{F}$ of functions, in particular, convex sets of signed finite measures defined by linear constraints as in (1.13) extending some previous works (about the existence and characterization of the φ -projections on subsets of $\mathcal{M}^1$, the set of all p.m.'s) of Csiszár (1975, 1984), Liese (1977), Rüschendorf (1984), Rüschendorf (1987), Liese and Vajda (1987), Teboulle and Vajda (1993) and Csiszár (1995). We give also different versions of dual representations of the φ -divergences viewed as convex functions on the space of all signed finite measures which integrate an arbitrary class of functions. When the set Ω is defined by linear constraints as in (1.13), we consider the dual problem, and we obtain the equality $\inf(1.14) = \sup(1.15)$ with dual attainment under different assumptions without constraint qualification. Additional conditions are given to obtain similar results which apply in the two important KL_m -divergence and Hellinger divergence cases.

Enhancing M^1 to M is motivated by the following arguments : sometimes the φ -projection, say Q_1^* , of a p.m. P on subset of M^1 is not an “interior” point and we can not give in this case a definite description of Q_1^* , while the φ -projection, say Q^* , of a p.m. P on subset of M is an “interior” point, which allows to give a perfect characterization of the φ -projection Q^* . In the context of statistical estimation and tests using the empirical likelihood method (see Owen (2001)), or related ones to criterions defined through divergences (see Broniatowski and Keziou (2012)), the projection of the empirical measure P_n of a sample on a set Ω^1 of p.m.'s may make problems when the projection is not an interior point of $\Omega^1 \cap \text{dom} D_\varphi(\cdot, P_n)$. Enhancing M^1 to M , this difficulty does not hold any longer, and tests as well as estimation can be performed.

The rest of this Chapter is organized as follows : In Section 1.2, we consider the problem of existence of φ -projections on general closed sets of signed measures. Section 1.3 deals with the problem of characterization of the φ -projections on sets of signed measures, in particular, sets of signed measures defined by linear constraints. In Section 1.4, we give different dual representations of φ -divergences seen as

convex functions on the vector space of all signed finite measures which integrate a given class $\mathcal{F}$ of functions. In Section 1.5, we apply the results of Sections 1.2, 1.3 and 1.4, to obtain the dual equality $\inf(1.14) = \sup(1.15)$ with dual attainment, under different assumptions without constraint qualification.

1.2 Existence of φ -projections on sets of signed measures

In this Section, we give sufficient conditions for the existence of φ -projections of some p.m. P on sets Ω of signed finite measures which integrate some class of functions (see Theorems 1.4, 1.5 and 1.6, and Corollary 1.3 below). First, we give some notation and we establish a convenient topological context for this problem. Let $\mathcal{F}$ be some class of measurable real valued functions f (bounded or unbounded) defined on $\mathcal{X}$. Here, $\mathcal{F}$ is not assumed to be finite. Denote $\mathcal{F}_b$ the set of all bounded measurable real valued functions defined on $\mathcal{X}$, and $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ the linear span of $\mathcal{F} \cup \mathcal{F}_b$. Define the set

$$\mathcal{M}_{\mathcal{F}}^1 := \left\{ Q \in M^1 \text{ s.t. } \int_{\mathcal{X}} |f(x)| dQ(x) < \infty, \text{ for all } f \text{ in } \mathcal{F} \right\},$$

and the real vector space

$$\mathcal{M}_{\mathcal{F}} := \left\{ Q \in M \text{ s.t. } \int_{\mathcal{X}} |f(x)| d|Q|(x) < \infty, \text{ for all } f \text{ in } \mathcal{F} \right\},$$

in which $|Q|$ denotes the total variation of the signed finite measure Q . Note that, when $\mathcal{F} = \mathcal{F}_b$, then $\mathcal{M}_{\mathcal{F}}^1 = M^1$ and $\mathcal{M}_{\mathcal{F}} = M$.

Definition 1.3. Denote $\tau_{\mathcal{F}}$ the weakest topology on $\mathcal{M}_{\mathcal{F}}$ for which all mappings $Q \in \mathcal{M}_{\mathcal{F}} \mapsto \int f dQ$ are continuous when f belongs to $\mathcal{F} \cup \mathcal{F}_b$. Denote also $\tau_{\mathcal{M}}$ the weakest topology on $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ for which all mappings $f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle \mapsto \int_{\mathcal{X}} f(x) dQ(x)$ are continuous when $Q \in \mathcal{M}_{\mathcal{F}}$. We sometimes call $\tau_{\mathcal{F}}$ the topology induced by $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ on $\mathcal{M}_{\mathcal{F}}$, and likewise $\tau_{\mathcal{M}}$ the topology induced by $\mathcal{M}_{\mathcal{F}}$ on $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$.

A base of open neighborhoods, for any R in $\mathcal{M}_{\mathcal{F}}$ endowed with the $\tau_{\mathcal{F}}$ -topology, is given by

$$U(R, \mathcal{A}, \varepsilon) := \left\{ Q \in \mathcal{M}_{\mathcal{F}} \text{ s.t. } \max_{f \in \mathcal{A}} \left| \int_{\mathcal{X}} f(x) dR(x) - \int_{\mathcal{X}} f(x) dQ(x) \right| < \varepsilon \right\} \quad (1.29)$$

for $\varepsilon > 0$ and $\mathcal{A}$ a finite collection of functions in $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$.

We refer to [Dunford and Schwartz \(1962\)](#) Chapter 5, for various topologies induced by classes of functions. Note that the class $\mathcal{F}_b$ induces the so-called τ -topology (see e.g. [Groeneboom et al. \(1979\)](#) and [Gänssler \(1971\)](#)), and that $\mathcal{M}_{\mathcal{F}_b}$ is the whole vector space M .

The above $\tau_{\mathcal{F}}$ -topology on $\mathcal{M}_{\mathcal{F}}$ is indeed the natural and the most convenient one in order to handle projection properties. It has been introduced in the context of large deviation probabilities by [Eichelsbacher and Schmock \(2002\)](#) for the Kullback-Leibler divergence and it is used in Statistics in [Broniatowski \(2003\)](#), [Keziou \(2003a\)](#), and [Broniatowski and Keziou \(2009\)](#). Usually the sets which are to be considered in statistical applications are not compact but merely closed sets; a typical example is when they are defined by linear constraints as in (1.13). Hence, the set M_g is closed in $\mathcal{M}_{\mathcal{F}}$ endowed with the $\tau_{\mathcal{F}}$ -topology if the functions g_i (which may be bounded or unbounded) belong to $\mathcal{F}$; this motivates the choice of $\tau_{\mathcal{F}}$ -topology.

Proposition 1.1. Equip $\mathcal{M}_{\mathcal{F}}$ with the $\tau_{\mathcal{F}}$ -topology and $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ with the $\tau_{\mathcal{M}}$ -topology. Then, $\mathcal{M}_{\mathcal{F}}$ and $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ are Hausdorff locally convex topological vector spaces. Further, the topological dual space of $\mathcal{M}_{\mathcal{F}}$ is the set of all mappings $Q \in \mathcal{M}_{\mathcal{F}} \mapsto \int_{\mathcal{X}} f(x) dQ(x)$ when f belongs to $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$, and the topological dual space of $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ is the set of all mappings $f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle \mapsto \int_{\mathcal{X}} f(x) dQ(x)$ when Q belongs to $\mathcal{M}_{\mathcal{F}}$.

We denote $[\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}]$ and $[\langle \mathcal{F} \cup \mathcal{F}_b \rangle; \tau_{\mathcal{M}}]$ the two Hausdorff locally convex topological vector spaces endowed with the $\tau_{\mathcal{F}}$ -topology and the $\tau_{\mathcal{M}}$ -topology, respectively.

The following proposition shows that $Q \in [\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}] \mapsto D_{\varphi}(Q, P)$ is lower semi-continuous (l.s.c.), provided that the corresponding convex function φ is closed; see Broniatowski and Keziou (2004) Proposition 2.3 and Keziou (2003a) Proposition 2.1.

Proposition 1.2. *The divergence function $Q \mapsto D_{\varphi}(Q, P)$ from $[\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}]$ to $[0, +\infty]$ is l.s.c.*

By the lower semi-continuity of the function $Q \in [\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}] \mapsto D_{\varphi}(Q, P)$, the following result holds.

Theorem 1.3. *Let P be some p.m. and Ω some compact subset of $[\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}]$. Then there exists at least one φ -projection of P on Ω .*

Using some similar arguments to those used in the proof of Theorem 2.4 in Liese (1977) or Proposition 8.5 in Liese and Vajda (1987) and Fenchel or Hölder inequalities, we state general results for the existence of φ -projections of some p.m. P on closed sets Ω of $[\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}]$ (see Theorems 1.5 and 1.6 below). First, in the following Theorem, we give a version of Theorem 2.4 in Liese (1977) or Proposition 8.5 in Liese and Vajda (1987).

Theorem 1.4. *Let Ω be some closed set in $[\mathcal{M}; \tau]$. Assume that the following two conditions*

$$D_{\varphi}(\Omega, P) := \inf_{Q \in \Omega} D_{\varphi}(Q, P) < \infty^6 \quad (1.30)$$

and

$$\lim_{|x| \rightarrow \infty} \frac{\varphi(x)}{|x|} = +\infty \quad (1.31)$$

hold. Then there exists at least one φ -projection of P on Ω .

Remark 1.3. *For sets Ω of p.m.'s defined by linear constraints, sufficient conditions for the existence of KL-projections are presented in Csiszár (1975) Theorem 3.1, Corollary 3.1 and Theorem 3.3. Sufficient conditions of the existence of φ -projections on sets of p.m.'s satisfying linear equality or inequality constraints are given in Csiszár (1995) Theorem 3.*

Remark 1.4. *By Eichelsbacher and Schmock (2002), whenever for all $\alpha > 0$ and all $f \in \mathcal{F}$, it holds $\int_{\mathcal{X}} \exp(\alpha|f(x)|) dP(x) < \infty$, then the level sets $\{Q \in \mathcal{M}_{\mathcal{F}}^1 \text{ s.t. } KL(Q, P) \leq c\}$ are compact in $[\mathcal{M}_{\mathcal{F}}^1; \tau_{\mathcal{F}}]$ for all real c . Therefore, for any $\tau_{\mathcal{F}}$ -closed set $\Omega \subset \mathcal{M}_{\mathcal{F}}^1$ for which $KL(\Omega, P) < \infty$, the projection of P on Ω exists; see Eichelsbacher and Schmock (2002) Lemma 2.1.*

Using Fenchel inequality and some similar arguments to that in Eichelsbacher and Schmock (2002) Lemma 2.1, We generalize the result of Csiszár (1995) Theorem 3 and the result in Remark 1.4 about the existence of projections, to the class of φ -divergences and to $\tau_{\mathcal{F}}$ -closed sets of signed measures.

Theorem 1.5. *Let Ω be some closed set in $\mathcal{M}_{\mathcal{F}}$ equipped with the $\tau_{\mathcal{F}}$ -topology. Suppose that the following three assumptions*

$$D_{\varphi}(\Omega, P) < \infty, \quad (1.32)$$

$$\lim_{|x| \rightarrow \infty} \frac{\varphi(x)}{|x|} = +\infty \quad (1.33)$$

$$\text{and for every } f \in \mathcal{F} \text{ and every } \alpha > 0, \int_{\mathcal{X}} \varphi^*(\alpha|f(x)|) dP(x) < \infty \quad (1.34)$$

hold. Then there exists at least one φ -projection of P on Ω .

⁶Note that this is equivalent to the condition: there exists $Q \in \Omega$ such that $D_{\varphi}(Q, P)$ is finite.

Using Hölder inequality, we give in the following Theorem an other result of existence of φ -projection on closed set in $[\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}]$. In the sequel, $\|\cdot\|_k$ denotes the usual norm of the vector space $L_k(\mathcal{X}, P)$, $1 \leq k \leq +\infty$.

Theorem 1.6. *Let Ω be some closed set in $\mathcal{M}_{\mathcal{F}}$ equipped with the $\tau_{\mathcal{F}}$ -topology. Assume that the following conditions*

$$D_{\varphi}(\Omega, P) < \infty, \quad (1.35)$$

there exists numbers $1 < r, k < +\infty$ such that $r^{-1} + k^{-1} = 1$,

$$\lim_{|x| \rightarrow \infty} \frac{\varphi(x)}{|x|^r} > 0, \quad \text{and for every } f \in \mathcal{F}, \quad \|f\|_k < \infty$$

hold. Then there exists at least one φ -projection of P on Ω .

Note that the above results do not apply in the case of KL_m and Hellinger divergences since the condition $\lim_{|x| \rightarrow \infty} \frac{\varphi(x)}{|x|} = +\infty$ does not hold. The following Corollary applies without assumption $\lim_{|x| \rightarrow \infty} \frac{\varphi(x)}{|x|} = +\infty$, in particular, in the KL_m and Hellinger divergences cases.

Corollary 1.3. *Let Ω be a closed set in $[M; \tau]$. If the following condition*

$$\text{there exists } u, l \in L_1(\mathcal{X}, P) \text{ such that } u \leq \frac{dQ}{dP} \leq l \text{ (} P - a.e.) \text{ for all } Q \in \Omega \cap \text{dom} D_{\varphi} \quad (1.36)$$

holds, then there exists at least one φ -projection of P on Ω , whenever $D_{\varphi}(\Omega, P)$ is finite.

1.3 Characterization of φ -projections on sets of signed measures

In this Section, we extend known results pertaining to the characterization of the φ -projections as can be found in Rüschendorf (1984), Rüschendorf (1987), Liese and Vajda (1987), (see also Csiszár (1975) and Csiszár (1984) for the characterization of KL -projections). These authors have characterized the φ -projections on subsets of M^1 . We expose similar results when considering subsets of M . We first consider the case of general subsets Ω of M and then the case of convex subsets of M defined by linear constraints. Throughout the section, we assume that the convex function φ is differentiable.

1.3.1 on general sets

We will use the following assumption

$$\begin{aligned} &\text{There exists } 0 < \delta < 1 \text{ such that for all } c \text{ in } [1 - \delta, 1 + \delta], \text{ we can find numbers } c_1, c_2, c_3 \\ &\text{such that } \varphi(cx) \leq c_1 \varphi(x) + c_2 |x| + c_3, \text{ for all real } x. \end{aligned} \quad (1.37)$$

Remark 1.5. *Condition (1.37) holds for all power divergences including KL , KL_m and Hellinger divergences. Note also that condition (1.37) implies that a_{φ} equals 0 or $-\infty$ and b_{φ} equals $+\infty$.*

Theorem 1.7. *Let Ω be a subset of M and Q^* be a signed measure in $\Omega \cap \text{dom} D_{\varphi}$. Then*

(a) *Sufficient conditions, for Q^* to be a φ -projection of P on Ω , are that (i) $\varphi'(q^*)q \in L_1(\mathcal{X}, P)$ and (ii) $\int_{\mathcal{X}} \varphi'(q^*) dQ^* \leq \int_{\mathcal{X}} \varphi'(q^*) dQ$, for all Q in $\Omega \cap \text{dom} D_{\varphi}$.*

(b) *If condition (1.37) holds and Ω is convex, then these conditions are necessary as well.*

1.3.2 on sets defined by linear constraints

In this subsection, we consider the problems of existence and characterization of φ -projections of some p.m. P on linear set S of measures in M defined by arbitrary family of constraints. So, let $\mathcal{G}$ denote a collection (finite or infinite, countable or not) of real valued functions defined on $(\mathcal{X}, \mathcal{B})$. The class $\mathcal{G}$ is assumed to contain the function $\mathbb{1}_{\mathcal{X}}$. The set S is defined by

$$S := \left\{ Q \in \mathcal{M}_{\mathcal{G}}(P) \text{ such that } \int_{\mathcal{X}} dQ = 1 \text{ and } \int_{\mathcal{X}} g dQ = 0, \text{ for all } g \text{ in } \mathcal{G} \setminus \{\mathbb{1}_{\mathcal{X}}\} \right\}. \quad (1.38)$$

The following result states the explicit form of Q^* , a φ -projection of P on S , when it exists.

Theorem 1.8.

- (a) Let Q^* be some finite measure in $S \cap \text{dom} D_{\varphi}$. A sufficient condition, for Q^* to be a φ -projection of P on S , is that there exists numbers $c_1, \dots, c_d \in \mathbb{R}$ and functions $g_1, \dots, g_d \in \mathcal{G}$ such that $\varphi'(q^*(x)) = c_1 g_1(x) + \dots + c_d g_d(x)$ (P -a.e.).
- (b) Assume that condition (1.37) holds. Then, any φ -projection, say Q^* , of P on S , if it exists, satisfies $\varphi'(q^*)$ belongs to $\overline{\langle \mathcal{G} \rangle}$, the closure of $\langle \mathcal{G} \rangle$ in $L_1(\mathcal{X}, |Q^*|)$.

If $\mathcal{G}$ is a finite collection of functions in $L_1(\mathcal{X}, |Q^*|)$, then the vector space $\langle \mathcal{G} \rangle$ is closed in $L_1(\mathcal{X}, |Q^*|)$. So, from the above Theorem, we can state the following Corollary :

Corollary 1.4. Let $\mathcal{G} := \{\mathbb{1}_{\mathcal{X}}, g_1, \dots, g_l\}$ be a finite collection of measurable functions on $\mathcal{X}$. Then the following results hold.

- (a) Let Q^* be some measure in $S \cap \text{dom} D_{\varphi}$. A sufficient condition, for Q^* to be a φ -projection of P on S , is that there exists some constant $c \in \mathbb{R}^{1+l}$ s.t. $\varphi' \left(\frac{dQ^*}{dP}(x) \right) = c_0 + \sum_{i=1}^l c_i g_i(x)$ (P -a.e.).
- (b) Assume that condition (1.37) holds. Then any φ -projection, say Q^* , of P on S , if it exists, satisfies : there exists some constant $c \in \mathbb{R}^{1+l}$ s.t. $\varphi' \left(\frac{dQ^*}{dP}(x) \right) = c_0 + \sum_{i=1}^l c_i g_i(x)$ ($|Q^*|$ -a.e.).

It should be noticed that the above Theorem and Corollary do not provide a definite description of the projected measure; indeed, it does not give any information on the support of $|Q^*|$. However, if $\varphi(0) = +\infty$ (which holds for example for the KL_m -divergence), then any φ -projection Q^* of P on some set Ω , if it exists, has obviously the same support as P when $D_{\varphi}(\Omega, P)$ is finite. Furthermore, we prove in the following Lemma that if $\varphi'(0) = -\infty$ (which holds for instance in the case of KL , KL_m and Hellinger divergences), then any φ -projection of P on some convex set Ω when it exists has the same support as P . First, state the following Corollary which applies in the KL_m -divergence case.

Corollary 1.5. Let $\mathcal{G}$ be defined as in Corollary 1.4. Assume that assumption (1.37) holds. Suppose additionally that $\varphi(0) = +\infty$, and let Q^* be some p.m. in $S \cap \text{dom} D_{\varphi}$. Then the following result holds : Q^* is a φ -projection of P on S iff there exists some constant $c \in \mathbb{R}^{1+l}$ such that

$$\varphi' \left(\frac{dQ^*}{dP}(x) \right) = c_0 + \sum_{i=1}^l c_i g_i(x) \quad (P - \text{a.e.}).$$

Lemma 1.1. Assume that condition (1.37) holds, $a_{\varphi} = 0$ and $\varphi'(0) = -\infty$. Let Ω be some convex set of signed finite measures. If there exists some $Q_0 \in \Omega \cap \text{dom} D_{\varphi}$ such that $\frac{dQ_0}{dP} > 0$ (P -a.e.), then any φ -projection, say Q^* , of P on Ω , if it exists, has the same support as P , i.e., $\frac{dQ^*}{dP} > 0$ (P -a.e.).

We can now state, from the above Theorem, the following Corollary which applies in the case of KL , KL_m and Hellinger divergences.

Corollary 1.6. *Let $\mathcal{G}$ be defined as in Corollary 1.4. Assume that assumption (1.37) holds. Suppose additionally that $a_\varphi = 0$ and $\varphi'(0) = -\infty$. If there exists some $Q_0 \in S \cap \text{dom} D_\varphi$ such that $\frac{dQ_0}{dP} > 0$ (P -a.e.), then the following result holds : a p.m. Q^* in $S \cap \text{dom} D_\varphi$ is a φ -projection of P on S iff there exists some constant $c \in \mathbb{R}^{1+l}$ s.t. $\varphi' \left(\frac{dQ^*}{dP}(x) \right) = c_0 + \sum_{i=1}^l c_i g_i(x)$ (P -a.e.).*

Remark 1.6. *Versions of Theorem 1.8, for sets of p.m.'s, have been proved by Csiszár (1975) and Csiszár (1984) for the Kullback-Leibler divergence, and by Rüschendorf (1984) and Rüschendorf (1984) for φ -divergences between p.m.'s. Broniatowski and Keziou (2006) give the proof of Theorem 1.8, in the present context, that is when the set S , see (1.38), is a subset of signed finite measures.*

1.4 Fenchel duality for φ -divergences

We refer to Fenchel (1949), Moreau (1962), Brøndsted (1964), Rockafellar (1968, 1974) and Ekeland and Témam (1999) for the notion of Fenchel duality of general convex functions on general vector spaces. We consider this notion for φ -divergences functionals $Q \mapsto D_\varphi(Q, P)$ viewed as convex functions on the vector space of signed finite measures $\mathcal{M}_{\mathcal{F}}$; we give different versions of dual representations for φ -divergences (see Theorems 1.9, 1.10, 1.11 and 1.12). In view of the result of Proposition 1.1, we identify the topological dual space of $[\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}]$ with $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ and the topological dual space of $[\mathcal{F} \cup \mathcal{F}_b; \tau_{\mathcal{M}}]$ with $\mathcal{M}_{\mathcal{F}}$. Hence, the Fenchel-Legendre transform (i.e., the convex conjugate) of the convex function $Q \in [\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}] \mapsto D_\varphi(Q, P) \in [0, +\infty]$ is defined as follows

$$D_\varphi^*(\cdot) : f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle \mapsto D_\varphi^*(f) := \sup_{Q \in \mathcal{M}_{\mathcal{F}}} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - D_\varphi(Q, P) \right\}, \quad (1.39)$$

which is convex and lower semi-continuous⁷ w.r.t. the $\tau_{\mathcal{M}}$ -topology, the weak topology induced on $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ by $\mathcal{M}_{\mathcal{F}}$.

By the lower semi-continuity of the convex function $Q \in [\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}] \mapsto D_\varphi(Q, P) \in [0, +\infty]$ (see Proposition 1.2 above), applying the Fenchel duality theory (see e.g. Rockafellar (1968), Fenchel (1949) or Dembo and Zeitouni (1998) Lemma 4.5.8), we can state the following result for any φ -divergence.

Theorem 1.9. *The function $Q \in [\mathcal{M}_{\mathcal{F}}; \tau_{\mathcal{F}}] \mapsto D_\varphi(Q, P) \in [0, +\infty]$ is the conjugate of its conjugate $f \in [\langle \mathcal{F} \cup \mathcal{F}_b \rangle; \tau_{\mathcal{M}}] \mapsto D_\varphi^*(f)$ defined by (1.39). In other words, the φ -divergence $D_\varphi(Q, P)$ admits the dual representation*

$$D_\varphi(Q, P) = \sup_{f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - D_\varphi^*(f) \right\}, \text{ for all } Q \in \mathcal{M}_{\mathcal{F}}, \quad (1.40)$$

where $D_\varphi^*(\cdot)$ is defined by (1.39).

Consider now the problem of computing $D_\varphi^*(f)$ (in particular the equality $D_\varphi^*(f) = \int_{\mathcal{X}} \varphi^*(f(x)) dP(x)$), and the problems of existence, uniqueness and characterization of a dual optimal solution in (1.40) (i.e., a function $\bar{f} \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$ realizing the supremum in (1.40)), when $D_\varphi(Q, P)$ is finite.

In the following Proposition, when φ is strictly convex and differentiable, we give the explicit form of $D_\varphi^*(f)$ for all $f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$ such that $\text{Im} f \subseteq \text{Im} \varphi'$.

Proposition 1.3. *Assume that φ is strictly convex and differentiable, and that for all $f, g \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$ such that $\text{Im} f \subseteq \text{Im} \varphi'$, the integrals*

$$\int_{\mathcal{X}} |g(x)| \left| \varphi'^{-1}(f(x)) \right| dP(x) \text{ and } \int_{\mathcal{X}} \varphi \left(\varphi'^{-1}(f(x)) \right) dP(x) \text{ are finite.} \quad (1.41)$$

Then, for all $f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$ s.t. $\text{Im} f \subseteq \text{Im} \varphi'$, we have

$$D_\varphi^*(f) \text{ is finite, and } D_\varphi^*(f) = \int_{\mathcal{X}} \varphi^*(f(x)) dP(x) = \int_{\mathcal{X}} \left[f \varphi'^{-1}(f) - \varphi \left(\varphi'^{-1}(f) \right) \right] dP. \quad (1.42)$$

⁷Note that the conjugate of a convex function is always l.s.c. w.r.t. the weak topology.

Remark 1.7. If the convex function $Q \in \mathcal{M}_{\mathcal{F}} \mapsto D_{\varphi}(Q, P) \in [0, +\infty]$ is proper, i.e.,

$$\text{there exists at least one measure } Q_0 \text{ in } \mathcal{M}_{\mathcal{F}}(P) \text{ s.t. } D_{\varphi}(Q_0, P) \text{ is finite,} \quad (1.43)$$

then the integral $\int_{\mathcal{X}} \varphi^*(f(x)) dP(x)$ is well defined for all $f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$. Indeed, for all $f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$ and for all $x \in \mathcal{X}$, by Fenchel's inequality, we have

$$\varphi^*(f(x)) \geq f(x) \frac{dQ_0}{dP}(x) - \varphi\left(\frac{dQ_0}{dP}(x)\right). \quad (1.44)$$

The RHS term belong to $L_1(\mathcal{X}, P)$ by assumption (1.43). Hence, the integral $\int_{\mathcal{X}} \varphi^*(f) dP$ is well defined. Moreover, we have $-\infty < \int_{\mathcal{X}} \varphi^*(f) dP \leq +\infty$ for all $f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$. Hence, from Theorem 1.9 we can state the following result.

Theorem 1.10. Assume that φ is differentiable. Then, for all $Q \in \mathcal{M}_{\mathcal{F}}$ such that $D_{\varphi}(Q, P)$ is finite and $\varphi'\left(\frac{dQ}{dP}\right)$ belongs to $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$, the φ -divergence $D_{\varphi}(Q, P)$ admits the dual representation

$$D_{\varphi}(Q, P) = \sup_{f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - \int_{\mathcal{X}} \varphi^*(f(x)) dP(x) \right\}, \quad (1.45)$$

and the function $\bar{f} := \varphi'\left(\frac{dQ}{dP}\right)$ is a dual optimal solution. Furthermore, if φ is essentially smooth, then $\bar{f}$ is the unique dual optimal solution (P -a.e.).

Theorem 1.10 remains valid if we substitute the vector space $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ by the arbitrary class of functions $\mathcal{F}$. We state this result in the following Theorem.

Theorem 1.11. Assume that φ is differentiable. Let $\mathcal{F}$ be an arbitrary class of measurable real valued functions on $\mathcal{X}$. Then, for all $Q \in \mathcal{M}_{\mathcal{F}}$ such that $D_{\varphi}(Q, P)$ is finite and $\varphi'\left(\frac{dQ}{dP}\right)$ belongs to $\mathcal{F}$, the φ -divergence $D_{\varphi}(Q, P)$ admits the dual representation

$$D_{\varphi}(Q, P) = \sup_{f \in \mathcal{F}} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - \int_{\mathcal{X}} \varphi^*(f(x)) dP(x) \right\}, \quad (1.46)$$

and the function $\bar{f} := \varphi'\left(\frac{dQ}{dP}\right)$ is a dual optimal solution. Furthermore, if φ is essentially smooth, then $\bar{f}$ is the unique dual optimal solution (P -a.e.).

Remark 1.8. Theorem 1.11, with an appropriate choice of the class $\mathcal{F}$, has been used by Keziou (2003a) and Broniatowski and Keziou (2009) to introduce a new common definition of the “minimum φ -divergence estimates” in discrete or continuous parametric models. Note that the “plug-in” minimum φ -divergence estimates introduced by Liese and Vajda (1987) Chapter 10 are defined only in discrete parametric models, see also Lindsay (1994) and Morales et al. (1995). The use of the dual representation (1.46) allows to give a common definition of the minimum φ -divergence estimates in discrete or continuous parametric models.

Remark 1.9. Other versions of dual representations of φ -divergences are given in Borwein and Lewis (1991) on $L_k(\mathcal{X}, P)$ spaces, in Borwein and Lewis (1993) on compact metric spaces, and in Léonard (2001a) on Orlicz spaces. See also Rockafellar (1968) for other convex integral functionals on some “decomposable” spaces.

Under the following integrability assumption

$$\int_{\mathcal{X}} |f(x)| dP(x) \text{ is finite, for all } f \in \mathcal{F}, \quad (1.47)$$

the convex function $Q \in [\mathcal{M}_{\mathcal{F}}(P); \tau_{\mathcal{F}}] \mapsto \varphi(Q, P) \in [0, +\infty]$ is proper. Its Fenchel-Legendre transform is

$$D_{\varphi}^*(\cdot) : f \in [\langle \mathcal{F} \cup \mathcal{F}_b \rangle; \tau_{\mathcal{M}}] \mapsto D_{\varphi}^*(f) := \sup_{Q \in \mathcal{M}_{\mathcal{F}}(P)} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - D_{\varphi}(Q, P) \right\} \in (-\infty, +\infty], \quad (1.48)$$

which is convex and lower semi-continuous. Following Rockafellar (1968), denote $L^* = \langle \mathcal{F} \cup \mathcal{F}_b \rangle$ and $L = \mathcal{M}_{\mathcal{F}}(P)$, then condition (1.47) implies that both L^* and L are decomposable. Hence, we can apply the Corollary of Theorem 2 in Rockafellar (1968), to obtain the following result :

Theorem 1.12. *Under assumption (1.47), the convex conjugate function $f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle \mapsto D_{\varphi}^*(f)$ defined by (1.48) is proper, it can be written as*

$$D_{\varphi}^*(f) = \int_{\mathcal{X}} \varphi^*(f(x)) dP(x), \text{ for all } f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle, \quad (1.49)$$

and the φ -divergence $D_{\varphi}(Q, P)$ admits the dual representation

$$D_{\varphi}(Q, P) = \sup_{f \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle} \left\{ \int_{\mathcal{X}} f(x) dQ(x) - \int_{\mathcal{X}} \varphi^*(f(x)) dP(x) \right\}, \text{ for all } Q \in \mathcal{M}_{\mathcal{F}}(P). \quad (1.50)$$

In particular, the function $Q \in [\mathcal{M}_{\mathcal{F}}(P); \tau_{\mathcal{F}}] \rightarrow D_{\varphi}(Q, P) \in [0, +\infty]$ is lower semi-continuous.

Remark 1.10. *The lower semi-continuity property of the function $Q \in [\mathcal{M}_{\mathcal{F}}(P); \tau_{\mathcal{F}}] \mapsto D_{\varphi}(Q, P) \in [0, +\infty]$ holds from Proposition 1.2 without assumption (1.47). On the other hand, Theorem 1.11 and 1.12 are of interest particularly when $D_{\varphi}(Q, P)$ is finite and the class $\mathcal{F}$ contains the function $f = \varphi'(dQ/dP)$. In Theorem 1.11, condition on Q , i.e., $\int_{\mathcal{X}} |\varphi'(dQ/dP)| d|Q| < \infty$, holds whenever $D_{\varphi}(Q, P)$ is finite and φ satisfies condition (1.37). However, in Theorem 1.12, these conditions do not inevitably imply assumption (1.47) if the class $\mathcal{F}$ contains $\varphi'(dQ/dP)$. It is the case, for example, when $D_{\varphi} = KL$, Q is a normal law and P is a Cauchy law. Indeed, $KL(Q, P)$ is finite, the assumption $\int |\log(dQ/dP)| dQ < \infty$ in Theorem 1.11 holds while the assumption $\int |\log(dQ/dP)| dP < \infty$ in Theorem 1.12 does not. This shows the interest of our Proposition 1.2 and Theorem 1.11.*

1.5 Applications to the minimization of divergences on sets of signed finite measures satisfying linear constraints

In this Section, we apply the results in Sections 1.2, 1.3 and 1.4 to the problem $\inf_{Q \in M_g} D_{\varphi}(Q, P)$, where M_g is defined in (1.13). Under different assumptions, we obtain the dual equality $\inf(1.14) = \sup(1.15)$ and results about the problems of existence, uniqueness and characterization of the dual optimal solution and the φ -projections of P on the set M_g . We state our results under the following assumptions:

$$\text{the convex function } \varphi \text{ is differentiable;} \quad (1.51)$$

$$\text{there exists at least one } \varphi\text{-projection } Q^* \text{ of } P \text{ on } M_g \text{ with the same support as } P. \quad (1.52)$$

Theorem 1.13. *Assume that conditions (1.37), (1.51) and (1.52) hold. Then,*

- (a) *there exists $\bar{\lambda} \in \mathbb{R}^{1+l}$ such that $\varphi' \left(\frac{dQ^*}{dP}(x) \right) = \bar{\lambda}_0 + \sum_{i=1}^l \bar{\lambda}_i g_i(x)$ ($P - a.e.$),*
- (b) *the equality $\inf_{Q \in M_g} D_{\varphi}(Q, P) = \sup_{\lambda \in \mathbb{R}^{1+l}} \left\{ \lambda_0 - \int_{\mathcal{X}} \varphi^* (\lambda^{\top} g(x)) dP(x) \right\}$ holds, and $\bar{\lambda}$ is a dual optimal solution. Furthermore, if the function φ is essentially smooth, then the dual optimal solution $\bar{\lambda}$ is unique.*

Remark 1.11. *Under assumptions of Theorem 1.13, the φ -projection of P on M_g is characterized without supposing that $\bar{\lambda}$ is an interior point of $\text{dom} D_{\varphi}^*$. Furthermore, the dual equality holds and the dual optimal solution is attained. Sufficient conditions for assumption (1.52) are given in Corollary 1.7 and Proposition 1.4 below.*

Corollary 1.7. *Assume that φ is differentiable and strictly convex. If there exists some $\bar{\lambda} \in \mathbb{R}^{1+l}$ such that*

$$\int_{\mathcal{X}} \varphi \left(\varphi'^{-1} \left(\bar{\lambda}^\top g(x) \right) \right) dP < \infty \text{ and } \int_{\mathcal{X}} g(x)^\top \varphi'^{-1} \left(\bar{\lambda}^\top g(x) \right) dP(x) = (1, 0, \dots, 0)^\top, \quad (1.53)$$

then,

- (a) *the measure Q^* , defined by $dQ^*(x) = \varphi'^{-1} \left(\bar{\lambda}^\top g(x) \right) dP(x)$, is the unique φ -projection of P on M_g .*
- (b) *the equality $\inf_{Q \in M_g} D_\varphi(Q, P) = \sup_{\lambda \in \mathbb{R}^{1+l}} \{ \lambda_0 - \int_{\mathcal{X}} \varphi^* (\lambda^\top g(x)) dP(x) \}$ holds, and $\bar{\lambda}$ is a dual optimal solution. Furthermore, if the function φ is essentially smooth, then the dual optimal solution $\bar{\lambda}$ is unique.*

In particular, the condition (1.53) holds if there exists a dual optimal solution $\bar{\lambda}$ which is an interior point of $\text{dom} D_\varphi^ := \{ \lambda \in \mathbb{R}^{1+l} \text{ s.t. } \int_{\mathcal{X}} |\varphi^* (\lambda^\top g(x))| dP(x) \text{ is finite} \}$.*

Remark 1.12. *Note that, if φ is differentiable and strictly convex, then under the assumption (1.37), the conditions (1.52) and (1.53) are equivalent; see Theorem 1.8 part (a) and (b).*

We give now in the following Proposition other sufficient conditions for assumption (1.52). The conditions are

$$D_\varphi(M_g, P) < \infty; \quad (1.54)$$

$$\lim_{|x| \rightarrow \infty} \frac{\varphi(x)}{|x|} = +\infty; \quad (1.55)$$

$$\text{for all } \alpha > 0, \text{ and all } i = 1, \dots, l, \quad \int_{\mathcal{X}} \varphi^* (\alpha |g_i(x)|) dP(x) < \infty; \quad (1.56)$$

there exists numbers $1 < r, k < +\infty$ such that $r^{-1} + k^{-1} = 1$,

$$\lim_{|x| \rightarrow \infty} \frac{\varphi(x)}{|x|^r} > 0, \text{ and for all } i = 1, \dots, l, \quad \|g_i\|_k < \infty; \quad (1.57)$$

$$\text{the functions } g_1, \dots, g_l \text{ belong to } L_\infty(\mathcal{X}, P); \quad (1.58)$$

$$\varphi(0) = +\infty; \quad (1.59)$$

$$a_\varphi = 0 \quad \text{and} \quad \varphi'(0) = -\infty; \quad (1.60)$$

$$\text{there exists some } Q_0 \in M_g \cap \text{dom} D_\varphi \text{ such that } \frac{dQ_0}{dP} > 0 \quad (P - a.e.). \quad (1.61)$$

Proposition 1.4.

- (a) *Under assumptions (1.54), (1.55), (1.56) and (1.59), condition (1.52) holds.*
- (b) *Condition (1.52) holds also under assumptions (1.54), (1.57) and (1.59).*
- (c) *Condition (1.52) holds also if, in part (a), (1.56) is replaced by (1.58) or/and if condition (1.59) is replaced by [(1.51), (1.60) and (1.61)].*
- (d) *Condition (1.52) holds also if, in part (b), condition (1.59) is replaced by [(1.51), (1.60) and (1.61)].*

Chapter 2

Multivariate divergences

This Chapter presents our results obtained in [Keziou \(2015\)](#) and [Keziou \(2016\)](#). We introduce multivariate divergences between K , $K \geq 1$, signed finite measures (s.f.m.) $(Q_1, \dots, Q_K)$ and a given reference probability measure P on a σ -field $(\mathcal{X}, \mathcal{B})$, extending the well known divergences between two measures. We investigate the Fenchel duality theory for the introduced multivariate divergences viewed as convex functionals on well chosen topological vector spaces of signed finite measures. We obtain new dual representations of these criteria, which we will use, in Part 2 Chapter 6, to define a new family of estimates and test statistics with multiple samples under semiparametric density ratio models. This family contains the estimate and test statistic obtained through dual empirical likelihood, c.f. [Qin \(1998\)](#) and [Keziou and Leoni-Aubin \(2008\)](#), in the particular case of two-sample problem. Moreover, the present approach allows obtaining the asymptotic properties of the estimates and test statistics both under the null and the alternative hypotheses. This leads to accurate approximations of the power function for any used criterion, including the dual empirical likelihood one, which is of its own interest. Moreover, the multivariate divergences context leads to a new definition of Information criterion, and by product a new model selection procedure with multiple samples under density ratio models, as well as a new procedure for multi-group classification, in statistical learning.

2.1 Introduction, notation and motivation

On a measurable space $(\mathcal{X}, \mathcal{B})$, let P_0 be a given reference probability distribution. Denote by M the real vector space of all signed finite measures (s.f.m.) on $(\mathcal{X}, \mathcal{B})$. Let

$$\psi(\cdot) : \mathbf{x} := (x_1, \dots, x_K)^\top \in \mathbb{R}^K \mapsto \psi(\mathbf{x}) \in [0, +\infty] \quad (2.1)$$

be any nonnegative proper¹ closed² convex function on $\mathbb{R}^K$, satisfying $\psi(\mathbf{1}_K) = 0$, where $\mathbf{1}_K := (1, \dots, 1)^\top \in \mathbb{R}^K$, and such that (s.t.) $\mathbf{1}_K$ is an interior point of the domain of $\psi(\cdot)$

$$\text{dom}_\psi := \{ \mathbf{x} \in \mathbb{R}^K \text{ s.t. } |\psi(\mathbf{x})| < \infty \}.$$

We define the ψ -multivariate-divergence between K -signed finite measures $(Q_1, \dots, Q_K)$ and a reference probability measure P_0 through

$$D_\psi(Q_1, \dots, Q_K; P_0) := \int_{\mathcal{X}} \psi \left(\frac{dQ_1}{dP_0}(x), \dots, \frac{dQ_K}{dP_0}(x) \right) dP_0(x), \quad (2.2)$$

¹ A function ψ is called proper if its domain is a non-empty set, and $\psi(\mathbf{x}) > -\infty$, for all $\mathbf{x}$.

² A function $\psi : \mathbf{x} \in \mathbb{R}^K \mapsto \psi(\mathbf{x}) \in \mathbb{R} \cup \{+\infty\}$ is called closed if its epigraph $\text{epi}_\psi := \{(\mathbf{x}, y) \in \mathbb{R}^K \times \mathbb{R} \text{ s.t. } \psi(\mathbf{x}) \leq y\}$ is a closed set in $\mathbb{R}^{K+1}$ endowed with its usual topology. Note that, for convex functions ψ on (finite dimension space) $\mathbb{R}^K$, the closeness of ψ is equivalent to the lower semi-continuity property, that is all level sets $\{\mathbf{x} \in \mathbb{R}^K \text{ s.t. } \psi(\mathbf{x}) \leq \alpha\}$, $\forall \alpha \in \mathbb{R}$, are closed in $\mathbb{R}^K$.

if $Q_1, \dots, Q_K$ are all absolutely continuous (a.c.) with respect to (w.r.t.) P_0 ; If there exists at least one measure Q_k not a.c.w.r.t. P_0 , we set $D_\psi(Q_1, \dots, Q_K; P_0) = +\infty$. In the above integral, $\frac{dQ_k}{dP_0}(\cdot)$ is the Radon-Nikodym derivative of Q_k w.r.t. P_0 . Obviously, if $K = 1$, we obtain the classical definition of φ -divergences between two measures (a signed finite measure Q_1 and a probability measure P_0). It is clear that any ψ -divergence is nonnegative, and that the functional $(Q_1, \dots, Q_K) \in M^K \mapsto D_\psi(Q_1, \dots, Q_K; P_0) \in [0, +\infty]$ is convex on the product vector space M^K , for all given distribution P_0 . Moreover, if the function $\psi(\cdot)$ is strictly convex on a neighborhood of the vector $\mathbf{1}_K$, then we can prove the following fundamental property

$$D_\psi(Q_1, \dots, Q_K; P_0) \geq 0, \text{ with } D_\psi(Q_1, \dots, Q_K; P_0) = 0 \text{ if and only if } Q_1 = \dots = Q_K = P_0. \quad (2.3)$$

A former different definition of multivariate divergences (between more than two probability measures) have been introduced by Györfi and Nemetz (1978), which requires a common dominating σ -finite measure for all probabilities. These criteria has been used by Zografos (1994), Zografos (1998) and Morales *et al.* (1998), to handle multi-sample problems in estimation and testing, for discrete distributions. We note that the definition of the multivariate divergences of Györfi and Nemetz (1978) is not appropriate to the present context, and it is more adequate for our sake to use the definition (2.2), in particulier, for applying the duality theory on the divergences functionals $(Q_1, \dots, Q_K) \in M^K \mapsto D_\psi(Q_1, \dots, Q_K; P_0)$. The introduction of the above multivariate divergences is motivated by the following problem. Consider the following statistical inference problem for multiple density ratio models (MDRM). Suppose we have at hand $1 + K$ random samples,

$$(X_{0,1}, \dots, X_{0,n_0}), (X_{1,1}, \dots, X_{1,n_1}), \dots, (X_{K,1}, \dots, X_{K,n_K}), \quad (2.4)$$

of $1 + K$ random variables (or vectors) $X_0, X_1, \dots, X_K$, defined on a probability space $(\Omega, \mathcal{A}, P)$ with values in $\mathcal{X} = \mathbb{R}^m, m \geq 1$, and distributions $P_0, Q_1, \dots, Q_K$, respectively. The MDRM postulates that

$$\frac{dQ_k}{dP_0}(x) = r_k(\boldsymbol{\theta}_k, x), \quad \text{for all } k = 1, \dots, K, \quad (2.5)$$

for some known function $r_k(\boldsymbol{\theta}_k, x)$, of $x \in \mathbb{R}^m$, and unknown vector-valued parameters $\boldsymbol{\theta}_k \in \Theta_k \subset \mathbb{R}^{1+d_k}$. We require, for all $k = 1, \dots, K$, the identifiability of the parameter $\boldsymbol{\theta}_k$, in the sense that

$$\text{if } r_k(\boldsymbol{\theta}_k, x) = r_k(\boldsymbol{\theta}'_k, x) \quad \forall x - P_0\text{-a.s.}, \quad \text{then } \boldsymbol{\theta}_k = \boldsymbol{\theta}'_k. \quad (2.6)$$

We denote then by $\boldsymbol{\theta}_T := (\boldsymbol{\theta}_{1T}^\top, \dots, \boldsymbol{\theta}_{KT}^\top)^\top$ the unique true value of the parameter $\boldsymbol{\theta} := (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_K^\top)^\top \in \Theta_1 \times \dots \times \Theta_K$. Typical examples of such MDRM models are the following exponential ones

$$r_k(\boldsymbol{\theta}_k, x) = \exp \left\{ \boldsymbol{\theta}_k^\top \underline{\mathbf{b}}_k(x) \right\}, \quad k = 1, \dots, K, \quad (2.7)$$

for some known vector-valued (basis) function

$$\underline{\mathbf{b}}_k(x) := \left(b_{k,0}(x), \mathbf{b}_k(x)^\top \right)^\top = (b_{k,0}(x), b_{k,1}(x), \dots, b_{k,d_k}(x))^\top \in \mathbb{R}^{1+d_k}, \quad x \in \mathbb{R}^m,$$

with values in $\mathbb{R}^{1+d_k}$, and corresponding unknown vector-valued parameters

$$\boldsymbol{\theta}_k := (\alpha_k, \boldsymbol{\beta}_k^\top)^\top \in \Theta_k \subset \mathbb{R}^{1+d_k}.$$

When the exponential MDRM (2.7) is considered, for technical raisons, we will represent the vector parameter using a different parametrization as $\boldsymbol{\theta}^\top := (\alpha_1, \dots, \alpha_K, \boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_K^\top)$, and will denote $\boldsymbol{\theta}_T^\top := (\alpha_{1T}, \dots, \alpha_{KT}, \boldsymbol{\beta}_{1T}^\top, \dots, \boldsymbol{\beta}_{KT}^\top)$ for the true value of the parameter; Θ denotes the new domain of the parameter $\boldsymbol{\theta}$. Note that the parameter space Θ is a subset of $\mathbb{R}^{K+d}$ with $d := \sum_{k=1}^K d_k$. We require also the first component of $\underline{\mathbf{b}}_k(\cdot)$, for each k , to be the constant function one, i.e., $b_{k,0}(\cdot) := \mathbf{1}_{\mathbb{R}^m}(\cdot)$, the indicator function of $\mathbb{R}^m$. This makes the corresponding first real true value α_{kT} , of the vector $\boldsymbol{\theta}_{kT} := (\alpha_{kT}, \boldsymbol{\beta}_{kT}^\top)^\top$, a normalization parameter, for all $k = 1 \dots, K$. In particular, we have $\alpha_{kT} = 0$,

whenever $\beta_{kT} = \mathbf{0}$. However, the value $\alpha_k \neq \alpha_{kT}$ is assumed to be free, for all $k = 1, \dots, K$. We will assume then, for all $k = 1, \dots, K$, that the parameter space Θ_k is such that its interior $\text{int}(\Theta_k)$ in $\mathbb{R}^{1+d_k}$ is not void. The basis function vector $\mathbf{b}_k(\cdot)$ in (2.7) may depend on k , with different dimensions $d_1, \dots, d_K$. Observe also that, the distribution P_0 is unspecified. Only the density ratios are specified through parametric functions. The components of $\mathbf{b}_k(\cdot)$ in (2.7) may be chosen, for instance, to be the first $1 + d_k$ elements of basis functions. Many other interesting examples of density ratio models, are give in the book by [Sugiyama et al. \(2012a\)](#), for the two-sample case, in particular, in large dimension (i.e., when the dimension m of the variable space is large). Consider the test problem of the null hypothesis

$$\mathcal{H}_0 : Q_1 = \dots = Q_K = P_0 \quad \text{against the alternative} \quad \mathcal{H}_1 : \exists Q_k \neq P_0, \quad (2.8)$$

which is equivalent, in the context of the exponential MDRM, to

$$\mathcal{H}_0 : \beta_{1T} = \dots = \beta_{KT} = \mathbf{0} \quad \text{against the alternative} \quad \mathcal{H}_1 : \exists \beta_{kT} \neq \mathbf{0}, \quad (2.9)$$

since $\beta_{kT} = \mathbf{0}$ implies $\alpha_{kT} = 0$, for all $k = 1, \dots, K$. In view of the basic property (2.3), in order to perform a statistical test of the above hypotheses, we can use an estimate of $D_\psi(Q_1, \dots, Q_K; P_0)$, and reject the null hypothesis $\mathcal{H}_0$ when the obtained estimate takes large value. The plug-in estimate of $D_\psi(Q_1, \dots, Q_K; P_0)$ is given by

$$\tilde{D}_\psi(Q_1, \dots, Q_K; P_0) := D_\psi(\hat{Q}_1, \dots, \hat{Q}_K; \hat{P}_0), \quad (2.10)$$

obtained from the definition (2.2) of the ψ -divergence by replacing the distribution $P_0, Q_1, \dots, Q_K$ by their corresponding empirical measures, namely,

$$\hat{P}_0(\cdot) := \frac{1}{n_0} \sum_{i=1}^{n_0} \delta_{X_{0,i}}(\cdot), \hat{Q}_1(\cdot) := \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{X_{1,i}}(\cdot), \dots, \hat{Q}_K(\cdot) := \frac{1}{n_K} \sum_{i=1}^{n_K} \delta_{X_{K,i}}(\cdot), \quad (2.11)$$

associated to the multiple samples given in (2.4). Unfortunately, the above estimate is not generally well defined, since, the absolute continuity condition is not satisfied (even for large sample sizes, if the distributions are continuous, or discrete with unbounded support). Moreover, the estimator does not make use of the semiparametric form of the MDRM. In the following Section, we will give new dual representations for the ψ -divergences, extending to the present multivariate case the dual representations of divergences between two measures, see [Keziou \(2003a\)](#) and [Broniatowski and Keziou \(2006\)](#), and we will expose how to use the obtained dual representations to give well defined estimates exploiting the form of the MDRM, in the present multiple sample context. We will apply the Fenchel duality theory for the convex functionals $(Q_1, \dots, Q_K) \mapsto D_\psi(Q_1, \dots, Q_K; P_0)$ on well chosen product vector space of signed finite measures, and their corresponding topological dual vector spaces of functions, to be specified in the following Sections. First, we recall the duality result of convex functions on finite dimensional vector space, and then the more general case of possibly infinite dimension.

2.2 Fenchel duality of convex functions on finite dimensional vector space

We refer to [Rockafellar \(1970\)](#) for the proof of the following duality theorem.

Theorem 2.1. *Let $g(\cdot) : \mathbf{x} \in \mathbb{R}^K \mapsto g(\mathbf{x}) \in]-\infty, +\infty]$ be a proper convex function. Denote by $g^*(\cdot)$ its convex conjugate, called also Fenchel-Legendre transform, defined by*

$$g^*(\cdot) : \mathbf{t} \in \mathbb{R}^K \mapsto g^*(\mathbf{t}) := \sup_{\mathbf{x} \in \mathbb{R}^K} \{ \langle \mathbf{t}, \mathbf{x} \rangle - g(\mathbf{x}) \} \in]-\infty, +\infty], \quad (2.12)$$

where $\langle \mathbf{t}, \mathbf{x} \rangle := \mathbf{t}^\top \mathbf{x} := \sum_{i=1}^K t_i x_i$ is the inner product on $\mathbb{R}^K$. Consider the convex conjugate of the proper convex function $g^*(\cdot)$, i.e., the function

$$g^{**}(\cdot) : \mathbf{x} \in \mathbb{R}^K \mapsto g^{**}(\mathbf{x}) := \sup_{\mathbf{t} \in \mathbb{R}^K} \{ \langle \mathbf{t}, \mathbf{x} \rangle - g^*(\mathbf{t}) \} \in]-\infty, +\infty].$$

If $g(\cdot)$ is lower semi-continuous (l.s.c.), then we have for all $\mathbf{x} \in \mathbb{R}^K$, $g(\mathbf{x}) = g^{**}(\mathbf{x})$, namely,

$$g(\mathbf{x}) = \sup_{\mathbf{t} \in \mathbb{R}^K} \{ \langle \mathbf{t}, \mathbf{x} \rangle - g^*(\mathbf{t}) \}, \quad \text{for all } \mathbf{x} \in \mathbb{R}^K, \quad (2.13)$$

where $g^*(\cdot)$ is given by (2.12).

The above theorem means that any proper convex function on $\mathbb{R}^K$, if it is l.s.c., can be represented as the convex conjugate of its convex conjugate. By applying the above theorem, we obtain for the function $\psi(\cdot)$, the representation

$$\psi(\mathbf{x}) = \sup_{\mathbf{t} \in \mathbb{R}^K} \{ \langle \mathbf{x}, \mathbf{t} \rangle - \psi^*(\mathbf{t}) \}, \quad \text{for all } \mathbf{x} \in \mathbb{R}^K, \quad (2.14)$$

where

$$\forall \mathbf{t} := (t_1, \dots, t_K)^\top \in \mathbb{R}^K, \quad \psi^*(\mathbf{t}) := \sup_{\mathbf{x} \in \mathbb{R}^K} \{ \langle \mathbf{t}, \mathbf{x} \rangle - \psi(\mathbf{x}) \}. \quad (2.15)$$

We give in the following lines, some properties of the convex conjugate $\psi^*(\cdot)$ of $\psi(\cdot)$. For the proofs, we can extend, to the present multivariate case, the arguments in Section 26 of Rockafellar (1970). One can show that $\psi^*(\cdot)$, in turn, is convex, closed, and proper, since $\psi^*(\mathbf{t}) > -\infty$, for all $\mathbf{t} \in \mathbb{R}^K$, and $\psi^*(\mathbf{0}) = 0$. It takes its values in $] -\infty, +\infty]$. Moreover, the strict convexity of $\psi(\cdot)$ on its domain dom_ψ is equivalent to the condition that $\psi^*(\cdot)$ is “essentially smooth”, i.e., it satisfies both conditions : differentiability on $\text{int}(\text{dom}_{\psi^*})$ and the condition that, if the boundary points set of dom_{ψ^*} is not void, the directional derivatives, for all boundary point $\mathbf{a}$ of dom_{ψ^*} , satisfy

$$\lim_{\mathbf{y} \rightarrow \mathbf{a}, \mathbf{y} \in \text{dom}_{\psi^*}} \psi^{*\prime}(\mathbf{y}; \mathbf{x} - \mathbf{y}) := \lim_{\mathbf{y} \rightarrow \mathbf{a}, \mathbf{y} \in \text{dom}_{\psi^*}} \lim_{\lambda \downarrow 0} \frac{\psi^*(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) - \psi^*(\mathbf{y})}{\lambda} = +\infty, \forall \mathbf{x} \in \text{dom}_{\psi^*}. \quad (2.16)$$

In (2.16), $\psi^{*\prime}(\mathbf{y}; \mathbf{x} - \mathbf{y})$ denotes the derivative of $\psi^*(\cdot)$ at the point $\mathbf{y}$ in the direction $\mathbf{x} - \mathbf{y}$, which always exists, may be infinite, since $\psi^*(\cdot)$ is convex. Conversely, $\psi(\cdot)$ is essentially smooth if and only if $\psi^*(\cdot)$ is strictly convex on its domain dom_{ψ^*} . Assume that ψ is differentiable on $\text{int}(\text{dom}_\psi)$, and denote by $\psi'_{x_k}(\cdot)$ the partial derivative $\frac{\partial \psi}{\partial x_k}(\cdot)$, for all $k = 1, \dots, K$. Then, we can give the explicit expression of $\psi^*(\mathbf{t})$ for all $\mathbf{t} \in \{(\psi'_{x_1}(\mathbf{x}), \dots, \psi'_{x_K}(\mathbf{x})) \text{ s.t. } \mathbf{x} \in \text{int}(\text{dom}_\psi)\}$. Indeed, one can show that, $\forall \mathbf{x} \in \text{int}(\text{dom}_\psi) \subset \mathbb{R}^K$,

$$\psi^*(\psi'_{x_1}(\mathbf{x}), \dots, \psi'_{x_K}(\mathbf{x})) = \sum_{k=1}^K x_k \psi'_{x_k}(\mathbf{x}) - \psi(\mathbf{x}). \quad (2.17)$$

2.3 Fenchel duality of convex functionals on possibly infinite dimensional vector spaces

We refer to Dembo and Zeitouni (1998), for the proof of the following duality theorem, for convex functionals on more general vector spaces, with possibly infinite dimension.

Theorem 2.2. *Let $\mathcal{S}$ be a locally convex Hausdorff topological vector space (l.c.h.t.v.s.), and $g(\cdot) : \mathcal{S} \mapsto] -\infty, +\infty]$ be a proper convex function. Denote by $g^*(\cdot)$ the convex conjugate of $g(\cdot)$, i.e., the proper convex function on $\mathcal{S}^*$ (the topological dual of $\mathcal{S}$) with values in $] -\infty, +\infty]$, defined by*

$$g^*(\cdot) : \ell \in \mathcal{S}^* \mapsto g^*(\ell) := \sup_{x \in \mathcal{S}} \{ \ell(x) - g(x) \}. \quad (2.18)$$

If $g(\cdot)$ is l.s.c. on $\mathcal{S}$, then we have the following dual representation for $g(\cdot)$

$$g(x) = \sup_{\ell \in \mathcal{S}^*} \{ \ell(x) - g^*(\ell) \}, \quad \forall x \in \mathcal{S}. \quad (2.19)$$

2.4 Dual representations of the multivariate ψ -divergences

We will apply the above theorem in order to obtain dual representations for ψ -divergences. We will look at ψ -divergences as convex functionals $(Q_1, \dots, Q_K) \mapsto D_\psi(Q_1, \dots, Q_K; P_0)$ on the product vector space of K well chosen vector spaces of signed finite measures. Here, P_0 is a fixed reference probability measure. The question is how to choose the vector spaces $\mathcal{S}$ (of signed finite measures) and the topology, on which the ψ -divergence functionals satisfy all conditions of the above duality theorem? At the same time, we need to determine the corresponding topological dual space, in order to get sufficiently explicit representations. We give answers to these questions in the following. Let $\mathcal{F} := \mathcal{F}_1 \times \dots \times \mathcal{F}_K$ be a collection of measurable functions defined on $\mathcal{X}$ with values in $\mathbb{R}^K$. We will denote by $f_1, \dots, f_K$ the K components of any function $\mathbf{f} \in \mathcal{F}$. The spaces $\mathcal{X}$ and $\mathbb{R}^K$ are endowed, respectively, by the σ -field $\mathcal{B}$ and the Borel σ -field $\mathcal{B}(\mathbb{R}^K)$, induced by the usual topology of $\mathbb{R}^K$. Denote by $\mathcal{F}_b$ the real vector space of all bounded measurable functions on $\mathcal{X}$ with values in $\mathbb{R}^K$. Define the real (product) vector space

$$\mathcal{M}_{\mathcal{F}} := \left\{ (Q_1, \dots, Q_K) \in M^K : \int_{\mathcal{X}} |f_k(x)| d|Q_k(x)| < +\infty, \forall k = 1, \dots, K, \forall (f_1, \dots, f_K) \in \mathcal{F} \right\}, \quad (2.20)$$

and denote by $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ the real vector space induced by the collection $\mathcal{F} \cup \mathcal{F}_b$. We will use now the following weak-topology. We equip $\mathcal{M}_{\mathcal{F}}$ with the $\tau_{\mathcal{F}}$ -topology, the weakest one that makes any linear mapping $(Q_1, \dots, Q_K) \in \mathcal{M}_{\mathcal{F}} \mapsto \sum_{k=1}^K \int_{\mathcal{X}} f_k(x) dQ_k(x) \in \mathbb{R}$ continuous, for all $\mathbf{f} \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$. A base of open neighborhoods for any measure vector $(R_1, \dots, R_K) \in \mathcal{M}_{\mathcal{F}}$ is given by

$$N((R_1, \dots, R_K), \mathcal{A}, \epsilon) := \left\{ (Q_1, \dots, Q_K) \in \mathcal{M}_{\mathcal{F}} \text{ s.t. } \max_{\mathbf{f} \in \mathcal{A}} \left| \sum_{k=1}^K \int f_k dR_k - \sum_{k=1}^K \int f_k dQ_k \right| < \epsilon \right\}, \quad (2.21)$$

where $\epsilon > 0$, and $\mathcal{A}$ is a finite collection of functions $\mathbf{f} := (f_1, \dots, f_K)^\top$ in $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$. We refer to Chapter 5 in [Dunford and Schwartz \(1962\)](#) for different weak topologies induced by different classes of functions. Then, we have (the proof is given in [Keziou \(2015\)](#) and [Keziou \(2016\)](#)).

Proposition 2.1. (a) *The real vector space $\mathcal{M}_{\mathcal{F}}$, equipped with the $\tau_{\mathcal{F}}$ -topology, is locally convex Hausdorff topological vector space; and the topological dual of $\mathcal{M}_{\mathcal{F}}$ is given by*

$$\mathcal{M}_{\mathcal{F}}^* = \left\{ (Q_1, \dots, Q_K) \in \mathcal{M}_{\mathcal{F}} \mapsto \sum_{k=1}^K \int_{\mathcal{X}} f_k(x) dQ_k(x); \mathbf{f} \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle \right\} \equiv \langle \mathcal{F} \cup \mathcal{F}_b \rangle;$$

(b) *The proper convex function $(Q_1, \dots, Q_K) \in \mathcal{M}_{\mathcal{F}} \mapsto D_\psi(Q_1, \dots, Q_K; P_0)$ is lower semi-continuous.*

In view of the above proposition, the Fenchel-Legendre transform of the proper convex function $D_\psi(\cdot; P_0) : (Q_1, \dots, Q_K) \in \mathcal{M}_{\mathcal{F}} \mapsto D_\psi(Q_1, \dots, Q_K; P_0)$ is given by

$$\begin{aligned} D_\psi^*(\cdot) : \mathbf{f} \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle &\mapsto D_\psi^*(\mathbf{f}) := \sup_{(Q_1, \dots, Q_K) \in \mathcal{M}_{\mathcal{F}}} \left\{ \sum_{k=1}^K \int_{\mathcal{X}} f_k(x) dQ_k(x) - D_\psi(Q_1, \dots, Q_K; P_0) \right\} \\ &= \sup_{(Q_1, \dots, Q_K) \in \mathcal{M}_{\mathcal{F}}^{P_0}} \left\{ \sum_{k=1}^K \int_{\mathcal{X}} f_k(x) dQ_k(x) - \int_{\mathcal{X}} \psi \left(\frac{dQ_1}{dP_0}(x), \dots, \frac{dQ_K}{dP_0}(x) \right) dP_0(x) \right\} \end{aligned}$$

where, in the last equality, the supremum is restricted to the subset

$$\mathcal{M}_{\mathcal{F}}^{P_0} := \{(Q_1, \dots, Q_K) \in \mathcal{M}_{\mathcal{F}} \text{ s.t. } Q_k \text{ is a.c.w.r.t. } P_0, \forall k = 1, \dots, K\} \quad (2.22)$$

since $D_\psi(Q_1, \dots, Q_K; P_0) = +\infty$ for all $(Q_1, \dots, Q_K) \notin \mathcal{M}_{\mathcal{F}}^{P_0}$. In the following proposition, we give, under some assumptions, explicit formula for the convex conjugate $D_\psi^*(\cdot)$. Moreover, we will prove that $D_\psi^*(\cdot) \equiv D_{\psi^*}(\cdot)$, in the sense that

$$D_\psi^*(\mathbf{f}) = \int_{\mathcal{X}} \psi^*(f_1(x), \dots, f_K(x)) dP_0(x), \quad \forall \mathbf{f} := (f_1, \dots, f_K)^\top \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle.$$

Proposition 2.2. Assume that, $\psi(\cdot)$ is differentiable on the interior of its domain dom_ψ , and denote by $\psi'_{x_k}(\cdot)$, the partial derivatives $\frac{\partial \psi}{\partial x_k}(\cdot), \forall k = 1 \dots, K$. Then, for all $(Q_1, \dots, Q_K) \in \mathcal{M}_{\mathcal{F}}$ such that $D_\psi(Q_1, \dots, Q_K; P_0) < \infty$ and $(\psi'_{x_1}(dQ_1/dP_0, \dots, dQ_K/dP_0), \dots, \psi'_{x_K}(dQ_1/dP_0, \dots, dQ_K/dP_0)) \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle$, the ψ -divergence $D_\psi(Q_1, \dots, Q_K; P_0)$ admits the dual representation

$$D_\psi(Q_1, \dots, Q_K; P_0) = \sup_{\mathbf{f} \in \langle \mathcal{F} \cup \mathcal{F}_b \rangle} \left\{ \sum_{k=1}^K \int_{\mathcal{X}} f_k(x) dQ_k(x) - \int_{\mathcal{X}} \psi^*(f_1(x), \dots, f_K(x)) dP_0(x) \right\}, \quad (2.23)$$

and the function

$$\bar{\mathbf{f}}(x) := \left(\psi'_{x_1} \left(\frac{dQ_1}{dP_0}(x), \dots, \frac{dQ_K}{dP_0}(x) \right), \dots, \psi'_{x_K} \left(\frac{dQ_1}{dP_0}(x), \dots, \frac{dQ_K}{dP_0}(x) \right) \right), x \in \mathcal{X}, \quad (2.24)$$

is a dual optimal solution. Furthermore, if $\psi(\cdot)$ is essentially smooth, then the dual solution $\bar{\mathbf{f}}(\cdot)$ above is unique (P_0 -a.s.).

The above proposition remains valid if the vector space $\langle \mathcal{F} \cup \mathcal{F}_b \rangle$ is restricted to the class of functions $\mathcal{F}$. We state this result in the following theorem.

Theorem 2.3. Assume that, $\psi(\cdot)$ is differentiable on the interior of its domain dom_ψ . Then, for all $(Q_1, \dots, Q_K) \in \mathcal{M}_{\mathcal{F}}$ such that $D_\psi(Q_1, \dots, Q_K; P_0) < \infty$ and $(\psi'_{x_1}(dQ_1/dP_0, \dots, dQ_K/dP_0), \dots, \psi'_{x_K}(dQ_1/dP_0, \dots, dQ_K/dP_0)) \in \mathcal{F}$, the ψ -divergence $D_\psi(Q_1, \dots, Q_K; P_0)$ admits the dual representation

$$D_\psi(Q_1, \dots, Q_K; P_0) = \sup_{\mathbf{f} \in \mathcal{F}} \left\{ \sum_{k=1}^K \int_{\mathcal{X}} f_k(x) dQ_k(x) - \int_{\mathcal{X}} \psi^*(f_1(x), \dots, f_K(x)) dP_0(x) \right\}, \quad (2.25)$$

and the function

$$\bar{\mathbf{f}}(x) := \left(\psi'_{x_1} \left(\frac{dQ_1}{dP_0}(x), \dots, \frac{dQ_K}{dP_0}(x) \right), \dots, \psi'_{x_K} \left(\frac{dQ_1}{dP_0}(x), \dots, \frac{dQ_K}{dP_0}(x) \right) \right), x \in \mathcal{X},$$

is a dual optimal solution. Furthermore, if $\psi(\cdot)$ is essentially smooth, then the dual solution $\bar{\mathbf{f}}(\cdot)$ above is unique (P_0 -a.s.).

Part 2 : New methods in statistical inference

Chapter 3

Parametric Estimation and Tests

We introduce estimation and test procedures through divergence optimization for parametric models not necessary discrete. This approach is based on our dual representations for divergences, presented in Part 1 Chapter 1. We treat point estimation and tests for simple and composite hypotheses, extending maximum likelihood technique. An other view at the maximum likelihood approach, for estimation and test, is given. We prove existence and consistency of the proposed estimates. The limiting distributions of the estimates and test statistics (including the classical likelihood ratio one) are given both under the null and the alternative hypotheses, and approximation of the power functions is deduced. A new procedure of constructing confidence regions, when the parameter may be a boundary value of the parameter space, is proposed. A solution to the irregularity problem of the generalized likelihood ratio test pertaining to the number of components in a mixture is given, and a new test, based on χ^2 -divergence on signed finite measures and duality technique, is proposed. For proofs and simulation results, we refer to [Broniatowski and Keziou \(2009\)](#).

3.1 Introduction and notation

In this Chapter, we consider the problems of estimation and tests using φ -divergences. An i.i.d. sample $X_1, \dots, X_n$, with common unknown distribution P , is observed and some p.m. Q is given. We aim to estimate $D_\varphi(Q, P)$ and, more generally, $\inf_{Q \in \Omega} D_\varphi(Q, P)$, where Ω is some set of measures, as well as the measure Q^* achieving the infimum on Ω . In the parametric context, these problems can be well defined and lead to new results in estimation and tests, extending classical notions.

Statistical examples and motivations

Let Q_0 and P be two p.m.'s with common support S . Introduce a finite partition $A_1, \dots, A_k$ of S (when S is finite this partition is the support of Q_0). The quantization method consists in approximating $D_\varphi(Q_0, P)$ by $\sum_{j=1}^k \varphi\left(\frac{Q_0(A_j)}{P(A_j)}\right) P(A_j)$ which is estimated by $\widetilde{D}_\varphi(Q_0, P) = \sum_{j=1}^k \varphi\left(\frac{Q_0(A_j)}{P_n(A_j)}\right) P_n(A_j)$, where P_n is the empirical measure associated to the data. In this vein, goodness of fit tests have been proposed by [Zografos *et al.* \(1990\)](#) for fixed number of classes, and by [Menéndez *et al.* \(1998\)](#) and [Györfi and Vajda \(2002\)](#) when the number of classes depends on the sample size. We refer to [Pardo \(2006\)](#) which treats these problems extensively and contains many more references.

Let $\{P_\theta; \theta \in \Theta\}$ be some parametric model with Θ a set in $\mathbb{R}^d$. On the basis of an i.i.d. sample $X_1, \dots, X_n$ with distribution P_{θ_T} , we want to estimate θ_T , the unknown true value of the parameter and perform statistical tests on the parameter using φ -divergences. When all p.m.'s P_θ share the same finite support S , [Liese and Vajda \(1987\)](#), [Lindsay \(1994\)](#) and [Morales *et al.* \(1995\)](#) introduced the so-called “Minimum φ -divergences estimates” (M φ DE's), Minimum Disparity Estimators in [Lindsay \(1994\)](#), of the parameter

θ_T , defined by

$$\tilde{\theta}_\varphi := \arg \inf_{\theta \in \Theta} D_\varphi(P_\theta, P_n), \quad (3.1)$$

where $D_\varphi(P_\theta, P_n)$ is the plug-in estimate of $D_\varphi(P_\theta, P_{\theta_T})$ for a given $\theta \in \Theta$. Various parametric tests can be performed based on the previous estimates of φ -divergences; see e.g. Lindsay (1994), Morales *et al.* (1995), Pardo *et al.* (2002) and Menéndez *et al.* (2003). The class of estimates (3.1) contains the maximum likelihood estimate (MLE). Indeed, when $\varphi(x) = \varphi_0(x) = -\log x + x - 1$, we obtain

$$\tilde{\theta}_{KL_m} := \arg \inf_{\theta \in \Theta} KL_m(P_\theta, P_n) = \arg \inf_{\theta \in \Theta} \sum_{j \in S} -\log(P_\theta(j))P_n(j) = MLE. \quad (3.2)$$

The $M\varphi$ DE's (3.1) are motivated by the fact that a suitable choice of the divergence may lead to an estimate more robust than the ML one (see e.g. Lindsay (1994), Basu and Lindsay (1994) and Jiménez and Shao (2001)).

When interested in testing hypotheses $\mathcal{H}_0 : \theta_T = \theta_0$ against alternatives $\mathcal{H}_1 : \theta_T \neq \theta_0$, where θ_0 is a given value, we can use the statistic $D_\varphi(P_{\theta_0}, P_n)$, the plug-in estimate of the φ -divergence between P_{θ_0} and P_{θ_T} , rejecting $\mathcal{H}_0$ for large values of the statistic; see e.g. Cressie and Read (1984). In the case when $\varphi(x) = -\log x + x - 1$, the corresponding test based on $KL_m(P_{\theta_0}, P_n)$ does not coincide with the generalized likelihood ratio one, which is defined through the generalized likelihood ratio (GLR) $\lambda_n := 2 \log \frac{\sup_{\theta \in \Theta} \prod_{i=1}^n p_\theta(X_i)}{\prod_{i=1}^n p_{\theta_0}(X_i)}$. The new estimate $\widehat{KL}_m(P_{\theta_0}, P_{\theta_T})$ of $KL_m(P_{\theta_0}, P_{\theta_T})$, which is proposed in the present Chapter, leads to the generalized likelihood ratio test; see Remark 3.7 below.

When the support S is continuous, the plug-in estimates (3.1) are not well defined; Basu and Lindsay (1994) investigate the so-called “minimum disparity estimators” (MDE’s) for continuous models, through some common kernel smoothing method of P_n and P_θ . When $\varphi(x) = -\log x + x - 1$, this estimate clearly, due to smoothing, does not coincide generally with the ML one. Also, the test based on the associated estimator of the $KL_m(P_{\theta_0}, P_{\theta_T})$ is different from the generalized likelihood ratio one. Further, their estimates poses the problem of the choice of the kernel and the window. For Hellinger distance, we refer to Beran (1977). For nonparametric goodness-of-fit test, Berlinet *et al.* (1998) and Berlinet (1999) proposed a test based on the estimation of the KL_m -divergence using the smoothed kernel estimate of the density. The extension of their results to other divergences remains an open problem; see Berlinet (1999), Györfi *et al.* (1998), and Berlinet *et al.* (1998). All those tests are stated for simple null hypotheses; the case of composite null hypotheses seems difficult to handle by the above technique. In the present Chapter, we treat this problem in the parametric setting.

When the support S is discrete infinite or continuous, then the plug-in estimate $D_\varphi(P_\theta, P_n)$ usually takes infinite value when no use is done of some partition-based approximation. In Broniatowski (2003), a new estimation procedure is proposed in order to estimate the KL -divergence between some set of p.m.’s Ω and some p.m. P , without making use of any partitioning nor smoothing, but merely making use of the well known “dual” representation of the KL -divergence as the Fenchel-Legendre transform of the moment generating function. Extending the paper by Broniatowski (2003), we will use the new dual representations of φ -divergences, see Keziou (2003a) Theorem 2.1 and Broniatowski and Keziou (2006) Theorem 4.4, to define the minimum φ -divergence estimates in both discrete and continuous parametric models. These representations are the starting point for the definition of estimates of the parameter θ_T , which we will call “minimum dual φ -divergence estimates” ($MD\varphi$ DE’s). They are defined in parametric models $\{P_\theta; \theta \in \Theta\}$, where the p.m.’s P_θ do not necessarily have finite support; it can be discrete or continuous, bounded or not. Also the same representations will be applied in order to estimate $D_\varphi(P_{\theta_0}, P_{\theta_T})$ and $\inf_{\theta \in \Theta_0} D_\varphi(P_\theta, P_{\theta_T})$ where θ_0 is a given value in Θ and Θ_0 is a given subset of Θ , which leads to various simple and composite tests pertaining to θ_T . When $\varphi(x) = -\log x + x - 1$, the $MD\varphi$ D estimate coincides with the maximum likelihood one (see Remark 3.2 below); since our approach includes also test procedures, it will be seen that with this peculiar choice for the function φ , we recover the classical likelihood ratio test for simple hypotheses and for composite hypotheses (see Remark 3.7

and Remark 3.10 below). A similar approach has been proposed by Liese and Vajda (2006); see their formula (118). In any case, an exhaustive study of $M\varphi$ DE's seems necessary, in a way that would include both the discrete and the continuous support cases. This is precisely the main scope of this Chapter. The remainder is organized as follows. In Section 2, we recall the dual representations of φ -divergences obtained in Keziou (2003a) Theorem 2.1 and Broniatowski and Keziou (2006) Theorem 2.4. Section 3 presents, through the dual representation of φ -divergences, various estimates and tests in the parametric framework and deals with their asymptotic properties both under the null and the alternative hypotheses. The existence and consistency of the proposed estimates are proved using similar arguments as developed in Qin and Lawless (1994) Lemma 1. We use the limiting laws of the proposed test statistics, in a similar way to Morales and Pardo (2001), to give approximations to the power functions of the tests (including the GLR one). Observe that the power functions of the likelihood ratio type tests are not generally known; one of our contributions is to provide explicit power functions in the general case for simple or composite hypotheses. As a by-product, we obtain the minimal sample size which ensures a given power, for quite general simple or composite hypotheses. In Section 4, we give a solution to the irregularity problem of the GLR test of the number of components in a mixture; we propose a new test based on the χ^2 -divergence on signed finite measures, and a new procedure for constructing confidence regions for the parameter in the case where θ_T may be a boundary value of the parameter space Θ . We sometimes write Pf for $\int f dP$ for any measure P and any function f , when defined. For more details, proofs and many simulation results, we refer to our paper Broniatowski and Keziou (2009).

3.2 Parametric estimation and tests through minimum φ -divergence approach and duality technique

We consider an identifiable parametric model $\{P_\theta; \theta \in \Theta\}$, defined on the Borel σ -field $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$, where Θ is some set in $\mathbb{R}^d$. For brevity, we write $D_\varphi(\theta, \alpha)$ instead of $D_\varphi(P_\theta, P_\alpha)$. We assume that for any θ in Θ , P_θ has density p_θ with respect to some dominating σ -finite measure λ , which can be either with countable support or not. Assume further that the support S of the p.m. P_θ does not depend upon θ . On the basis of an i.i.d. sample $X_1, \dots, X_n$, with distribution P_{θ_T} , we intend to estimate θ_T , the true unknown value of the parameter, which is assumed to be an interior point of the parameter space Θ . We will consider strictly convex essentially smooth functions φ . We will use the following assumption

$$\int_{\mathbb{R}^m} \left| \varphi' \left(\frac{p_\theta(x)}{p_\alpha(x)} \right) \right| dP_\theta(x) < \infty. \quad (3.3)$$

Note that if the function φ satisfies

$$\begin{aligned} &\text{there exists } 0 < \delta < 1 \text{ such that for all } c \text{ in } [1 - \delta, 1 + \delta], \text{ one can find numbers } c_1, c_2, c_3 \text{ such that} \\ &\varphi(cx) \leq c_1 \varphi(x) + c_2 |x| + c_3, \text{ for all real } x, \end{aligned} \quad (3.4)$$

then the assumption (3.3) is satisfied whenever $D_\varphi(\theta, \alpha) < \infty$; see e.g. Broniatowski and Keziou (2006) Lemma 3.2. Also the real convex functions φ_γ (1.5), associated to the class of power divergences, all satisfy the condition (3.4), including all standard divergences. For a given $\theta \in \Theta$, consider the class of functions

$$\mathcal{F} = \mathcal{F}_\theta := \left\{ x \in \mathbb{R}^m \mapsto \varphi' \left(\frac{p_\theta(x)}{p_\alpha(x)} \right) \in \mathbb{R}; \alpha \in \Theta \right\}. \quad (3.5)$$

Applying Theorem 1.11, when assumption (3.3) holds for any $\alpha \in \Theta$, we obtain

$$D_\varphi(\theta, \theta_T) = \sup_{f \in \mathcal{F}_\theta} \left\{ \int_{\mathbb{R}^m} f(x) dP_\theta(x) - \int_{\mathbb{R}^m} \varphi^*(f(x)) dP_{\theta_T}(x) \right\},$$

which, by (1.10), can be written as

$$D_\varphi(\theta, \theta_T) = \sup_{\alpha \in \Theta} \left\{ \int_{\mathbb{R}^m} \varphi' \left(\frac{p_\theta}{p_\alpha} \right) dP_\theta - \int_{\mathbb{R}^m} \left[\frac{p_\theta}{p_\alpha} \varphi' \left(\frac{p_\theta}{p_\alpha} \right) - \varphi \left(\frac{p_\theta}{p_\alpha} \right) \right] dP_{\theta_T} \right\}. \quad (3.6)$$

Furthermore, the supremum in this display is unique and it is achieved at $\alpha = \theta_T$, independently upon the value of θ . Hence, it is reasonable to estimate $D_\varphi(\theta, \theta_T) := \int \varphi(p_\theta/p_{\theta_T}) dP_{\theta_T}$, the φ -divergence between P_θ and P_{θ_T} , by

$$\widehat{D}_\varphi(\theta, \theta_T) := \sup_{\alpha \in \Theta} \left\{ \int_{\mathbb{R}^m} \varphi' \left(\frac{p_\theta}{p_\alpha} \right) dP_\theta - \int_{\mathbb{R}^m} \left[\frac{p_\theta}{p_\alpha} \varphi' \left(\frac{p_\theta}{p_\alpha} \right) - \varphi \left(\frac{p_\theta}{p_\alpha} \right) \right] dP_n \right\}, \quad (3.7)$$

in which we have replaced P_{θ_T} by its estimate P_n , the empirical measure associated to the data. For a given $\theta \in \Theta$, since the supremum in (3.6) is unique and it is achieved at $\alpha = \theta_T$, define the following class of M-estimates of θ_T

$$\widehat{\alpha}_\varphi(\theta) := \arg \sup_{\alpha \in \Theta} \left\{ \int_{\mathbb{R}^m} \varphi' \left(\frac{p_\theta}{p_\alpha} \right) dP_\theta - \int_{\mathbb{R}^m} \left[\frac{p_\theta}{p_\alpha} \varphi' \left(\frac{p_\theta}{p_\alpha} \right) - \varphi \left(\frac{p_\theta}{p_\alpha} \right) \right] dP_n \right\} \quad (3.8)$$

which we call “dual φ -divergence estimates” (D φ DE’s); (in the sequel, for simplicity, we sometimes write $\widehat{\alpha}$ instead of $\widehat{\alpha}_\varphi(\theta)$). Further, we have $\inf_{\theta \in \Theta} D_\varphi(\theta, \theta_T) = D_\varphi(\theta_T, \theta_T) = 0$ and the infimum in this display is unique and it is achieved at $\theta = \theta_T$. It follows that a natural definition of minimum φ -divergence estimates of θ_T , which we will call “minimum dual φ -divergence estimates” (MD φ DE’s), is

$$\widehat{\theta}_\varphi := \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Theta} \left\{ \int_{\mathbb{R}^m} \varphi' \left(\frac{p_\theta}{p_\alpha} \right) dP_\theta - \int_{\mathbb{R}^m} \left[\frac{p_\theta}{p_\alpha} \varphi' \left(\frac{p_\theta}{p_\alpha} \right) - \varphi \left(\frac{p_\theta}{p_\alpha} \right) \right] dP_n \right\}. \quad (3.9)$$

In order to simplify formulas (3.7), (3.8) and (3.9), define the functions

$$g(\theta, \alpha) : x \in \mathbb{R}^m \mapsto g(\theta, \alpha, x) := \frac{p_\theta(x)}{p_\alpha(x)} \varphi' \left(\frac{p_\theta(x)}{p_\alpha(x)} \right) - \varphi \left(\frac{p_\theta(x)}{p_\alpha(x)} \right) \in \mathbb{R}, \quad (3.10)$$

$$f(\theta, \alpha) : x \in \mathbb{R}^m \mapsto f(\theta, \alpha, x) := \varphi' \left(\frac{p_\theta(x)}{p_\alpha(x)} \right) \in \mathbb{R} \quad (3.11)$$

and

$$h(\theta, \alpha) : x \in \mathbb{R}^m \mapsto h(\theta, \alpha, x) := P_\theta f(\theta, \alpha) - g(\theta, \alpha, x) \in \mathbb{R}. \quad (3.12)$$

Hence, (3.7), (3.8) and (3.9) can be written as follows

$$\widehat{D}_\varphi(\theta, \theta_T) := \sup_{\alpha \in \Theta} P_n h(\theta, \alpha), \quad (3.13)$$

$$\widehat{\alpha}_\varphi(\theta) := \arg \sup_{\alpha \in \Theta} P_n h(\theta, \alpha) \quad (3.14)$$

and

$$\widehat{\theta}_\varphi := \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Theta} P_n h(\theta, \alpha). \quad (3.15)$$

Formula (3.6) can be written then as

$$D_\varphi(\theta, \theta_T) = \sup_{\alpha \in \Theta} P_{\theta_T} h(\theta, \alpha). \quad (3.16)$$

If the supremum in (3.14) is not unique, we define the estimate $\widehat{\alpha}_\varphi(\theta)$ as any value of $\alpha \in \Theta$ that maximizes the function $\alpha \in \Theta \mapsto P_n h(\theta, \alpha)$. Also, if the infimum in (3.15) is not unique, the estimate $\widehat{\theta}_\varphi$ is defined to be any value of $\theta \in \Theta$ that minimizes the function $\theta \mapsto \sup_{\alpha \in \Theta} P_n h(\theta, \alpha)$. Conditions assuring the existences of the above estimates are given in Sections 3.1 and 3.2 below.

Remark 3.1. For the L_1 distance, i.e., when $\varphi(x) = |x - 1|$, formula (3.6) does not apply since the corresponding φ function is not differentiable. However, using the general dual representation of divergences given in [Keziou \(2003a\)](#) Proposition 2.2 and [Broniatowski and Keziou \(2006\)](#) Theorem 4.1, we can obtain an explicit formula for L_1 distance avoiding the differentiability assumption. A methodology on estimation and testing in L_1 distance has been proposed by [Devroye and Lugosi \(2001\)](#), and its consequences for composite hypothesis testing and for model selection based density estimates for nested classes of densities are presented in [Devroye et al. \(2002\)](#) and [Biau and Devroye \(2005\)](#).

Remark 3.2. (*An other view at the ML estimate*). The maximum likelihood estimate belongs to both classes of estimates (3.14) and (3.15). Indeed, it is obtained when $\varphi(x) = -\log x + x - 1$, that is as the dual modified KL-divergence estimate or as the minimum dual modified KL-divergence estimate, i.e., $MLE = D(KL_m)DE = MD(KL_m)DE$. Indeed, we have $P_\theta f(\theta, \alpha) = 0$ and $P_n h(\theta, \alpha) = -\int \log\left(\frac{p_\theta}{p_\alpha}\right) dP_n$. Hence by definitions (3.8) and (3.9), we get

$$\hat{\alpha}_{KL_m}(\theta) = \arg \sup_{\alpha \in \Theta} - \int \log\left(\frac{p_\theta}{p_\alpha}\right) dP_n = \arg \sup_{\alpha \in \Theta} \int \log(p_\alpha) dP_n = MLE$$

independently upon θ , and

$$\hat{\theta}_{KL_m} = \arg \inf_{\theta \in \Theta} \sup_{\alpha \in \Theta} - \int \log\left(\frac{p_\theta}{p_\alpha}\right) dP_n = \arg \sup_{\theta \in \Theta} \int \log(p_\theta) dP_n = MLE.$$

So, the MLE can be seen as the estimate of θ_T that minimizes the estimate of the KL_m -divergence between the parametric model $\{P_\theta; \theta \in \Theta\}$ and the p.m. P_{θ_T} .

3.2.1 The asymptotic properties of $\hat{\alpha}_\varphi(\theta)$ and $\widehat{D}_\varphi(\theta, \theta_T)$ for a given θ in Θ

This Section deals with the asymptotic properties of the estimates (3.13) and (3.14). We will use similar arguments as those presented in van der Vaart (1998) Section 5.2 and 5.6 under classical conditions, for the study of M-estimates. In the sequel, we assume that condition (3.3) holds for any $\alpha \in \Theta$, and use $\|\cdot\|$ to denote the Euclidean norm in $\mathbb{R}^d$. Consider the following conditions

- (c.1) The estimate $\hat{\alpha}_\varphi(\theta)$ exists;
- (c.2) $\sup_{\alpha \in \Theta} |P_n h(\theta, \alpha) - P_{\theta_T} h(\theta, \alpha)|$ converges to zero a.s. (resp. in probability);
- (c.3) for any positive ϵ , there exists some positive η such that for all $\alpha \in \Theta$ satisfying $\|\alpha - \theta_T\| > \epsilon$ we have $P_{\theta_T} h(\theta, \alpha) < P_{\theta_T} h(\theta, \theta_T) - \eta$.

Remark 3.3. Condition (c.1) is fulfilled for example if the function $\alpha \in \Theta \mapsto P_n h(\theta, \alpha)$ is continuous and Θ is compact. Condition (c.2) is satisfied if $\{x \mapsto h(\theta, \alpha, x); \alpha \in \Theta\}$ is a Glivenko-Cantelli class of functions. Condition (c.3) means that the maximizer $\alpha = \theta_T$ of the function $\alpha \mapsto P_{\theta_T} h(\theta, \alpha)$ is well-separated. This condition holds, for example, when the function $\alpha \in \Theta \mapsto P_{\theta_T} h(\theta, \alpha)$ is strictly concave and Θ is convex, which is the case for the following two examples.

Example 3.1. Consider the case $\varphi(x) = -\log x + x - 1$ and the normal model $\{\mathcal{N}(\alpha, 1); \alpha \in \Theta = \mathbb{R}\}$. Hence, we obtain

$$P_{\theta_T} h(\theta, \alpha) = \frac{1}{2}(\theta - \theta_T)^2 - \frac{1}{2}(\alpha - \theta_T)^2. \quad (3.17)$$

We see that condition (c.3) is satisfied; we can choose $\eta = \frac{\epsilon^2}{2}$.

Example 3.2. Consider the case $\varphi(x) = -\log x + x - 1$ and the exponential model $\{p_\alpha(x) = \alpha \exp(-\alpha x); \alpha \in \Theta = \mathbb{R}_+^*\}$. Hence, we obtain

$$P_{\theta_T} h(\theta, \alpha) = -\log \theta + \frac{\theta}{\theta_T} + \log \alpha - \frac{\alpha}{\theta_T}, \quad (3.18)$$

which is strictly concave (in α). Hence, condition (c.3) is satisfied.

Proposition 3.1. (a) Under assumption (c.1-2), the estimate $\widehat{D}_\varphi(\theta, \theta_T)$ converges a.s. (resp. in probability) to $D_\varphi(\theta, \theta_T)$.

(b) Assume that the assumptions (c.1-2-3) hold. Then the estimate $\hat{\alpha}_\varphi(\theta)$ converges in probability to θ_T .

Assume that θ_T is an interior point of Θ , the convex function φ has continuous derivatives up to 4th order, and the density $p_\alpha(x)$ has continuous partial derivatives up to 3th order (for all x λ -a.e.). Denote I_{θ_T} the Fisher information matrix $I_{\theta_T} := \int_{\mathbb{R}^m} \frac{p'_{\theta_T}(x)p'_{\theta_T}(x)^\top}{p_{\theta_T}(x)} d\lambda(x)$. In the following theorem, we give the limiting laws of the estimates $\hat{\alpha}_\varphi(\theta)$ and $\widehat{D}_\varphi(\theta, \theta_T)$. We will use the following assumptions.

- (A.0) The estimate $\hat{\alpha}_\varphi(\theta)$ exists and is consistent;
- (A.1) There exists a neighborhood $N(\theta_T)$ of θ_T such that the first and second order partial derivatives (w.r.t α) of $f(\theta, \alpha, x)p_\theta(x)$ are dominated on $N(\theta_T)$ by some λ -integrable functions. The third order partial derivatives (w.r.t α) of $h(\theta, \alpha, x)$ are dominated on $N(\theta_T)$ by some P_{θ_T} -integrable functions;
- (A.2) The integrals $P_{\theta_T} \|(\partial/\partial\alpha)h(\theta, \theta_T)\|^2$ and $P_{\theta_T} \|(\partial^2/\partial\alpha^2)h(\theta, \theta_T)\|$ are finite, and the matrix $P_{\theta_T}(\partial^2/\partial\alpha^2)h(\theta, \theta_T)$ is non singular;
- (A.3) The integral $P_{\theta_T}h(\theta, \theta_T)^2$ is finite.

Theorem 3.1. *Assume that assumptions (A.0-1-2) hold. Then, we have*

- (a) $\sqrt{n}(\hat{\alpha}_\varphi(\theta) - \theta_T)$ converges in distribution to a centered multivariate normal random variable with covariance matrix

$$V_\varphi(\theta, \theta_T) = S^{-1}MS^{-1} \quad (3.19)$$

with $S := -P_{\theta_T}(\partial^2/\partial\alpha^2)h(\theta, \theta_T)$ and $M := P_{\theta_T}(\partial/\partial\alpha)h(\theta, \theta_T)(\partial/\partial\alpha)^\top h(\theta, \theta_T)$.
If $\theta_T = \theta$, then $V_\varphi(\theta, \theta_T) = V(\theta_T) = I_{\theta_T}^{-1}$.

- (b) If $\theta_T = \theta$, then the statistic $\frac{2n}{\varphi''(1)}\widehat{D}_\varphi(\theta, \theta_T)$ converges in distribution to a χ^2 random variable with d degrees of freedom.
- (c) If additionally assumption (A.3) holds, then when $\theta \neq \theta_T$, we have $\sqrt{n}(\widehat{D}_\varphi(\theta, \theta_T) - D_\varphi(\theta, \theta_T))$ converges in distribution to a centered normal random variable with variance

$$\sigma_\varphi^2(\theta, \theta_T) = P_{\theta_T}h(\theta, \theta_T)^2 - (P_{\theta_T}h(\theta, \theta_T))^2. \quad (3.20)$$

Remark 3.4. Our first result (Proposition 3.1 above) provides a general solution for the consistency of the global maximum (3.14) under strong but usual conditions, also difficult to be checked; see van der Vaart (1998) Chapter 5. Moreover, in practice, the optimization in (3.14) is handled through gradient descent algorithms, depending on some initial guess $\alpha_0 \in \Theta$, which may provide a local maximum (not necessarily global) of $P_n h(\theta, \cdot)$. Hence, it is desirable to prove that in a “neighborhood” of θ_T there exists a maximum of $P_n h(\theta, \cdot)$ which indeed converges to θ_T ; this is the scope of Theorem 3.2, in the following subsection, which states that for some “good” α_0 (near θ_T) the algorithm provides a consistent estimate. It is well known that, in various classical models, the global maximizer of the likelihood function may not exist or be inconsistent. Typical examples are provided in mixture models. Consider the Beta-mixture model, given in Ferguson (1982) Section 3, $p_\theta(x) = \theta g(x|1, 1) + (1 - \theta)g(x|\gamma(\theta), \beta(\theta))$, where $\Theta = [1/2, 1]$, $g(x|\gamma(\theta), \beta(\theta))$ is the $Be(\gamma, \beta)$ -density and $\gamma(\theta) = \theta\delta(\theta)$ and $\beta(\theta) = (1 - \theta)\delta(\theta)$ with $\delta(\theta) \rightarrow +\infty$ sufficiently fast as $\theta \rightarrow 1$. The ML estimate converges to 1 (a.s.) whatever the value of θ_T in Θ ; see Ferguson (1982) Section 3 for the proof. However, if we take for example $\theta_T = 3/4$, Theorem 3.2 hereafter proves the existence and consistency of a sequence of local maximizers under weak assumptions which hold for this example. Other motivations for the results of Theorem 3.2 are given in Remark 3.5 below.

Existence, consistency and limiting laws of a sequence of local maxima

We use similar arguments as developed in [Qin and Lawless \(1994\)](#) Lemma 1. Assume that θ_T is an interior point of Θ , the convex function φ has continuous derivatives up to 4th order, and the density $p_\alpha(x)$ has continuous partial derivatives up to 3th order (for all $x \lambda - a.e$). In the following theorem, we state the existence and the consistency of a sequence of local maxima $\tilde{\alpha}_\varphi(\theta)$ and $\widetilde{D}_\varphi(\theta, \theta_T)$. We give also their limiting laws.

Theorem 3.2. *Assume that assumptions (A.1) and (A.2) hold. Then, we have*

(a) *Let $B(\theta_T, n^{-1/3}) := \{\alpha \in \Theta; \|\alpha - \theta_T\| \leq n^{-1/3}\}$. Then, as $n \rightarrow \infty$, with probability one, the function $\alpha \mapsto P_n h(\theta, \alpha)$ attains its maximum value at some point $\tilde{\alpha}_\varphi(\theta)$ in the interior of the ball B , and satisfies $P_n(\partial/\partial\alpha)h(\theta, \tilde{\alpha}_\varphi(\theta)) = 0$.*

(b) *$\sqrt{n}(\tilde{\alpha}_\varphi(\theta) - \theta_T)$ converges in distribution to a centered multivariate normal random variable with covariance matrix*

$$V_\varphi(\theta, \theta_T) = S^{-1}MS^{-1}. \quad (3.21)$$

(c) *If $\theta_T = \theta$, then the statistic $\frac{2n}{\varphi''(1)}\widetilde{D}_\varphi(\theta, \theta_T)$ converges in distribution to a χ^2 random variable with d degrees of freedom.*

(d) *If additionally assumption (A.3) holds, then when $\theta \neq \theta_T$, we have*

$\sqrt{n}(\widetilde{D}_\varphi(\theta, \theta_T) - D_\varphi(\theta, \theta_T))$ converges in distribution to a centered normal random variable with variance $\sigma_\varphi^2(\theta, \theta_T)$.

Remark 3.5. *The results of this theorem are motivated by the following statements*

- *The estimates $\tilde{\alpha}_\varphi(\theta)$ can be calculated if the statistician disposes of some preknowledge of the true unknown parameter θ_T .*
- *The hypotheses are satisfied for a large class of parametric models for which the support does not depend upon θ , such normal, log normal, exponential, Gamma, Beta, Weibull, ... etc; see for example [van der Vaart \(1998\)](#) Paragraph 5.43.*
- *The maps $h(\theta, \alpha) : x \mapsto h(\theta, \alpha, x)$ and $(\theta, \alpha) \mapsto P_{\theta_T}h(\theta, \alpha)$ are allowed to take the value $-\infty$; for example, take $\varphi(x) = -\log x + x - 1$, and consider the model*

$$\{P_\alpha = \alpha \text{Cauchy}(0) + (1 - \alpha)\mathcal{N}(0, 1); \alpha \in \Theta\},$$

with $\Theta = [0, 1]$ and $\theta_T = 1/2$. Then, $P_{\theta_T}h(\theta, 1) = -\infty$ for all $\theta \in]0, 1[$.

- *The theorem states both existence, consistency and asymptotic normality of the estimates.*
- *The estimate $\tilde{\alpha}_\varphi(\theta)$ may exist and be consistent whereas $\hat{\alpha}_\varphi(\theta)$ does not in many cases.*
- *One interesting situation also is if the map $\alpha \in \Theta \mapsto P_n h(\theta, \alpha) = 0$ is strictly concave and Θ is convex; the estimates $\tilde{\alpha}_\varphi(\theta)$ and $\hat{\alpha}_\varphi(\theta)$ are the same.*

Remark 3.6. *Using Theorem 3.1 part (c), the estimate $\widehat{D}_\varphi(\theta_0, \theta_T)$ can be used to perform statistical tests (asymptotically of level ϵ) of the null hypothesis $\mathcal{H}_0 : \theta_T = \theta_0$ against the alternative $\mathcal{H}_1 : \theta_T \neq \theta_0$ for a given value θ_0 . Since $D_\varphi(\theta_0, \theta_T)$ is nonnegative and takes value zero only when $\theta_T = \theta_0$, the tests are defined through the critical region*

$$C_\varphi(\theta_0, \theta_T) := \left\{ \frac{2n}{\varphi''(1)}\widehat{D}_\varphi(\theta_0, \theta_T) > \chi_d^2(1 - \epsilon) \right\} \quad (3.22)$$

where $\chi_d^2(1 - \epsilon)$ is the $(1 - \epsilon)$ -quantile of the χ^2 -distribution with d degrees of freedom. Note that these tests are all consistent, since $\widehat{D}_\varphi(\theta_0, \theta_T)$ are n -consistent estimates of $D_\varphi(\theta_0, \theta_T) = 0$ under $\mathcal{H}_0$, and

$\sqrt{n}$ -consistent estimate of $D_\varphi(\theta_0, \theta_T) > 0$ under $\mathcal{H}_1$; see part (b) and (c) in Theorem 3.1 above. Further, the asymptotic result (c) in Theorem 3.1 can be used to give approximation of the power function $\theta_T \mapsto \beta(\theta_T) := P_{\theta_T}(C_\varphi(\theta_0, \theta_T))$. We obtain then the following approximation

$$\beta(\theta_T) \approx 1 - F_N \left(\frac{\sqrt{n}}{\sigma_\varphi(\theta_0, \theta_T)} \left[\frac{\varphi''(1)}{2n} \chi_d^2(1 - \epsilon) - D_\varphi(\theta_0, \theta_T) \right] \right) \quad (3.23)$$

where F_N is the cumulative distribution function of a normal random variable with mean zero and variance one. An important application of this approximation is the approximate sample size (3.24) below that ensures a power β for a given alternative $\theta_T \neq \theta_0$. Let n_0 be the positive root of the equation

$$\beta = 1 - F_N \left(\frac{\sqrt{n}}{\sigma_\varphi(\theta_0, \theta_T)} \left[\frac{\varphi''(1)}{2n} \chi_d^2(1 - \epsilon) - D_\varphi(\theta_0, \theta_T) \right] \right)$$

i.e., $n_0 = \frac{(a+b) - \sqrt{a(a+2b)}}{2D_\varphi(\theta_0, \theta_T)^2}$ where $a = \sigma_\varphi^2(\theta_0, \theta_T) [F_N^{-1}(1 - \beta)]^2$ and $b = \varphi''(1) \chi_d^2(1 - \epsilon) D_\varphi(\theta_0, \theta_T)$. The required sample size is then

$$n^* = [n_0] + 1 \quad (3.24)$$

where $[.]$ is used here to denote “integer part of”.

Remark 3.7. (An other view at the generalized likelihood ratio test and approximation of the power function through KL_m -divergence). In the particular case of the KL_m -divergence, i.e., when $\varphi(x) = \varphi_0(x) := -\log x + x - 1$, we obtain from (3.22) the critical area

$$C_{KL_m}(\theta_0, \theta_T) := \left\{ 2n \sup_{\alpha \in \Theta} P_n \log \left(\frac{p_\alpha}{p_{\theta_0}} \right) > \chi_d^2(1 - \epsilon) \right\} = \left\{ 2 \log \frac{\sup_{\alpha \in \Theta} \prod_{i=1}^n p_\alpha(X_i)}{\prod_{i=1}^n p_{\theta_0}(X_i)} > \chi_d^2(1 - \epsilon) \right\},$$

which is to say that the test obtained in this case is precisely the generalized likelihood ratio one. The power approximation and the approximate sample size guaranteeing a power β for a given alternative (for the GLRT) are given by (3.23) and (3.24), respectively, where φ is replaced by φ_0 and D_φ by KL_m .

3.2.2 The asymptotic behavior of the MD_φ DE's

We now explore the asymptotic properties of the estimates $\hat{\theta}_\varphi$ and $\hat{\alpha}_\varphi(\hat{\theta}_\varphi)$ defined in (3.15) and (3.14). We assume that condition (3.3) holds for any $\alpha, \theta \in \Theta$. We state consistency under the following assumptions

(c.4) The estimates $\hat{\theta}_\varphi$ and $\hat{\alpha}_\varphi(\hat{\theta}_\varphi)$ exist.

(c.5) $\sup_{\{\alpha, \theta \in \Theta\}} |P_n h(\theta, \alpha) - P_{\theta_T} h(\theta, \alpha)|$ tends to 0 in probability;

(a) for any positive ϵ , there exists some positive η , such that for any α in Θ with $\|\alpha - \theta_T\| > \epsilon$ and for all $\theta \in \Theta$, it holds $P_{\theta_T} h(\theta, \alpha) < P_{\theta_T} h(\theta, \theta_T) - \eta$;

(b) there exists a neighborhood of θ_T , say $N(\theta_T)$, such that for any positive ϵ , there exists some positive η such that for all $\alpha \in N(\theta_T)$ and all $\theta \in \Theta$ satisfying $\|\theta - \theta_T\| > \epsilon$, it holds $P_{\theta_T} h(\theta_T, \alpha) < P_{\theta_T} h(\theta, \alpha) - \eta$;

(c.6) there exists some neighborhood $N(\theta_T)$ of θ_T and a positive function H such that for all α in $N(\theta_T)$, $\|h(\theta_T, \alpha, x)\| \leq H(x)$ (P_{θ_T} -a.s.) with $P_{\theta_T} H < \infty$.

Remark 3.8. Condition (c.5) is fulfilled if $\{x \mapsto h(\theta, \alpha); (\theta, \alpha) \in \Theta^2\}$ is a Glivenko-Cantelli class of functions. Conditions (c.5.a) and (c.5.b) mean that the saddle-point (θ_T, θ_T) , of $(\theta, \alpha) \in \Theta \times \Theta \mapsto P_n h(\theta, \alpha)$, is well-separated. Note that theses two conditions are not very restrictive, they are satisfied for example when Θ is convex and the function $(\theta, \alpha) \in \Theta \times \Theta \mapsto P_n h(\theta, \alpha)$ is concave in α (for all θ) and convex in θ (for all α), which is the case for Examples 3.1 and 3.2 above, both conditions (c.5.a) and (c.5.b) are satisfied; we can take $\eta = \frac{\epsilon^2}{2}$.

Proposition 3.2. Assume that conditions (c.4-5-6) hold. Then,

(a) $\sup_{\theta \in \Theta} \|\hat{\alpha}_\varphi(\theta) - \theta_T\|$ tends to 0 in probability.

(b) The MD φ estimate $\hat{\theta}_\varphi$ converges to θ_T in probability.

Assume that θ_T is an interior point of Θ , the convex function φ has continuous derivatives up to 4th order, and the density $p_\theta(x)$ has continuous partial derivatives up to 3th order (for all x λ -a.e.). In the following theorem we state the asymptotic normality of the estimates $\hat{\theta}_\varphi$ and $\hat{\alpha}_\varphi(\hat{\theta}_\varphi)$. We will use the following assumptions

- (A.4) The estimates $\hat{\theta}_\varphi$ and $\hat{\alpha}_\varphi(\hat{\theta}_\varphi)$ exist and are consistent;
- (A.5) There exists a neighborhood $N(\theta_T)$ of θ_T such that the first and second order partial derivatives (w.r.t. α and θ) of $f(\theta, \alpha, x)p_\theta(x)$ are dominated on $N(\theta_T) \times N(\theta_T)$ by λ -integrable functions. The third partial derivatives (w.r.t. α and θ) of $h(\theta, \alpha, x)$ are dominated on $N(\theta_T) \times N(\theta_T)$ by some P_{θ_T} -integrable functions;
- (A.6) The integrals $P_{\theta_T} \|(\partial/\partial\alpha)h(\theta_T, \theta_T)\|^2$, $P_{\theta_T} \|(\partial/\partial\theta)h(\theta_T, \theta_T)\|^2$, $P_{\theta_T} \|(\partial^2/\partial\alpha^2)h(\theta_T, \theta_T)\|$, $P_{\theta_T} \|(\partial^2/\partial\theta^2)h(\theta_T, \theta_T)\|$ and $P_{\theta_T} \|(\partial^2/\partial\theta\partial\alpha)h(\theta_T, \theta_T)\|$ are finite, and the matrix I_{θ_T} is non singular.

Theorem 3.3. Assume that conditions (A.4-5-6) hold. Then, both $\sqrt{n}(\hat{\theta}_\varphi - \theta_T)$ and $\sqrt{n}(\hat{\alpha}_\varphi(\hat{\theta}_\varphi) - \theta_T)$ converge in distribution to a centered multivariate normal random variable with covariance matrix $V = I_{\theta_T}^{-1}$.

Existence, consistency and limiting laws of a sequence of local minima-maxima

Assume that θ_T is an interior point of Θ , the convex function φ has continuous derivatives up to 4th order, and the density $p_\theta(x)$ has continuous partial derivatives up to 3th order (for all x λ -a.e.). In the following theorem we state the existence and consistency of a sequence of local minima-maxima $\tilde{\theta}_\varphi$ and $\tilde{\alpha}_\varphi(\tilde{\theta}_\varphi)$. We give also their limiting laws.

Theorem 3.4. Assume that conditions (A.5) and (A.6) hold.

- (a) Let $B := \{\theta \in \Theta; \|\theta - \theta_T\| \leq n^{-1/3}\}$. Then, as $n \rightarrow \infty$, with probability one, the function $(\theta, \alpha) \mapsto P_n h(\theta, \alpha)$ attains its min-max value at some point $(\tilde{\theta}_\varphi, \tilde{\alpha}_\varphi(\tilde{\theta}_\varphi))$ in the interior of $B \times B$, and $P_n(\partial/\partial\alpha)h(\tilde{\theta}_\varphi, \tilde{\alpha}_\varphi(\tilde{\theta}_\varphi)) = 0$ and $P_n(\partial/\partial\theta)h(\tilde{\theta}_\varphi, \tilde{\alpha}_\varphi(\tilde{\theta}_\varphi)) = 0$;
- (b) Both $\sqrt{n}(\tilde{\theta}_\varphi - \theta_T)$ and $\sqrt{n}(\tilde{\alpha}_\varphi(\tilde{\theta}_\varphi) - \theta_T)$ converge in distribution to a centered multivariate normal random variable with covariance matrix $V = I_{\theta_T}^{-1}$.

3.2.3 Composite tests by minimum φ -divergence

Let Θ_0 be a subset of Θ . We assume that there exists an open set $B_0 \subset \mathbb{R}^{d-l}$ and mappings $r : \Theta \rightarrow \mathbb{R}^l$ and $s : B_0 \rightarrow \mathbb{R}^d$ such that the matrices $R(\theta) := \left[\frac{\partial}{\partial\theta_i} r(\theta) \right]$ and $S(\beta) := \left[\frac{\partial}{\partial\beta_i} s(\beta) \right]$ exist, with elements continuous, and are of rank l and $(d-l)$, respectively, $\Theta_0 = \{s(\beta); \beta \in B_0\}$ and $r(\theta) = 0$ for all $\theta \in \Theta_0$. Consider the composite null hypothesis

$$\mathcal{H}_0 : \theta_T \in \Theta_0 \text{ versus } \mathcal{H}_1 : \theta_T \in \Theta \setminus \Theta_0. \quad (3.25)$$

This is equivalent to $\mathcal{H}_0 : \theta_T \in s(B_0)$ versus $\mathcal{H}_1 : \theta_T \in \Theta \setminus s(B_0)$. Using (3.16), the φ -divergence $D_\varphi(\Theta_0, \theta_T)$, between the set of distributions $\{P_\theta$ such that $\theta \in \Theta_0\}$ and the p.m. P_{θ_T} , can be written as $D_\varphi(\Theta_0, \theta_T) = \inf_{\theta \in \Theta_0} \sup_{\alpha \in \Theta} P_{\theta_T} h(\theta, \alpha)$. Hence, it can be estimated by

$$\widehat{D}_\varphi(\Theta_0, \theta_T) := \inf_{\theta \in \Theta_0} \widehat{D}_\varphi(\theta, \theta_T) := \inf_{\theta \in \Theta_0} \sup_{\alpha \in \Theta} P_n h(\theta, \alpha).$$

We use $\widehat{D}_\varphi(\Theta_0, \theta_T)$ to perform statistical test pertaining to (3.25). Since $D_\varphi(\Theta_0, \theta_T) := \inf_{\theta \in \Theta_0} D_\varphi(\theta, \theta_T)$ is positive under $\mathcal{H}_1$ and takes value 0 only under $\mathcal{H}_0$ (provided that the infimum is attained on Θ_0), we reject $\mathcal{H}_0$ whenever $\widehat{D}_\varphi(\Theta_0, \theta_T)$ takes large values. The following theorem provides the limit distribution of $\widehat{D}_\varphi(\Theta_0, \theta_T)$ under the null hypothesis $\mathcal{H}_0$.

Theorem 3.5. *Let us assume that the conditions in Theorem 3.3 are satisfied. Under $\mathcal{H}_0$, the statistics $\frac{2n}{\varphi''(1)} \widehat{D}_\varphi(\Theta_0, \theta_T)$ converge in distribution to a χ^2 -random variable with l degrees of freedom.*

The following theorem gives the limit laws of the test statistics $\frac{2n}{\varphi''(1)} \widehat{D}_\varphi(\Theta_0, \theta_T)$ under the alternative hypothesis $\mathcal{H}_1 : \theta_T \in \Theta \setminus \Theta_0$. We will use the following assumptions.

- (C.1) The minimum of $\theta \mapsto D_\varphi(\theta, \theta_T)$ on Θ_0 is attained at some point, say $\theta^* := s(\beta^*)$ with $\beta^* \in B_0$; uniqueness then follows by strict convexity of φ and model identifiability assumption;
- (C.2) There exists a neighborhood $N(\beta^*)$ of β^* and a neighborhood $N(\theta_T)$ of θ_T such that the first and second order partial derivatives (w.r.t. α and β) of $f(s(\beta), \alpha, x)p_{s(\beta)}(x)$ are dominated on $N(\beta^*) \times N(\theta_T)$ by λ -integrable functions. The third partial derivatives (w.r.t. β and α) of $h(s(\beta), \alpha, x)$ are dominated on $N(\beta^*) \times N(\theta_T)$ by some P_{θ_T} -integrable functions;
- (C.3) The integrals $P_{\theta_T} \|(\partial/\partial\alpha)h(s(\beta^*), \theta_T)\|^2$, $P_{\theta_T} \|(\partial/\partial\beta)h(s(\beta^*), \theta_T)\|^2$, $P_{\theta_T} \|(\partial^2/\partial\alpha^2)h(s(\beta^*), \theta_T)\|$, $P_{\theta_T} \|(\partial^2/\partial\beta^2)h(s(\beta^*), \theta_T)\|$ and $P_{\theta_T} \|(\partial^2/\partial\beta\partial\alpha)h(s(\beta^*), \theta_T)\|$ are finite, and the matrix

$$A := \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

is non singular, where $A_{11} := P_{\theta_T}(\partial^2/\partial\beta^2)h(s(\beta^*), \theta_T)$, $A_{22} := P_{\theta_T}(\partial^2/\partial\alpha^2)h(s(\beta^*), \theta_T)$ and $A_{12} = A_{21}^T := P_{\theta_T}(\partial^2/\partial\beta\partial\alpha)h(s(\beta^*), \theta_T)$.

- (C.4) The integral $P_{\theta_T} \|h(s(\beta^*), \theta_T)\|^2$ is finite.

Denote $\widehat{\beta}_\varphi$ and $\widehat{\alpha}_\varphi(\widehat{\beta}_\varphi)$ the min-max optimal solution of $\widehat{D}_\varphi(\Theta_0, \theta_T) := \inf_{\beta \in B_0} \sup_{\alpha \in \Theta} P_n h(s(\beta), \alpha)$, and let $B(\beta^*, n^{-1/3}) := \{\beta \in B_0; \|\beta - \beta^*\| \leq n^{-1/3}\}$, $c_n := (\widehat{\beta}_\varphi^\top, \widehat{\alpha}_\varphi(\widehat{\beta}_\varphi)^\top)^\top$, $c^* := (\beta^{*T}, \theta_T^\top)^\top$ and F the matrix defined by

$$F := P_{\theta_T} \begin{bmatrix} (\partial/\partial\beta)h(s(\beta^*), \theta_T) \\ (\partial/\partial\alpha)h(s(\beta^*), \theta_T) \end{bmatrix} \begin{bmatrix} (\partial/\partial\beta)h(s(\beta^*), \theta_T) \\ (\partial/\partial\alpha)h(s(\beta^*), \theta_T) \end{bmatrix}^\top.$$

- (C.5) The estimates $\widehat{\beta}_\varphi$ and $\widehat{\alpha}_\varphi(\widehat{\beta}_\varphi)$ exist and are consistent estimators for β^* and θ_T respectively.

Theorem 3.6. *Assume that conditions (C.1-2-3-4-5) hold. Then, under the alternative hypothesis $\mathcal{H}_1$, we have*

- (a) $\sqrt{n}(c_n - c^*)$ converges in distribution to a centered multivariate normal random variable with covariance matrix $V = A^{-1}FA^{-1}$;
- (b) If additionally the condition (C.6) holds, then $\sqrt{n}(\widehat{D}_\varphi(\Theta_0, \theta_T) - D_\varphi(\Theta_0, \theta_T))$ converges in distribution to a centered normal random variable with variance

$$\sigma_\varphi^2(\beta^*, \theta_T) = P_{\theta_T} h(s(\beta^*), \theta_T)^2 - (P_{\theta_T} h(s(\beta^*), \theta_T))^2. \quad (3.26)$$

Remark 3.9. *Using Theorem 3.5, the estimate $\widehat{D}_\varphi(\Theta_0, \theta_T)$ can be used to perform statistical tests (asymptotically of level ϵ) of the null hypothesis $\mathcal{H}_0 : \theta_T \in \Theta_0$ against the alternative $\mathcal{H}_1 : \theta_T \in \Theta \setminus \Theta_0$. Since $D_\varphi(\Theta_0, \theta_T)$ is nonnegative and takes value zero only when $\theta_T \in \Theta_0$, the tests are defined through the critical region*

$$C_\varphi(\Theta_0, \theta_T) := \left\{ \frac{2n}{\varphi''(1)} \widehat{D}_\varphi(\Theta_0, \theta_T) > \chi_l^2(1 - \epsilon) \right\}, \quad (3.27)$$

where $\chi_l^2(1 - \epsilon)$ is the $(1 - \epsilon)$ -quantile of the χ^2 -distribution with l degrees of freedom. Note that these tests are all consistent, since $\widehat{D}_\varphi(\Theta_0, \theta_T)$ are n -consistent estimates of $D_\varphi(\Theta_0, \theta_T) = 0$ under $\mathcal{H}_0$, and $\sqrt{n}$ -consistent estimate of $D_\varphi(\Theta_0, \theta_T) > 0$ under $\mathcal{H}_1$; see Theorem 3.5 and Theorem 3.6 part (c). Further, the asymptotic result (c) in theorem 3.6 above can be used to give an approximation to the power function $\theta_T \mapsto \beta(\theta_T) := P_{\theta_T}(C_\varphi(\Theta_0, \theta_T))$. We obtain then the following approximation

$$\beta(\theta_T) \approx 1 - F_N \left(\frac{\sqrt{n}}{\sigma_\varphi(\beta^*, \theta_T)} \left[\frac{\varphi''(1)}{2n} \chi_l^2(1 - \epsilon) - D_\varphi(\Theta_0, \theta_T) \right] \right) \quad (3.28)$$

where F_N is the cumulative distribution function of a normal variable with mean zero and variance one. An important application of this approximation is the approximate sample size (3.29) below that ensures a power β for a given alternative $\theta_T \in \Theta \setminus \Theta_0$. Let n_0 be the positive root of the equation

$$\beta = 1 - F_N \left(\frac{\sqrt{n}}{\sigma_\varphi(\beta^*, \theta_T)} \left[\frac{\varphi''(1)}{2n} \chi_l^2(1 - \epsilon) - D_\varphi(\Theta_0, \theta_T) \right] \right)$$

i.e., $n_0 = \frac{(a+b) - \sqrt{a(a+2b)}}{2D_\varphi(\Theta_0, \theta_T)^2}$ where $a = \sigma_\varphi^2(\beta^*, \theta_T) [F_N^{-1}(1 - \beta)]^2$ and $b = \varphi''(1) \chi_l^2(1 - \epsilon) D_\varphi(\Theta_0, \theta_T)$. The required sample size is then

$$n^* = [n_0] + 1 \quad (3.29)$$

where $[.]$ is used here to denote “integer part of”.

Remark 3.10. (An other view at the generalized likelihood ratio test for composite hypotheses, and approximation of the power function through KL_m -divergence). In the particular case of the KL_m -divergence, i.e., when $\varphi(x) = \varphi_0(x) := -\log x + x - 1$, we obtain from (3.27) the critical area

$$C_{KL_m}(\Theta_0, \theta_T) = \left\{ 2 \log \frac{\sup_{\alpha \in \Theta} \prod_{i=1}^n p_\alpha(X_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n p_\theta(X_i)} > \chi_l^2(1 - \epsilon) \right\},$$

which is to say that the test obtained in this case is precisely the generalized likelihood ratio test associated to (3.25). The power approximation and the approximate sample size guaranteeing a power β for a given alternative (for the GLRT) are given by (3.28) and (3.29), respectively, where φ is replaced by φ_0 and D_φ by KL_m .

3.3 Non regular models. A simple solution for mixture models

The test problem for the number of components of a finite mixture has been extensively treated when the total number of components k is equal to 2, leading to a satisfactory solution; the limit distribution of the generalized likelihood ratio statistic is non standard, since it is $0.5\delta_0 + 0.5\chi_1^2$, a mixture of a Dirac mass at 0 and a χ_1^2 with weights equal to 1/2; see e.g. Titterington *et al.* (1985) and Self and Liang (1987). When $k > 2$, the problem is much more involved. Self and Liang (1987) obtained the limit distribution of the generalized likelihood ratio statistic, which is non standard and complex. This result yields formidable numerical difficulties for the calculation of the critical value of the test. In Section 5.1 below, we propose a unified treatment for all these cases, with simple and standard limit distribution both when the parameter θ_T is an interior or a boundary point of the parameter space Θ . On the other hand, confidence regions for the mixture parameter θ_T even when $k = 2$ are intractable through the generalized likelihood ratio statistic. Indeed, the asymptotic law of the generalized likelihood ratio statistic depends heavily on the fact that θ is a boundary or an interior point of the parameter space. For example, when $k = 2$, the limiting distribution of the generalized likelihood ratio statistic is $0.5\delta_0 + 0.5\chi_1^2$ when $\theta = 0$ and χ_1^2 when $0 < \theta < 1$. Therefore, the confidence level is not defined uniquely. At the opposite, we will prove in Section 5.3 that the proposed dual χ^2 -statistic yields quite standard confidence regions even when $k > 2$. Let $\{P_{a_1}^{(1)}; a_1 \in A_1\}, \dots, \{P_{a_k}^{(k)}; a_k \in A_k\}$ be k -parametric models where $A_1, \dots, A_k$ are k ($k \geq 2$) sets in $\mathbb{R}^{d_1}, \dots, \mathbb{R}^{d_k}$ and $d_1, \dots, d_k \in \mathbb{N}^*$. Denote P_θ the mixture model

$$P_\theta := \sum_{i=1}^k w_i P_{a_i}^{(i)} \quad (3.30)$$

where $0 \leq w_i \leq 1$, $\sum w_i = 1$ and

$$\theta \in \Theta := \left\{ (w_1, \dots, w_k, a_1, \dots, a_k)^\top \in [0, 1]^k \times A_1 \times \dots \times A_k \text{ such that } \sum_{i=1}^k w_i = 1 \right\}, \quad (3.31)$$

and assume that the model is identifiable. Let $k_0 \in \{1, \dots, k-1\}$. We test if $(k - k_0)$ components in (3.30) have null coefficients. We assume that their labels are $k_0 + 1, \dots, k$. Denote Θ_0 the subset of Θ defined by $\Theta_0 := \{\theta \in \Theta \text{ such that } w_{k_0+1} = \dots = w_k = 0\}$. On the basis of an i.i.d sample $X_1, \dots, X_n$ with distribution P_{θ_T} , $\theta_T \in \Theta$, we intend to perform tests of the hypothesis

$$\mathcal{H}_0 : \theta_T \in \Theta_0 \text{ against the alternative } \mathcal{H}_1 : \theta_T \in \Theta \setminus \Theta_0. \quad (3.32)$$

It is known that the generalized likelihood ratio test, based on the statistic

$$2 \log \lambda := 2 \log \frac{\sup_{\theta \in \Theta} \prod_{i=1}^n p_\theta(X_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n p_\theta(X_i)}, \quad (3.33)$$

is not valid for this problem, since the asymptotic approximation by χ^2 -distribution does not hold in this case; the problem is due to the fact that the null value of θ_T is not in the interior of the parameter space Θ . We clarify now this problem. For simplicity, consider a mixture of two known densities p_0 and p_1 with $p_0 \neq p_1$:

$$p_\theta = (1 - \theta)p_0 + \theta p_1 \text{ where } \theta \in \Theta := [0, 1]. \quad (3.34)$$

Given i.i.d. data $X_1, \dots, X_n$, with distribution P_{θ_T} , $\theta_T \in [0, 1]$, consider the test problem

$$\mathcal{H}_0 : \theta_T = 0 \text{ against the alternative } \mathcal{H}_1 : \theta_T > 0. \quad (3.35)$$

The generalized likelihood ratio statistic for this test problem is $W_n(0) := 2 \log \frac{L(\hat{\theta})}{L(0)}$, where $L(\theta) := \prod_{i=1}^n [(1 - \theta)p_0(X_i) + \theta p_1(X_i)]$ for all $\theta \in [0, 1]$, and $\hat{\theta}$ is the MLE of θ . Using the strict concavity of the function $\theta \in [0, 1] \mapsto l(\theta) := \log L(\theta)$, it is clear that $\hat{\theta} = 0$ whenever $l'_+(0)$, the derivative on the right at $\theta = 0$ of $\theta \mapsto l(\theta)$, is nonpositive. Hence, we can write

$$\begin{aligned} P_0 \{W_n = 0\} &\geq P_0 \{\hat{\theta} = 0\} = P_0 \{l'_+(0) \leq 0\} = P_0 \left\{ \sum_{i=1}^n \frac{p_0(X_i)}{p_1(X_i)} - n \leq 0 \right\} \\ &= P_0 \left\{ \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \frac{p_0(X_i)}{p_1(X_i)} - 1 \right) \leq 0 \right\} \end{aligned} \quad (3.36)$$

which, by the CLT, tends to $1/2$ (if $1 \neq E(Y_i^2) < \infty$ where $Y_i := p_0(X_i)/p_1(X_i)$) since the random variables Y_i are i.i.d with $E(Y_i) = 1$ under $\mathcal{H}_0$. This proves that the convergence in distribution of the generalized likelihood ratio statistic $W_n(0)$ to a χ^2 random variable (under $\mathcal{H}_0$) does not hold. Under suitable regularity conditions we can prove that the limit distribution of the statistic W_n in (3.3) is $0.5\delta_0 + 0.5\chi_1^2$, a mixture of the χ^2 -distribution and the Dirac measure at zero; see Self and Liang (1987). Moreover, in the case of more than two components and $k - k_0 \geq 2$, the asymptotic distribution of the GLR statistic (3.33) under $\mathcal{H}_0$ is complicate and not standard (not a χ^2 distribution) which poses some difficulty in determining the critical value that will give correct asymptotic size; see Self and Liang (1987). On the other hand, the likelihood ratio statistic

$$W_n(\theta) := 2 \log \frac{L(\hat{\theta})}{L(\theta)} \quad (3.37)$$

can not be used to construct asymptotic confidence region for the parameter θ_T since its limiting law is not the same when $\theta_T = 0$ and $\theta_T > 0$.

3.3.1 A simple solution to the problem of testing the number of components in a mixture

We propose the following simple solution. Consider the following set of signed finite measures

$$p_\theta := (1 - \theta)p_0 + \theta p_1, \text{ where } \theta \in \mathbb{R}. \quad (3.38)$$

This set (of signed finite measures with mass one) obviously contains the mixture model (3.34). In particular, the null value of θ_T (i.e., $\theta_T = 0$) is an interior point of the parameter space $\mathbb{R}$. The likelihood ratio test (for a model of signed measures) cannot be used since the log-likelihood $l(\theta)$ may be infinite (when $\theta < 0$ or $\theta > 1$). In the context of divergences, this means that the estimate $\widehat{KL}_m(P_0, P_{\theta_T})$ may be infinite if we consider the model (3.38), which is due to the fact that the corresponding convex function $\varphi(x) = -\log x + x - 1$ is infinite on $\mathbb{R}_-$. This suggests to use a divergence associated to a convex function φ which is finite on all $\mathbb{R}$, for instance, the χ^2 -divergence (which is associated to the convex function $\varphi(x) = \frac{1}{2}(x - 1)^2$). So, in order to perform a test asymptotically of level ϵ for (3.35), we propose to use the following estimate of the χ^2 -divergence between P_0 and P_{θ_T}

$$\widetilde{\chi^2}(0, \theta_T) = \sup_{\alpha \in \Theta_e} \{P_0 f(0, \alpha) - P_n g(0, \alpha)\}, \quad (3.39)$$

where $f(0, \alpha) = p_0/p_\alpha - 1$ and $g(0, \alpha) = 1/2(p_0/p_\alpha + 1)(p_0/p_\alpha - 1)$ as a consequence of definitions (9.25) and (3.10), and Θ_e is the new parameter space which we define as follows

$$\Theta_e := \left\{ \alpha \in \mathbb{R} \text{ such that } \int |f(0, \alpha, x)| dP_0(x) \text{ is finite} \right\}.$$

The value of the parameter θ_T under the null hypothesis $\mathcal{H}_0$, i.e., $\theta_T = 0$, is in the interior of the new parameter space Θ_e which is generally non void. Hence, under conditions of Theorem 3.1 where Θ is replaced by Θ_e and θ by zero, under $\mathcal{H}_0$ the statistic $2n\widetilde{\chi^2}(0, \theta_T)$ converges in distribution to a χ^2 random variable with one degree of freedom; the critical region takes then the form

$$CR := \left\{ 2n\widetilde{\chi^2}(0, \theta_T) > \chi_1^2(1 - \epsilon) \right\}, \quad (3.40)$$

where $\chi_1^2(1 - \epsilon)$ is the $(1 - \epsilon)$ -quantile of the χ^2 -distribution with one degree of freedom. Obviously other divergences which are associated to convex functions finite on all $\mathbb{R}$ can be used. The use of the χ^2 -divergence is recommended. Indeed, for regular cases (for example for multinomial goodness-of-fit tests) χ^2 -test is equivalent (in Pitman sense) to the generalized likelihood ratio one; see also Cressie and Read (1984) Sections 3.1 and 3.2 for other motivations in favor of the χ^2 criterion.

3.3.2 Confidence regions for the mixture parameters

We propose the following solution to the confidence region problem when the parameter may be a boundary value of the parameter space. The estimate

$$\widetilde{\chi^2}(\theta, \theta_T) = \sup_{\alpha \in \Theta_e(\theta)} \{P_\theta f(\theta, \alpha) - P_n g(\theta, \alpha)\}, \quad (3.41)$$

where $\Theta_e(\theta) := \left\{ \alpha \in \mathbb{R} \text{ such that } \int |f(\theta, \alpha)| dP_\theta \text{ is finite} \right\}$, can be used to construct asymptotic confidence region for the parameter θ_T with level $(1 - \epsilon)$ defined by

$C := \left\{ \theta \in \Theta \text{ such that } 2n\widetilde{\chi^2}(\theta, \theta_T) \leq \chi_1^2(1 - \epsilon) \right\}$. In fact, $\lim_{n \rightarrow \infty} P_{\theta_T}(\theta_T \in C) = 1 - \epsilon$ both when $\theta_T = 0$ or $\theta_T > 0$ since the statistic $2n\widetilde{\chi^2}(\theta_T, \theta_T)$ converges in distribution to χ^2 -random variable with one degree of freedom both when $\theta_T = 0$ or $\theta_T > 0$. We give now the form of the critical region and the confidence region in the multivariate case, i.e., in the case of the general model (3.30). For all $\theta \in \Theta$, define the set $\Theta_e(\theta) := \left\{ \alpha \in \mathbb{R}^k \times A_1 \times \cdots \times A_k \text{ such that } \sum_{i=1}^k \alpha_i = 1 \text{ and } \int |f(\theta, \alpha, x)| dP_\theta(x) \text{ is finite} \right\}$,

and the statistic $\widetilde{\chi}^2(\Theta_0, \theta_T) := \inf_{\theta \in \Theta_0} \widetilde{\chi}^2(\theta, \theta_T) := \inf_{\theta \in \Theta_0} \sup_{\alpha \in \Theta_e(\theta)} \{P_\theta f(\theta, \alpha) - P_n g(\theta, \alpha)\}$. Under some conditions similar to that in Theorems 3.1, 3.2 and 3.3, we can prove, under the null hypothesis $\mathcal{H}_0$ in (3.32), that the statistic $2n\widetilde{\chi}^2(\Theta_0, \theta_T)$ converges in distribution to χ^2 random variable with $(k - k_0)$ degrees of freedom. Also, the statistic $2n\widetilde{\chi}^2(\theta, \theta_T)$ when $\theta = \theta_T$ converges in distribution to χ^2 -random variable with $d := k - 1 + d_1 + \dots + d_k$ degrees of freedom in both case when θ_T is a boundary value or not. Hence, the critical region is given by $CR := \left\{2n\widetilde{\chi}^2(\Theta_0, \theta_T) > \chi_{k-k_0}^2(1 - \epsilon)\right\}$, and $C := \left\{\theta \in \Theta \text{ such that } 2n\widetilde{\chi}^2(\theta, \theta_T) \leq \chi_d^2(1 - \epsilon)\right\}$ is an asymptotic confidence region for θ_T of level ϵ both when θ_T is a boundary value or not.

3.4 Concluding remarks and possible developments

We have addressed parametric estimation and test problems. We have introduced new estimation and test procedures, using the obtained conjugate expressions of φ -divergences, for general parametric models, not necessarily discrete, avoiding the smoothing approach. The procedure leads to optimal estimates for the parameter model and for the divergences. It includes both the discrete (finite or infinite) and the continuous support cases. It extends the maximum likelihood method for both estimation and test problems. Moreover, the procedure and the divergences framework allow to obtain the limiting laws of the proposed estimates and the test statistics both under the null and the alternative (simple or composite) hypotheses, including the generalized likelihood ratio statistic. As a by-product, we obtain explicit power functions in a general case for simple or composite parametric test problems, and approximations of the minimal sample size which guarantees a desired power for a given alternative. A new test and new asymptotic confidence regions are proposed in the case where the parameter may be a boundary value of the parameter space. Many problems remain to be studied in the future, such as the choice of the divergence which leads to “optimal” (in some sense) estimate or test in terms of efficiency and robustness, Bartlett correctability and the large deviation properties of the proposed test statistics.

Chapter 4

Estimation and test under moment condition models

This Chapter presents our paper [Broniatowski and Keziou \(2012\)](#). We introduce estimation and test procedures through divergence minimization for models satisfying linear constraints with unknown parameter. These procedures extend the empirical likelihood (EL) method and share common features with generalized empirical likelihood approach. Using our previous results in [Broniatowski and Keziou \(2006\)](#), see Chapter 1, we give a precise characterization of duality, for the proposed class of estimates and test statistics, which is used to derive their limiting distributions (including the EL estimate and the EL ratio statistic) both under the null hypotheses and under alternatives or misspecification. An approximation to the power function is deduced as well as the sample size which ensures a desired power for a given alternative. For more detail, proofs and simulation results, we refer to [Broniatowski and Keziou \(2012\)](#).

4.1 Introduction and notation

Statistical models are often defined through estimating equations $\mathbb{E}[g(X, \theta)] = 0$, where $\mathbb{E}[\cdot]$ denotes the mathematical expectation, $g := (g_1, \dots, g_l)^\top \in \mathbb{R}^l$ is some specified vector valued function of a random vector $X \in \mathbb{R}^m$ and a parameter vector $\theta \in \Theta \subset \mathbb{R}^d$. Examples of such models are numerous, see e.g. [Qin and Lawless \(1994\)](#), [Haberman \(1984\)](#), [Sheehy \(1987\)](#), [McCullagh and Nelder \(1983\)](#), [Owen \(2001\)](#) and the references therein. Denoting M^1 the collection of all probability measures (p.m.) on the measurable space $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$, the submodel $\mathcal{M}_\theta^1$, associated to a given value $\theta \in \Theta$, consists of all distributions Q satisfying l linear constraints induced by the vector valued function $g(\cdot, \theta)$, namely, $\mathcal{M}_\theta^1 := \{Q \in M^1 \text{ such that } \int_{\mathbb{R}^m} g(x, \theta) dQ(x) = 0\}$, with $l \geq d$. The statistical model which we consider can be written as

$$\mathcal{M}^1 := \bigcup_{\theta \in \Theta} \mathcal{M}_\theta^1. \quad (4.1)$$

Let $X_1, \dots, X_n$ denote an i.i.d sample of X with unknown distribution P_0 . We denote θ_T , if it exists, the value of the parameter such that P_0 belongs to $\mathcal{M}_{\theta_T}^1$, namely, the value satisfying $\mathbb{E}[g(X, \theta_T)] = 0$, and we assume obviously that θ_T is unique. This work addresses the two following natural questions.

Problem 1: Does P_0 belong to the model $\mathcal{M}^1$?

Problem 2: When P_0 is in the model, which is the value θ_T of the parameter for which $\mathbb{E}[g(X, \theta_T)] = 0$? Also can we perform tests about θ_T ? Can we construct confidence areas for θ_T ?

We note that these problems have been investigated by many authors. [Hansen \(1982\)](#) considered generalized method of moments (GMM). [Hansen et al. \(1996\)](#) introduced the continuous updating (CU) estimate. The empirical likelihood (EL) approach, developed by [Owen \(1988\)](#) and [Owen \(1990\)](#), has been investigated in the context of model (4.1) by [Qin and Lawless \(1994\)](#) and [Imbens \(1997\)](#) introduc-

ing the EL estimate. The recent literature in econometrics focusses on such models; [Smith \(1997\)](#), [Newey and Smith \(2004\)](#) provided a class of estimates called generalized empirical likelihood (GEL) estimates which contains the EL and the CU ones. [Schennach \(2007\)](#) discussed the asymptotic properties of the empirical likelihood estimate under misspecification; the author showed the important fact that the EL estimate may cease to be root n consistent when both the functions g_j defining the moments conditions and the support of P_0 are unbounded. Among other results pertaining to EL, [Newey and Smith \(2004\)](#) stated that EL estimate enjoys optimality properties in term of efficiency when bias corrected among all GEL estimates including the GMM one. Moreover, [Corcoran \(1998\)](#) and [Baggerly \(1998\)](#) proved that in a class of minimum discrepancy statistics (called power divergence statistics), EL ratio is the only one that is Bartlett correctable. Confidence areas for the parameter θ_T have been considered in the seminal paper by [Owen \(1990\)](#). Problems 1 and 2 have been handled via EL and GEL approaches in [Qin and Lawless \(1994\)](#), [Smith \(1997\)](#) and [Newey and Smith \(2004\)](#) under the null hypothesis $\mathcal{H}_0 : P_0 \in \mathcal{M}^1$; the limiting distributions of the GEL estimates and the GEL test statistics have been obtained under the model and under the null hypotheses. [Imbens \(1997\)](#) discusses the asymptotic properties of the EL and exponential tilting estimates under misspecification and give the formula of the asymptotic variance, using dual characterizations, without presenting the hypotheses under which their results hold. [Chen et al. \(2007\)](#) give the limiting distribution of the EL estimate under misspecification as well as the EL ratio statistic between a parametric model and a moment condition model. The paper by [Kitamura \(2007\)](#) gives a discussion of duality for GEL estimates under moment condition models. [Bertail \(2006\)](#) uses duality to study, under the model, the asymptotic properties of the EL ratio statistic and its Bartlett correctability; the author extends the results to semiparametric problems with infinite-dimensional parameters.

The main contribution of the present Chapter is the precise characterization of duality for a large class of estimates and test statistics (including GEL and EL ones) and its use in deriving the limiting properties of both the estimates and the test statistics under misspecification and under alternative hypotheses. Moreover,

- 1) The approach which we develop is based on *minimum discrepancy estimates*, which extends the EL method and has common features with minimum distance and GEL techniques, using merely divergences. We present a wide class of estimates, test statistics and confidence regions for the parameter θ_T as well as various test statistics for *Problems 1 and 2*, all depending on the choice of the divergence;
- 2) The limiting distribution of the EL test statistic under the alternative and under misspecification remains up to date an open problem. The present work fills this gap; indeed, we give the limiting distributions of the proposed estimates and test statistics (including the EL ones) both under the null hypotheses, under alternatives and under misspecification;
- 3) The limiting distributions of the test statistics under the alternatives and misspecification are used to give an approximation to the power function and the sample size which ensures a desired power for a given alternative;
- 4) We extend confidence region (C.R.) estimation techniques based on EL (see [Owen \(1990\)](#)), providing a wide range of such C.R.'s, each one depending upon a specific divergence.

From the point of view of the statistical criterion under consideration, the main advantage, of using a divergence based approach and duality, lays in the fact that it leads to asymptotic properties of the estimates and test statistics under the alternative, including misspecification, which cannot be achieved through the classical EL context. In the case of parametric models of densities, [White \(1982\)](#) studied the asymptotic properties of the parametric maximum likelihood estimate and the parametric likelihood ratio statistic under misspecification; [Broniatowski and Keziou \(2009\)](#) stated the consistency and obtained the limiting distributions of the minimum divergence estimates and the corresponding test statistics (including the parametric likelihood ones) both under the null hypotheses and the alternatives, from which

they deduced an approximation to the power function. In this work, we extend the above results for the proposed class of estimates and test statistics (including the EL ones) in the context of semiparametric models (4.1).

The rest is organized as follows. Section 4.2 describes the statistical divergences used in the sequel. Section 4.3 is devoted to the description of the proposed estimation and test procedures. In Section 4.4, we adapt the Lagrangian duality formalism to the context of statistical divergence, and we use it to give practical formulas (for the study and the numerical computation) of the proposed estimates and test statistics. Section 4.5 deals with the asymptotic properties of the estimates and the test statistics under the model and under misspecification. For proofs and simulations results, we refer to our paper Broniatowski and Keziou (2012) Section 6.

4.2 Statistical divergences

We first set some general definitions and notations. Let P be some p.m. on the measurable space $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$. Denote by M the space of all signed finite measures (s.f.m.) on $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$. Let φ be a convex function from $\mathbb{R}$ onto $[0, +\infty]$ with $\varphi(1) = 0$, and s.t. its domain, $\text{dom}_\varphi := \{x \in \mathbb{R} \text{ s.t. } \varphi(x) < \infty\} = (a_\varphi, b_\varphi)$, is an interval, with endpoints $a_\varphi < 1 < b_\varphi$, which may be bounded or unbounded, open or not. We assume that φ is proper and closed¹. For any s.f.m. $Q \in M$, the φ -divergence between Q and the p.m. P , when Q is absolutely continuous with respect to (a.c.w.r.t) P , is defined through

$$D_\varphi(Q, P) := \int_{\mathbb{R}^m} \varphi\left(\frac{dQ}{dP}(x)\right) dP(x), \quad (4.2)$$

in which $\frac{dQ}{dP}(\cdot)$ denotes the Radon-Nikodym derivative. When Q is not a.c.w.r.t. P , we set $D_\varphi(Q, P) := +\infty$. For any p.m. P , the mapping $Q \in M \mapsto D_\varphi(Q, P)$ is convex and takes nonnegative values. When $Q = P$ then $D_\varphi(Q, P) = 0$. Furthermore, if the function $x \mapsto \varphi(x)$ is strictly convex on a neighborhood of $x = 1$, then

$$D_\varphi(Q, P) = 0 \text{ if and only if } Q = P. \quad (4.3)$$

4.3 Minimum divergence estimates

Let $X_1, \dots, X_n$ denote an i.i.d. sample of a random vector $X \in \mathbb{R}^m$ with distribution P_0 . Let P_n be the empirical measure pertaining to this sample, namely, $P_n(\cdot) := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(\cdot)$, where $\delta_x(\cdot)$ denotes the Dirac measure at point x , for all x . We will endow our statistical approach in the global context of s.f.m.'s with total mass 1 satisfying l linear constraints

$$\mathcal{M}_\theta := \left\{ Q \in M \text{ such that } \int_{\mathbb{R}^m} dQ(x) = 1 \text{ and } \int_{\mathbb{R}^m} g(x, \theta) dQ(x) = 0 \right\} \quad (4.4)$$

and

$$\mathcal{M} := \bigcup_{\theta \in \Theta} \mathcal{M}_\theta, \quad (4.5)$$

sets of signed finite measures that replace $\mathcal{M}_\theta^1$ and $\mathcal{M}^1$. Enhancing the model (4.1) to the above one (4.5) bears a number of improvements upon existing results; this is argued at the end of the present Section; see also Remark 4.5 below. The “plug-in” estimate of $D_\varphi(\mathcal{M}_\theta, P_0)$ is

$$\widehat{D}_\varphi(\mathcal{M}_\theta, P_0) := \inf_{Q \in \mathcal{M}_\theta} D_\varphi(Q, P_n) = \inf_{Q \in \mathcal{M}_\theta} \int_{\mathbb{R}^m} \varphi\left(\frac{dQ}{dP_n}(x)\right) dP_n(x). \quad (4.6)$$

¹The closedness of φ means that if a or b are finite then $\varphi(x) \rightarrow \varphi(a)$ when $x \downarrow a$, and $\varphi(x) \rightarrow \varphi(b)$ when $x \uparrow b$. Note that, this is equivalent to the fact that the level sets $\{x \in \mathbb{R}; \varphi(x) \leq \alpha\}$, $\forall \alpha \in \mathbb{R}$, are closed in $\mathbb{R}$ endowed with the usual topology.

If the projection $Q_\theta^{(n)}$ of P_n on $\mathcal{M}_\theta$ exists, then it is clear that $Q_\theta^{(n)}$ is a s.f.m. (or possibly a p.m.) a.c.w.r.t. P_n ; this means that the support of $Q_\theta^{(n)}$ must be included in the set $\{X_1, \dots, X_n\}$. So, define the sets

$$\mathcal{M}_\theta^{(n)} := \left\{ Q \in \mathcal{M} \mid Q \text{ a.c.w.r.t. } P_n, \sum_{i=1}^n Q(X_i) = 1 \text{ and } \sum_{i=1}^n Q(X_i)g(X_i, \theta) = 0 \right\}, \quad (4.7)$$

which may be seen as subsets of $\mathbb{R}^n$. Then, the plug-in estimate (4.6) can be written as

$$\widehat{D}_\varphi(\mathcal{M}_\theta, P_0) = \inf_{Q \in \mathcal{M}_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^n \varphi(nQ(X_i)). \quad (4.8)$$

In the same way, $D_\varphi(\mathcal{M}, P_0) := \inf_{\theta \in \Theta} \inf_{Q \in \mathcal{M}_\theta} D_\varphi(Q, P_0)$ can be estimated by

$$\widehat{D}_\varphi(\mathcal{M}, P_0) := \inf_{\theta \in \Theta} \inf_{Q \in \mathcal{M}_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^n \varphi(nQ(X_i)). \quad (4.9)$$

By uniqueness of $\arg \inf_{\theta \in \Theta} D_\varphi(\mathcal{M}_\theta, P_0)$ and since the infimum is reached at $\theta = \theta_T$ under the model, we estimate θ_T through

$$\widehat{\theta}_\varphi := \arg \inf_{\theta \in \Theta} \inf_{Q \in \mathcal{M}_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^n \varphi(nQ(X_i)). \quad (4.10)$$

Enhancing $\mathcal{M}^1$ to $\mathcal{M}$ and accordingly extensions in the definitions of the φ functions on $\mathbb{R}$ and the φ -divergences on the whole space of s.f.m.'s $\mathcal{M}$, is motivated by the following arguments:

- If the domain (a_φ, b_φ) of the function φ is included in $[0, +\infty[$ then minimizing over $\mathcal{M}^1$ or over $\mathcal{M}$ leads to the same estimates and test statistics. It is the case of the KL_m , KL , modified χ^2 and Hellinger divergences;
- Let θ be a given value in Θ . Denote $Q_\theta^{(1,n)}$ and $Q_\theta^{(n)}$, respectively, the projection of P_n on $\mathcal{M}_\theta^1$ and on $\mathcal{M}_\theta$. If $Q_\theta^{(1,n)}$ satisfies $0 < Q_\theta^{(1,n)}(X_i) < 1$, for all $i = 1, \dots, n$, then $Q_\theta^{(1,n)} = Q_\theta^{(n)}$. Therefore, in this case, both approaches leads also to the same estimates and test statistics;
- It may occur that for some θ in Θ and some $i = 1, \dots, n$, $Q_\theta^{(1,n)}(X_i)$ is a boundary value of $[0, 1]$, hence the first order conditions are not met which makes a real difficulty for the calculation of the estimates over the sets of p.m. $\mathcal{M}_\theta^1$ and $\mathcal{M}^1$. However, when $\mathcal{M}^1$ is replaced by $\mathcal{M}$, then this problem does not hold any longer in particular when $\text{dom}_\varphi = \mathbb{R}$, which is the case for the χ^2 -divergence. Other arguments are given in Remark 4.5 below.

The empirical likelihood paradigm (see Owen (1988), Owen (1990), Qin and Lawless (1994) and Owen (2001)), enters as a special case of the statistical issues related to estimation and tests based on φ -divergences with $\varphi(x) = \varphi_0(x) := -\log x + x - 1$, namely, on KL_m -divergence. Indeed, it is straightforward to see that the empirical log-likelihood ratio statistic for testing $\mathcal{H}_0 : P_0 \in \mathcal{M}$ against $\mathcal{H}_1 : P_0 \notin \mathcal{M}$, in the context of φ -divergences, can be written as $2n\widehat{D}_{KL_m}(\mathcal{M}, P_0)$; and that the EL estimate of θ_T can be written as $\widehat{\theta}_{KL_m} = \arg \inf_{\theta \in \Theta} \widehat{D}_{KL_m}(\mathcal{M}_\theta, P_0)$; see Remark 4.3 below. In the case of the power functions $\varphi = \varphi_\gamma$, the corresponding estimates (4.10) belong to the class of GEL estimates introduced by Smith (1997) and Newey and Smith (2004), and (4.8) in this case are the empirical Cressie-Read statistics introduced by Baggerly (1998) and Corcoran (1998); see Remark 4.4 below.

The constrained optimization problems (4.8), (4.9) and (4.10) can be transformed into unconstrained ones making use of some arguments of “duality” which we briefly state below from Rockafellar (1970). On the other hand, the obtaining of asymptotic statistical results of the estimates and the test statistics, under misspecification or under alternative hypotheses, requires handle existence conditions and characterization of the projection of P_0 on the submodel $\mathcal{M}_\theta$ or on the model $\mathcal{M}$. This also will be considered through duality, along the following Section.

4.4 Dual representation of φ -divergences under constraints

This Section is central for our purposes. Indeed, it provides the explicit form of the proposed estimates by transforming the constrained problems (4.8) to unconstrained ones, using Lagrangian duality which is a classical tool in optimization theory. This Section adapts this formalism to the context of φ -divergences and the present statistical setting. The Lagrangian “dual” problem, corresponding to the “primal” one

$$\inf_{Q \in \mathcal{M}_\theta} D_\varphi(Q, P_0) \quad (4.11)$$

and its empirical counterpart (4.8), make use of the Fenchel-Legendre transform of φ , defined through

$$\varphi^* : t \in \mathbb{R} \mapsto \varphi^*(t) := \sup_{x \in \mathbb{R}} \{tx - \varphi(x)\}. \quad (4.12)$$

We refer to Section 1.1, for the properties of $\varphi^*(\cdot)$ and to Table 1.1 for explicit expressions of φ^* for some standard divergences. In all the sequel, we assume that φ is strictly convex and essentially smooth on its domain. The “dual” problems associated to (4.11) and (4.8) are respectively

$$\sup_{t \in \mathbb{R}^{1+l}} \left\{ t_0 - \int_{\mathbb{R}^m} \varphi^*\left(t_0 + \sum_{j=1}^l t_j g_j(x, \theta)\right) dP_0(x) \right\}, \quad (4.13)$$

and

$$\sup_{t \in \mathbb{R}^{1+l}} \left\{ t_0 - \frac{1}{n} \sum_{i=1}^n \varphi^*\left(t_0 + \sum_{j=1}^l t_j g_j(X_i, \theta)\right) \right\}. \quad (4.14)$$

In the following Propositions 4.1 and 4.2, we state sufficient conditions under which the primal problems (4.11) and (4.8) coincide respectively with the dual ones (4.13) and (4.14).

Proposition 4.1. *Let θ be a given value in Θ . If there exists Q_0 in $\mathcal{M}_\theta^{(n)}$ such that*

$$a_\varphi < nQ_0(X_i) < b_\varphi, \quad \text{for all } i = 1, \dots, n, \quad (4.15)$$

then

$$\inf_{Q \in \mathcal{M}_\theta^{(n)}} D_\varphi(Q, P_n) = \sup_{t \in \mathbb{R}^{1+l}} \left\{ t_0 - \frac{1}{n} \sum_{i=1}^n \varphi^*\left(t_0 + \sum_{j=1}^l t_j g_j(X_i, \theta)\right) \right\} \quad (4.16)$$

with dual attainment. Conversely, if there exists some dual optimal solution $\hat{t} := (\hat{t}_0, \hat{t}_1, \dots, \hat{t}_l)^\top \in \mathbb{R}^{1+l}$ such that

$$a_{\varphi^*} < \hat{t}_0 + \sum_{j=1}^l \hat{t}_j g_j(X_i, \theta) < b_{\varphi^*}, \quad \text{for all } i = 1, \dots, n, \quad (4.17)$$

then the equality (4.16) holds, and the unique optimal solution of the primal problem $\inf_{Q \in \mathcal{M}_\theta^{(n)}} D_\varphi(Q, P_n)$, namely, the projection of P_n on $\mathcal{M}_\theta^{(n)}$, is given by $Q_\theta^{(n)}(X_i) = \frac{1}{n} \varphi'^{-1}(\hat{t}_0 + \sum_{j=1}^l \hat{t}_j g_j(X_i, \theta))$, $i = 1, \dots, n$, where $\hat{t} := (\hat{t}_0, \hat{t}_1, \dots, \hat{t}_l)^\top$ is solution of the system of equations

$$\begin{cases} 1 - \frac{1}{n} \sum_{i=1}^n \varphi'^{-1}(\hat{t}_0 + \sum_{j=1}^l \hat{t}_j g_j(X_i, \theta)) &= 0, \\ -\frac{1}{n} \sum_{i=1}^n g_j(X_i, \theta) \varphi'^{-1}(\hat{t}_0 + \sum_{j=1}^l \hat{t}_j g_j(X_i, \theta)) &= 0, \quad j = 1, \dots, l. \end{cases}$$

Remark 4.1. For the χ^2 -divergence, we have $a_\varphi = -\infty$ and $b_\varphi = +\infty$. Hence, condition (4.15) holds whenever $\mathcal{M}_\theta^{(n)}$ is not void. More generally, the above Proposition holds for any φ -divergence with $\text{dom}_\varphi = \mathbb{R}$.

Remark 4.2. Assume that $g(x, \theta) := (x - \theta)^\top$. So, for any divergence D_φ with $\text{dom}_\varphi =]0, +\infty[$, which is the case of the modified χ^2 divergence and the modified Kullback-Leibler divergence (or equivalently EL method), condition (4.15) means that θ is an interior point of the convex hull of the data $(X_1, \dots, X_n)$. This is precisely what is checked in Owen (1990), p. 100, for the EL method; see also Owen (2001).

For the asymptotic counterpart of the above results we have; see Theorem 1.1 in Section 1.1. See also Theorem 1.13, Corollary 1.7 and Proposition 1.4 in Section 1.5 for the same results under different assumptions.

Proposition 4.2. *Let θ be a given value in Θ . Assume that $\int_{\mathbb{R}^m} |g_j(x, \theta)| dP_0(x) < \infty$, for all $j = 1, \dots, l$. If there exists Q_0 in $\mathcal{M}_\theta$ with $D_\varphi(Q_0, P_0) < \infty$ and²*

$$a_\varphi < \inf_{x \in \mathbb{R}^m} \frac{dQ_0}{dP_0}(x) \leq \sup_{x \in \mathbb{R}^m} \frac{dQ_0}{dP_0}(x) < b_\varphi, \quad P_0 - a.s., \quad (4.18)$$

then

$$\inf_{Q \in \mathcal{M}_\theta} D_\varphi(Q, P_0) = \sup_{t \in \mathbb{R}^{1+l}} \left\{ t_0 - \int_{\mathbb{R}^m} \varphi^*(t_0 + \sum_{j=1}^l t_j g_j(x, \theta)) dP_0(x) \right\} \quad (4.19)$$

with dual attainment. Conversely, if there exists some dual optimal solution t^* which is an interior point of the set

$$\left\{ t \in \mathbb{R}^{1+l} \text{ such that } \int_{\mathbb{R}^m} |\varphi^*(t_0 + \sum_{j=1}^l t_j g_j(x, \theta))| dP_0(x) < \infty \right\}, \quad (4.20)$$

then the dual equality (4.19) holds, and the unique optimal solution, say Q_θ^* , of the primal problem $\inf_{Q \in \mathcal{M}_\theta} D_\varphi(Q, P_0)$, namely the projection of P_0 on $\mathcal{M}_\theta$, is given by $\frac{dQ_\theta^*}{dP_0}(x) = \varphi'^{-1}(t_0^* + \sum_{j=1}^l t_j^* g_j(x, \theta))$, where $t^* := (t_0^*, t_1^*, \dots, t_l^*)^\top$ is solution of the system of equations

$$\begin{cases} 1 - \int \varphi'^{-1}(t_0^* + \sum_{j=1}^l t_j^* g_j(x, \theta)) dP_0(x) &= 0, \\ - \int g_j(x, \theta) \varphi'^{-1}(t_0^* + \sum_{j=1}^l t_j^* g_j(x, \theta)) dP_0(x) &= 0, \quad j = 1, \dots, l. \end{cases} \quad (4.21)$$

Furthermore, t^* is unique if the functions $\mathbf{1}_{\mathbb{R}^m}, g_1(\cdot, \theta), \dots, g_l(\cdot, \theta)$ are linearly independent in the sense that $P_0 \left\{ x \in \mathbb{R}^m \mid t_0 + \sum_{j=1}^l t_j g_j(x, \theta) \neq 0 \right\} > 0$, for all $t \in \mathbb{R}^{1+l}$, with $t \neq 0$.

For sake of brevity and clearness, we must introduce some additional notations. In all the sequel, $\|x\|$ denotes the norm of x defined by $\|x\| := \sup_i |x_i|$ for any vector $x := (x_1, \dots, x_k)^\top \in \mathbb{R}^k$, and for any matrix A , the norm of A is defined by $\|A\| := \sup_{i,j} |a_{i,j}|$. Denote by $\bar{g}$ the vector valued function $\bar{g} := (\mathbf{1}_{\mathbb{R}^m}, g_1, \dots, g_l)^\top \in \mathbb{R}^{1+l}$. For any p.m. P on $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ and any real measurable function f from $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, denote $Pf := \int_{\mathbb{R}^m} f(x) dP(x)$. Let $t^\top \bar{g}(x, \theta) := t_0 + \sum_{j=1}^l t_j g_j(x, \theta)$ and

$$m(x, \theta, t) := t_0 - \varphi^*(t^\top \bar{g}(x, \theta)), \text{ for all } x \in \mathbb{R}^m, \theta \in \Theta \subset \mathbb{R}^d, t \in \mathbb{R}^{1+l}. \quad (4.22)$$

Note that the sup in (4.16) and (4.19) can be restricted, respectively, to the sets

$$\Lambda_\theta^{(n)} := \left\{ t \in \mathbb{R}^{1+l} \mid a_{\varphi^*} < t^\top \bar{g}(X_i, \theta) < b_{\varphi^*}, \text{ for all } i = 1, \dots, n \right\} \quad (4.23)$$

and

$$\Lambda_\theta := \left\{ t \in \mathbb{R}^{1+l} \mid \int_{\mathbb{R}^m} |\varphi^*(t_0 + \sum_{j=1}^l t_j g_j(x, \theta))| dP_0(x) < \infty \right\}. \quad (4.24)$$

²The strict inequalities (4.18) mean that $P_0 \left\{ x \in \mathbb{R}^m \mid \frac{dQ_0}{dP_0}(x) \leq a_\varphi \right\} = P_0 \left\{ x \in \mathbb{R}^m \mid \frac{dQ_0}{dP_0}(x) \geq b_\varphi \right\} = 0$.

In view of the above two Propositions 4.1 and 4.2, we redefine the estimates (4.8), (4.9) and (4.10) as follows

$$\widehat{D}_\varphi(\mathcal{M}_\theta, P_0) := \sup_{t \in \Lambda_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^n m(X_i, \theta, t) := \sup_{t \in \Lambda_\theta^{(n)}} P_n m(\theta, t), \quad (4.25)$$

$$\widehat{D}_\varphi(\mathcal{M}, P_0) := \inf_{\theta \in \Theta} \sup_{t \in \Lambda_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^n m(X_i, \theta, t) := \inf_{\theta \in \Theta} \sup_{t \in \Lambda_\theta^{(n)}} P_n m(\theta, t) \quad (4.26)$$

and

$$\widehat{\theta}_\varphi := \arg \inf_{\theta \in \Theta} \sup_{t \in \Lambda_\theta^{(n)}} \frac{1}{n} \sum_{i=1}^n m(X_i, \theta, t) := \arg \inf_{\theta \in \Theta} \sup_{t \in \Lambda_\theta^{(n)}} P_n m(\theta, t). \quad (4.27)$$

Remark 4.3. When $\varphi(x) = -\log x + x - 1$, then the estimate (4.10) clearly coincides with the EL one, so it can be seen as the value of the parameter which minimizes the KL_m -divergence between the model $\mathcal{M}$ and the empirical measure P_n of the data $X_1, \dots, X_n$. The statistic $2n\widehat{D}_{KL_m}(\mathcal{M}, P_0)$, see (4.9), coincides with the empirical likelihood ratio statistic associated to the null hypothesis $\mathcal{H}_0 : P_0 \in \mathcal{M}$ against the alternative $\mathcal{H}_1 : P_0 \notin \mathcal{M}$. The dual representation of $\widehat{D}_{KL_m}(\mathcal{M}, P_0)$, see (4.26) and (4.16), is

$$\widehat{D}_{KL_m}(\mathcal{M}, P_0) = \inf_{\theta \in \Theta} \sup_{t \in \Lambda_\theta^{(n)}} \left\{ t_0 + \frac{1}{n} \sum_{i=1}^n \log(1 - t_0 - \sum_{j=1}^l t_j g_j(X_i, \theta)) \right\}.$$

For a given $\theta \in \Theta$, the KL_m -projection $Q_\theta^{(n)}$, of P_n on $\mathcal{M}_\theta$, is given by (see Proposition 4.1)

$$\frac{1}{Q_\theta^{(n)}(X_i)} = n \left(1 - \widehat{t}_0 - \sum_{j=1}^l \widehat{t}_j g_j(X_i, \theta) \right), \quad i = 1, \dots, n,$$

which, multiplying by $Q_\theta^{(n)}(X_i)$ and summing upon $i = 1, \dots, n$, yields $\widehat{t}_0 = 0$. Therefore, t_0 can be omitted, and the above representation can be rewritten under the form

$$\widehat{D}_{KL_m}(\mathcal{M}, P_0) = \inf_{\theta \in \Theta} \sup_{t_1, \dots, t_l} \left\{ \frac{1}{n} \sum_{i=1}^n \log(1 + \sum_{j=1}^l t_j g_j(X_i, \theta)) \right\},$$

and then

$$\widehat{\theta}_{KL_m} = \widehat{\theta}_{EL} = \arg \inf_{\theta \in \Theta} \sup_{t_1, \dots, t_l} \left\{ \frac{1}{n} \sum_{i=1}^n \log(1 + \sum_{j=1}^l t_j g_j(X_i, \theta)) \right\} \quad (4.28)$$

in which the sup is taken over the set $\{(t_1, \dots, t_l)^\top \in \mathbb{R}^m \mid -1 < \sum_{j=1}^l t_j g_j(X_i, \theta) < +\infty, \forall i = 1, \dots, n\}$. The formula (4.28) is the ordinary dual representation of the EL estimate; see Qin and Lawless (1994) and Owen (2001).

Remark 4.4. Consider the power divergences, associated to the power functions φ_γ ; see (1.4) and (1.5). We will show that the estimates $\widehat{\theta}_{\varphi_\gamma}$ belong to the class of GEL estimators introduced by Smith (1997) and Newey and Smith (2004). The projection $Q_\theta^{(n)}$ of P_n on $\mathcal{M}_\theta$ is given by

$$Q_\theta^{(n)}(X_i) = \left((\gamma - 1)(\widehat{t}_0 + \sum_{j=1}^l \widehat{t}_j g_j(X_i, \theta)) + 1 \right)^{1/(\gamma-1)}, \quad i = 1, \dots, n.$$

Using the constraint $\sum_{i=1}^n Q_\theta^{(n)}(X_i) = 1$, we can explicit $\widehat{t}_0$ in terms of $\widehat{t}_1, \dots, \widehat{t}_l$, and hence the sup in the dual representation (4.27) can be reduced to a subset of $\mathbb{R}^l$, as in Newey and Smith (2004). When $\varphi(x) = \frac{1}{2}(x - 1)^2$, it is straightforward to see that the corresponding estimate $\widehat{\theta}_\varphi$ coincides with the continuous updating estimator of Hansen *et al.* (1996).

Remark 4.5. (Numerical calculation of the estimates and the specific role of the χ^2 - divergence). The computation of $\hat{t}(\theta)$ for fixed $\theta \in \Theta$ as defined in (4.21) is difficult when handling a generic divergence. In the particular case of χ^2 -divergence, i.e., when $\varphi(x) = \frac{1}{2}(x-1)^2$, optimizing on all s.f.m's, the system (4.21) is linear; we thus easily obtain an explicit form for $\hat{t}(\theta)$, which in turn allows for a single gradient descent when optimizing upon Θ . This procedure is useful in order to compute the estimates for all other divergences (for which the corresponding system is non linear) including EL, since it provides an easy starting point for the resulting double gradient descent. Moreover, Hjort *et al.* (2009) extend the EL approach, to more general moment condition models, allowing the number of constraints to increase with growing sample size. In this case, the computation of EL estimate is more complex, and the same idea as above can help to solve the problem.

4.5 Asymptotic properties of the estimates of the parameter and the divergences

4.5.1 Asymptotic properties under the model

This Section addresses Problems 1 and 2, aiming at testing the null hypothesis $\mathcal{H}_0 : P_0 \in \mathcal{M}$ against the alternative $\mathcal{H}_1 : P_0 \notin \mathcal{M}$. We derive the limiting distributions of the proposed test statistics which are the estimated divergences between the model $\mathcal{M}$ and P_0 . We also derive the limiting distributions of the estimates of θ_T . The following two Theorems 5.1 and 5.2 extend Theorems 3.1 and 3.2 in Newey and Smith (2004) to the context of divergence based approach. The Assumptions which we will consider match those of Theorems 3.1 and 3.2 in Newey and Smith (2004).

Assumption 4.1. a) $P_0 \in \mathcal{M}$ and $\theta_T \in \Theta$ is the unique solution of $\mathbb{E}[g(X, \theta)] = 0$; b) $\Theta \subset \mathbb{R}^d$ is compact; c) $g(X, \theta)$ is continuous at each $\theta \in \Theta$ with probability one; d) $\mathbb{E}[\sup_{\theta \in \Theta} \|g(X, \theta)\|^\alpha] < \infty$ for some $\alpha > 2$; e) the matrix $\Omega := \mathbb{E}[g(X, \theta_0)g(X, \theta_T)^\top]$ is nonsingular.

Theorem 4.1. Under Assumption 4.1, with probability approaching one as $n \rightarrow \infty$, the estimate $\hat{\theta}_\varphi$ exists, and converges to θ_T in probability. Moreover, $\frac{1}{n} \sum_{i=1}^n g(X_i, \hat{\theta}_\varphi) = O_P(1/\sqrt{n})$, $\hat{t}(\hat{\theta}_\varphi) := \arg \sup_{t \in \Lambda_{\hat{\theta}_\varphi}^{(n)}} P_n m(\hat{\theta}_\varphi, t)$ exists and belongs to $\text{int}(\Lambda_{\hat{\theta}_\varphi}^{(n)})$ with probability approaching one as $n \rightarrow \infty$, and $\hat{t}(\hat{\theta}_\varphi) = O_P(1/\sqrt{n})$.

In order to obtain asymptotic normality, we need some additional Assumptions. Denote by G the matrix $G := \mathbb{E}[\partial g(X, \theta_T)/\partial \theta]$.

Assumption 4.2. a) $\theta_T \in \text{int}(\Theta)$; b) with probability one, $g(X, \theta)$ is continuously differentiable in a neighborhood N_{θ_T} of θ_T , and $\mathbb{E}[\sup_{\theta \in N_{\theta_T}} \|\partial g(X, \theta)/\partial \theta\|] < \infty$; c) $\text{rank}(G) = d$.

Theorem 4.2. Assume that Assumptions 4.1 and 4.2 hold. Then,

- (a) $\sqrt{n}(\hat{\theta}_\varphi - \theta_T)$ converges in distribution to a centered normal random vector with covariance matrix $V := [G\Omega^{-1}G^\top]^{-1}$.
- (b) If $l > d$, then the statistic $2n\hat{D}_\varphi(\mathcal{M}, P_0)$ converges in distribution to a χ^2 -random variable with $(l - d)$ degrees of freedom.

Remark 4.6. The above Theorem allows to perform statistical tests (of the model) with asymptotic level $\alpha \in]0, 1[$. Consider the null hypothesis

$$\mathcal{H}_0 : P_0 \in \mathcal{M} \quad \text{against the alternative} \quad \mathcal{H}_1 : P_0 \notin \mathcal{M}. \quad (4.29)$$

The critical region is then $C_\varphi := \{2n\hat{D}_\varphi(\mathcal{M}, P_0) > \chi_{l-d}^2(1-\alpha)\}$ where $\chi_{l-d}^2(1-\alpha)$ is the $(1-\alpha)$ -quantile of the χ_{l-d}^2 -distribution. When $\varphi(x) = -\log x + x - 1$, it is straightforward to see that the corresponding test is the empirical likelihood ratio one; see Qin and Lawless (1994).

4.5.2 Asymptotic properties of the estimates of the divergences for a given value of the parameter

For a given $\theta \in \Theta$, consider the test problem of the null hypothesis $\mathcal{H}_0 : P_0 \in \mathcal{M}_\theta$ against two different families of alternative hypotheses: $\mathcal{H}_1 : P_0 \notin \mathcal{M}_\theta$ and $\mathcal{H}'_1 : P_0 \in \mathcal{M} \setminus \mathcal{M}_\theta$. Those two tests address different situations since $\mathcal{H}_1$ may include misspecification of the model. We give two different test statistics each pertaining to one of the situations and derive their limiting distributions both under $\mathcal{H}_0$ and under the alternatives. As a by product, we also derive confidence areas for the true value θ_T of the parameter. We will first state the convergence in probability of $\widehat{D}_\varphi(\mathcal{M}_\theta, P_0)$ to $D_\varphi(\mathcal{M}_\theta, P_0)$, and then we obtain the limiting distribution of $\widehat{D}_\varphi(\mathcal{M}_\theta, P_0)$ both when $P_0 \in \mathcal{M}_\theta$ and when $P_0 \notin \mathcal{M}_\theta$. Obviously, when $P_0 \in \mathcal{M}_\theta$, this means that $\theta = \theta_T$ since the true value θ_T of the parameter is assumed to be unique.

Assumption 4.3. a) $P_0 \in \mathcal{M}_\theta$ and θ is the unique solution of $\mathbb{E}[g(X, \theta)] = 0$; b) $\mathbb{E}[\|g(X, \theta)\|^\alpha] < \infty$ for some $\alpha > 2$; c) the matrix $\Omega := \mathbb{E}[g(X, \theta)g(X, \theta)^\top]$ is nonsingular.

Theorem 4.3. Under Assumption 4.3, we have

(a) $\widehat{t}(\theta) := \arg \sup_{t \in \Lambda_\theta^{(n)}} P_n m(\theta, t)$ exists and belongs to $\text{int}(\Lambda_\theta^{(n)})$ with probability approaching one as $n \rightarrow \infty$, and $\widehat{t}(\theta) = O_P(1/\sqrt{n})$;

(b) The statistic $2n\widehat{D}_\varphi(\mathcal{M}_\theta, P_0)$ converges in distribution to a χ^2_l -random variable.

In order to obtain the limiting distribution of the test statistic $2n\widehat{D}_\varphi(\mathcal{M}_\theta, P_0)$ under the alternative $\mathcal{H}_1 : P_0 \notin \mathcal{M}_\theta$, including misspecification, the following Assumption is needed.

Assumption 4.4. a) $P_0 \notin \mathcal{M}_\theta$, and $t^*(\theta) := \arg \sup_{t \in \Lambda_\theta} \mathbb{E}[m(X, \theta, t)]$ exists and is an interior point of Λ_θ ; b) $\mathbb{E}[\sup_{t \in N_{t^*(\theta)}} |m(X, \theta, t)|] < \infty$ for some compact set $N_{t^*(\theta)} \subset \Lambda_\theta$ such that $t^*(\theta) \in \text{int}(N_{t^*(\theta)})$; c) $\mathbb{1}_{\mathbb{R}^m}, g_1, \dots, g_l$ are linearly independent in the sense that $P_0 \left\{ x \in \mathbb{R}^m \mid t_0 + \sum_{j=1}^l t_j g_j(x, \theta) \neq 0 \right\} > 0$, for all $t \in \mathbb{R}^{1+l}$, with $t \neq 0$.

Remark 4.7. Assumption 4.4.c above ensures the strict concavity of the function $t \in \Lambda_\theta \mapsto \mathbb{E}[m(X, \theta, t)]$ on the convex set Λ_θ , which implies that $t^*(\theta)$ is unique. It can be replaced by the following one : there exists a neighborhood, $N_{t^*(\theta)} \subset \Lambda_\theta$, of $t^*(\theta)$, such that $\mathbb{E}[\sup_{t \in N_{t^*(\theta)}} \|\partial m(X, \theta, t)/\partial t\|] < \infty$, $\mathbb{E}[\sup_{t \in N_{t^*(\theta)}} \|\partial^2 m(X, \theta, t)/\partial t^2\|] < \infty$ and the matrix $\mathbb{E}[\partial^2 m(X, \theta, t^*(\theta))/\partial t^2]$ is nonsingular; which implies also that $t^*(\theta)$ is unique.

Theorem 4.4. Under Assumption 4.4, when $P_0 \notin \mathcal{M}_\theta$, we have

(a) $\widehat{t}(\theta)$ converges in probability to $t^*(\theta)$.

(b) $\widehat{D}_\varphi(\mathcal{M}_\theta, P_0)$ converges in probability to $D_\varphi(\mathcal{M}_\theta, P_0)$.

We now give the limiting distribution of the test statistic under $\mathcal{H}_1$. We need the following additional condition.

Assumption 4.5. a) There exists $N_{t^*(\theta)} \subset \Lambda_\theta$, some compact neighborhood of $t^*(\theta)$, such that

$$\mathbb{E}[\sup_{t \in N_{t^*(\theta)}} \|\partial m(X, \theta, t)/\partial t\|] < \infty, \quad \mathbb{E}[\sup_{t \in N_{t^*(\theta)}} \|\partial^2 m(X, \theta, t)/\partial t^2\|] < \infty;$$

b) as $\delta \rightarrow 0$,

$$\mathbb{E} \left\{ \sup_{\{t: \|t - t^*(\theta)\| \leq \delta\}} \left\| \partial^2 m(X, \theta, t)/\partial t^2 - \partial^2 m(X, \theta, t^*(\theta))/\partial t^2 \right\| \right\} \rightarrow 0;$$

c) $\mathbb{E}[m(X, \theta, t^*(\theta))^2] < \infty$, $\mathbb{E}[\|\partial m(X, \theta, t^*(\theta))/\partial t\|^2] < \infty$ and the matrix $\mathbb{E}[\partial^2 m(X, \theta, t^*(\theta))/\partial t^2]$ is nonsingular.

Remark 4.8. Assumption 4.5.b is used here to relax the condition on the third derivatives (in t) of the function $t \mapsto m(X, \theta, t)$.

Theorem 4.5. Under Assumptions 4.4 and 4.5, we have

- (a) $\sqrt{n}(\hat{t}(\theta) - t^*(\theta))$ converges in distribution to a centered normal random vector with covariance matrix $[\mathbb{E}[m''(X, \theta, t^*)]]^{-1} \mathbb{E}[m'(X, \theta, t^*)m'(X, \theta, t^*)^\top] [\mathbb{E}[m''(X, \theta, t^*)]]^{-1}$;
- (b) $\sqrt{n}(\hat{D}_\varphi(\mathcal{M}_\theta, P_0) - D_\varphi(\mathcal{M}_\theta, P_0))$ converges in distribution to a centered normal random variable with variance $\sigma^2(\theta) := \mathbb{E}[m(X, \theta, t^*(\theta))^2] - [\mathbb{E}[m(X, \theta, t^*(\theta))]]^2$.

Remark 4.9. Let θ be a given value in Θ . Consider the test of the null hypothesis

$$\mathcal{H}_0 : P_0 \in \mathcal{M}_\theta \quad \text{against} \quad \mathcal{H}_1 : P_0 \notin \mathcal{M}_\theta. \quad (4.30)$$

In view of Theorem 4.3 part (b), we reject $\mathcal{H}_0$ against $\mathcal{H}_1$, at asymptotic level $\alpha \in]0, 1[$, when $2n\hat{D}_\varphi(\mathcal{M}_\theta, P_0)$ exceeds $\chi_l^2(1 - \alpha)$, the $(1 - \alpha)$ -quantile of the χ_l^2 -distribution. Theorem 4.5 part (b) is useful to give an approximation to the power function $P_0 \notin \mathcal{M}_\theta \mapsto \beta(P_0) := P_0[2n\hat{D}_\varphi(\mathcal{M}_\theta, P_0) > \chi_l^2(1 - \alpha)]$. We obtain then the following approximation

$$\beta(P_0) \approx 1 - F_N\left(\frac{\sqrt{n}}{\sigma(\theta)} \left[\frac{\chi_l^2(1 - \alpha)}{2n} - D_\varphi(\mathcal{M}_\theta, P_0) \right]\right), \quad (4.31)$$

where F_N is the cumulative distribution function of the standard normal distribution. From this approximation, we can give the approximate sample size that ensures a desired power β for a given alternative $P_0 \notin \mathcal{M}_\theta$. Let n_0 be the positive root of the equation $\beta = 1 - F_N\left[\frac{\sqrt{n}}{\sigma(\theta)} \left(\frac{\chi_l^2(1 - \alpha)}{2n} - D_\varphi(\mathcal{M}_\theta, P_0) \right)\right]$, i.e., $n_0 = \frac{(a+b) - \sqrt{a(a+2b)}}{2D_\varphi(\mathcal{M}_\theta, P_0)^2}$ with $a := \sigma(\theta^*)^2 [F_N^{-1}(1 - \beta)]^2$ and $b := \chi_l^2(1 - \alpha)D_\varphi(\mathcal{M}_\theta, P_0)$. The required sample size is then $[n_0] + 1$, where $[n_0]$ denotes the integer part of n_0 .

Remark 4.10. (Generalized empirical likelihood ratio test). For testing $\mathcal{H}_0 : P_0 \in \mathcal{M}_\theta$ against the alternative $\mathcal{H}_1' : \mathcal{M} \setminus \mathcal{M}_\theta$, we propose to use the statistics

$$2nS_n^\varphi := 2n \left[\hat{D}_\varphi(\mathcal{M}_\theta, P_0) - \inf_{\theta \in \Theta} \hat{D}_\varphi(\mathcal{M}_\theta, P_0) \right], \quad (4.32)$$

which converge in distribution to a χ_d^2 -random variable under $\mathcal{H}_0$ when Assumptions 4.1 and 4.2 hold. This can be proved using similar arguments as in Theorems 5.2 and 4.3. We then reject $\mathcal{H}_0$ at asymptotic level α when $2nS_n^\varphi > \chi_d^2(1 - \alpha)$, the $(1 - \alpha)$ -quantile of the χ_d^2 -distribution. Under $\mathcal{H}_1'$ and when Assumptions 4.1, 4.2, 4.4 and 4.5 hold, as in Theorem 4.5, it can be proved that

$$\sqrt{n}(S_n^\varphi - D_\varphi(\mathcal{M}_\theta, P_0)) \quad (4.33)$$

converges to a centered normal random variable with variance

$$\sigma^2(\theta) := \mathbb{E}(m(X, \theta, t^*(\theta))^2) - (\mathbb{E}m(X, \theta, t^*(\theta)))^2.$$

So, as in the above Remark, we obtain the following approximation

$$\beta(P_0) \approx 1 - F_N\left(\frac{\sqrt{n}}{\sigma(\theta)} \left[\frac{\chi_d^2(1 - \alpha)}{2n} - D_\varphi(\mathcal{M}_\theta, P_0) \right]\right) \quad (4.34)$$

to the power function $P_0 \in \mathcal{M}/\mathcal{M}_\theta \mapsto \beta(P_0) := P_0[2nS_n^\varphi > \chi_d^2(1 - \alpha)]$. The approximated sample size required to achieve a desired power for a given alternative can be obtained in a similar way.

Remark 4.11. (Confidence region for the parameter). For a fixed level $\alpha \in]0, 1[$, using convergence (4.32), the set $\{\theta \in \Theta \text{ such that } 2nS_n^\varphi \leq \chi_d^2(1 - \alpha)\}$ is an asymptotic confidence region for θ_T where $\chi_d^2(1 - \alpha)$ is the $(1 - \alpha)$ -quantile of the χ_d^2 -distribution. It is straightforward to see that the confidence region obtained for the KL_m -divergence coincides with that of Owen (1991) and Qin and Lawless (1994).

4.5.3 Asymptotic properties under misspecification

We address Problem 1 stating the limiting distribution of the proposed test statistics under the alternative $\mathcal{H}_1 : P_0 \notin \mathcal{M}$. This needs the introduction of $Q_{\theta^*}^*$, the projection of P_0 on $\mathcal{M}$. Assumption 6 below ensures the existence of the “pseudo-true” value θ^* as well as the existence of the projection $Q_{\theta^*}^*$ of P_0 on $\mathcal{M}$, and states some necessary other regularity conditions. Proposition 4.2 above states the existence and characterization of the projection Q_θ^* of P_0 on $\mathcal{M}_\theta$, for a given $\theta \in \Theta$.

Assumption 4.6. a) Θ is compact, $\theta^* := \arg \inf_{\theta \in \Theta} \sup_{t \in \Lambda_\theta} \mathbb{E}[m(X, \theta, t)]$ exists and is unique; b) $g(X, \theta)$ is continuous at each $\theta \in \Theta$ with probability one; c) $\mathbb{E} \left[\sup_{\{\theta \in \Theta, t \in N_{t^*}(\theta)\}} |m(X, \theta, t)| \right] < \infty$, where $N_{t^*}(\theta) \subset \Lambda_\theta$ is a compact set such that $t^*(\theta) \in \text{int}(N_{t^*}(\theta))$; d) for all $\theta \in \Theta$, the functions $\mathbb{1}_{\mathbb{R}^m}, g_1, \dots, g_l$ are linearly independent in the sense that $P_0 \left\{ x \in \mathbb{R}^m \mid t_0 + \sum_{j=1}^l t_j g_j(x, \theta) \neq 0 \right\} > 0$, for all $t \in \mathbb{R}^{1+l}$ with $t \neq 0$.

Remark 4.12. Assumption 4.6.d ensures the strict concavity of the function $t \in \Lambda_\theta \mapsto \mathbb{E}[m(X, \theta, t)]$ on the convex set Λ_θ , which implies the uniqueness of $t^*(\theta)$, for all $\theta \in \Theta$. This Assumption can be replaced by the following one : for all $\theta \in \Theta$, there exists a neighborhood $N_{t^*}(\theta)$ of $t^*(\theta)$ such that $\mathbb{E}[\sup_{t \in N_{t^*}(\theta)} \|\partial m(X, \theta, t)/\partial t\|] < \infty$, $\mathbb{E}[\sup_{t \in N_{t^*}(\theta)} \|\partial^2 m(X, \theta, t)/\partial t^2\|] < \infty$, and the matrix $\mathbb{E}[\partial^2 m(X, \theta, t^*(\theta))/\partial t^2] < \infty$ is nonsingular, which implies the uniqueness of $t^*(\theta)$, for all $\theta \in \Theta$.

Theorem 4.6. Under Assumption 4.6, we have

- (a) $\|\hat{t}(\theta) - t^*(\theta)\|$ converges in probability to 0 uniformly in $\theta \in \Theta$;
- (b) $\hat{\theta}_\varphi$ converges in probability to θ^* ;
- (c) $\hat{D}_\varphi(\mathcal{M}, P_0)$ converges in probability to $D_\varphi(\mathcal{M}, P_0)$.

The asymptotic normality of the test statistics under misspecification requires the following additional conditions.

Assumption 4.7. a) $\theta^* \in \text{int}(\Theta)$; b) there exists $\mathcal{N} \subset \Theta \times \Lambda_\Theta$, some compact neighborhood of $(\theta^*, t^*(\theta^*))$, s.t. with probability one $(\theta, t) \in \mathcal{N} \mapsto m(X, \theta, t)$ is $\mathcal{C}^2$ and $\mathbb{E}[\sup_{(\theta, t) \in \mathcal{N}} \|\partial m(X, \theta, t)/\partial(\theta, t)\|] < \infty$, $\mathbb{E}[\sup_{(\theta, t) \in \mathcal{N}} \|\partial^2 m(X, \theta, t)/\partial(\theta, t)^2\|] < \infty$; c) as $\delta \rightarrow 0$,

$$\mathbb{E} \left\{ \sup_{\{(t, \theta); \|(t, \theta) - (t^*(\theta^*), \theta^*)\| \leq \delta\}} \left\| \partial^2 m(X, \theta, t)/\partial(\theta, t)^2 - \partial^2 m(X, \theta^*, t^*(\theta^*))/\partial(\theta, t)^2 \right\| \right\} \rightarrow 0;$$

d) $\mathbb{E}[m(X, \theta^*, t^*(\theta^*))^2]$, $\mathbb{E}[\|\partial m(X, \theta^*, t^*(\theta^*))/\partial t\|^2]$ and $\mathbb{E}[\|\partial m(X, \theta^*, t^*(\theta^*))/\partial \theta\|^2]$ are finite, and the matrix $S := \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$, is nonsingular, where $S_{11} := \mathbb{E}[\partial^2 m(X, \theta^*, t^*(\theta^*))/\partial t^2]$, $S_{12} = S_{21}^\top := \mathbb{E}[\partial^2 m(X, \theta^*, t^*(\theta^*))/\partial t \partial \theta]$ and $S_{22} := \mathbb{E}[\partial^2 m(X, \theta^*, t^*(\theta^*))/\partial \theta^2]$.

Remark 4.13. Assumption 4.7.c is used here to relax the condition on the third derivatives (in t and θ) of the function $(\theta, t) \mapsto m(X, \theta, t)$.

Theorem 4.7. Under Assumptions 4.6 and 4.7, we have

- (a) $\sqrt{n} \left(\hat{t}(\hat{\theta}_\varphi) - t^*(\theta^*), \hat{\theta}_\varphi - \theta^* \right)$ converges in distribution to a centered normal random vector with covariance matrix $W := S^{-1}MS^{-1}$, where $M := \mathbb{E} \left[\begin{bmatrix} \frac{\partial}{\partial t} m(X, \theta^*, t^*(\theta^*)) \\ \frac{\partial}{\partial \theta} m(X, \theta^*, t^*(\theta^*)) \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial t} m(X, \theta^*, t^*(\theta^*)) \\ \frac{\partial}{\partial \theta} m(X, \theta^*, t^*(\theta^*)) \end{bmatrix}^\top \right]$;
- (b) $\sqrt{n} \left(\hat{D}_\varphi(\mathcal{M}, P_0) - D_\varphi(\mathcal{M}, P_0) \right)$ converges in distribution to a centered normal random variable with variance $\sigma^2(\theta^*) := \mathbb{E}[m(X, \theta^*, t^*(\theta^*))^2] - [\mathbb{E}[m(X, \theta^*, t^*(\theta^*))]]^2$.

Remark 4.14. In the case of EL, i.e., when $\varphi(x) = -\log x + x - 1$, Assumption 4.6.c implies that

$$-\infty < \inf_{x \in \mathbb{R}^m} t_0 + t^\top g(x, \theta) \leq \sup_{x \in \mathbb{R}^m} t_0 + t^\top g(x, \theta) < 1 \quad (4.35)$$

for all $x \in \mathbb{R}^m - P_0$ -a.s., for all $\theta \in \Theta$ and for all $t \in N_{t^*(\theta)}$. This imposes a restriction on the model when the support of P_0 and the functions g_j are unbounded. Indeed, when the support of P_0 is for example the whole space $\mathbb{R}^m$, the condition above does not hold when g is unbounded. In this case, the EL estimate may cease to be consistent as it is stated by Schennach (2007) under misspecification. This is a potential problem for all divergences associated to φ -functions with domain of the form $(a_\varphi, +\infty[,] - \infty, b_\varphi)$ or (a_φ, b_φ) , where a_φ and b_φ are some finite real numbers; it is the case of modified χ^2 , Hellinger, KL and modified KL divergences. At the contrary, Assumption 6.c may be satisfied for other divergences associated to φ functions with $\text{dom}_\varphi = \mathbb{R}$ which is the case of χ^2 -divergence for example.

Remark 4.15. Theorem 4.7 part 2 is useful for the computation of the power function. For testing the null hypothesis $\mathcal{H}_0 : P_0 \in \mathcal{M}$ against the alternative $\mathcal{H}_1 : P_0 \notin \mathcal{M}$, the power function is

$$P_0 \notin \mathcal{M} \mapsto \beta(P_0) := P_0 \left[2n\widehat{D}_\varphi(\mathcal{M}, P_0) > \chi_{l-d}^2(1-\alpha) \right]. \quad (4.36)$$

Using Theorem 4.7 part 2, we obtain the following approximation to the power function (4.36)

$$\beta(P_0) \approx 1 - F_{\mathcal{N}} \left[\frac{\sqrt{n}}{\sigma(\theta^*)} \left(\frac{\chi_{l-d}^2(1-\alpha)}{2n} - D_\varphi(\mathcal{M}, P_0) \right) \right] \quad (4.37)$$

where $F_{\mathcal{N}}$ is the empirical cumulative distribution of the standard normal distribution. From the proxy value of $\beta(P_0)$ hereabove, the approximate sample size that ensures a given power β for a given alternative $P_0 \notin \mathcal{M}$ can be obtained as follows. Let n_0 be the positive root of the equation $\beta = 1 - F_{\mathcal{N}} \left[\frac{\sqrt{n}}{\sigma(\theta^*)} \left(\frac{\chi_{l-d}^2(1-\alpha)}{2n} - D_\varphi(\mathcal{M}, P_0) \right) \right]$, i.e., $n_0 = \frac{(a+b) - \sqrt{a(a+2b)}}{2D_\varphi(\mathcal{M}, P_0)^2}$, where $a := \sigma(\theta^*)^2 [F_{\mathcal{N}}^{-1}(1-\beta)]^2$ and $b := \chi_{l-d}^2(1-\alpha)D_\varphi(\mathcal{M}, P_0)$. The required sample size is then $\lfloor n_0 \rfloor + 1$.

4.6 Concluding remarks and possible developments

We have proposed new estimates and tests for model satisfying linear constraints with unknown parameter through divergence based methods which generalize the EL approach. This leads to the obtaining of the limiting distributions of the test statistics and the estimates under alternatives and under misspecification. Consistency of the test statistics under the alternatives is the starting point for the study of the optimality of the tests through Bahadur approach; also the generalized Neyman-Pearson optimality of EL test (as developed by Kitamura (2001)) can be adapted for empirical divergence based methods. Many other problems remain to be studied in the future such as the choice of the divergence which leads to optimal (in some sense) estimator or test in terms of efficiency and/or robustness. Preliminary simulation results show that Hellinger divergence enjoys good properties in terms of efficiency-robustness. Also comparisons under local alternatives should be developed, as well as the case where the number of constraints grows with the sample size.

Chapter 5

Statistical inference for semiparametric two-sample density ratio models

This Chapter presents our papers [Keziou and Leoni-Aubin \(2005\)](#) and [Keziou and Leoni-Aubin \(2008\)](#). We consider estimation and test problems for semiparametric two-sample density ratio models. The profile empirical likelihood (EL) poses an irregularity problem under the null hypothesis that the laws of the two samples are equal. We show that a “dual” form of the profile EL is well defined even under the null hypothesis. A statistical test, based on the dual form of the EL ratio statistic (ELRS), is then proposed. We give an interpretation for the dual form of the ELRS through the convex conjugate expressions of φ -divergences. The asymptotic properties of the test statistic are presented both under the null and the alternative hypotheses, and approximation of the power function of the test is deduced. We refer to [Keziou and Leoni-Aubin \(2008\)](#) for proofs and simulation results, illustrating the accuracy of the proposed estimates, tests and power approximations.

5.1 Introduction and notation

We consider the following problems : two-sample test for comparing two populations and estimation of the parameters for semiparametric density ratio models. We have at hand two samples : $X_{0,1}, \dots, X_{0,n_0}$ with distribution P and $X_{1,1}, \dots, X_{1,n_1}$ with distribution Q . Both distributions Q and P are defined on $\mathbb{R}^m$, $m \geq 1$. We consider the following semiparametric density ratio model

$$\frac{dQ}{dP}(x) = \exp \left\{ \alpha + \beta^\top \mathbf{b}(x) \right\} =: r(\boldsymbol{\theta}, x), \quad (5.1)$$

where $\boldsymbol{\theta}^\top := (\alpha, \beta^\top)$ is the parameter of interest which we suppose to belong to some open set $\Theta \subset \mathbb{R}^{1+d}$. We assume that the above model is identifiable, and denote by $\boldsymbol{\theta}_T^\top := (\alpha_T, \beta_T^\top)$ the true value of the parameter. For simplicity, we sometimes write $r(\boldsymbol{\theta}, x)$ instead of $\exp \left\{ \alpha + \beta^\top \mathbf{b}(x) \right\}$. $\mathbf{b}(\cdot)$ is a known (basis) function on $\mathbb{R}^m$ with values in $\mathbb{R}^d$. It often takes the form $\mathbf{b}(x) = (x, x^2, \dots, x^d)^\top$, if $m = 1$, and the model (5.1) is sometimes called “log-linear model” in this case. The supports of the two laws Q and P may be known or unknown, discrete or not, bounded or unbounded. We now give some statistical examples and motivations for model (5.1).

5.1.1 Comparison of two populations

In applications, we often come across with the problem of comparing two distributions. The use of the well known t -test requires to assume that both samples are normally distributed with unknown means and common (known or unknown) variance. The t -test enjoys several optimal properties, for example it is the uniformly most powerful unbiased test (see e.g. [Lehmann \(1986\)](#)). If both Q and P are normally

distributed with equal variance $Q = \mathcal{N}(\mu_2, \sigma^2)$ and $P = \mathcal{N}(\mu_1, \sigma^2)$, then, the ratio $\frac{dQ}{dP}$ takes the form

$$\frac{dQ}{dP}(x) = \exp\{\alpha + \beta x\}, \text{ where } \alpha = \frac{\mu_1^2 - \mu_2^2}{2\sigma^2} \text{ and } \beta = \frac{\mu_2 - \mu_1}{\sigma^2}. \quad (5.2)$$

It follows that testing the hypothesis $\mathcal{H}_0 : Q = P$ is equivalent to testing the parametric hypothesis $\mathcal{H}_0 : \beta_T = 0$. We underline that $\beta_T = 0$ implies $\alpha_T = 0$. [Kay and Little \(1987\)](#) and [Fokianos \(2003\)](#) observed that there are many practical examples for which the choice

$$\frac{dQ}{dP}(x) = \exp\left\{\alpha + \beta^\top \mathbf{b}(x)\right\}, \quad (5.3)$$

where $\mathbf{b}(\cdot)$ is an arbitrary (but known) function, is more appropriate. When the two samples $X_{0,1}, \dots, X_{0,n_0}$ and $X_{1,1}, \dots, X_{1,n_1}$ are independent, [Fokianos et al. \(2001\)](#) presented a Wald-type test, for the null hypothesis $\mathcal{H}_0 : Q = P$ or equivalently $\mathcal{H}_0 : \beta_T = 0$, where the test statistic is based on a “constrained” empirical likelihood estimate of the parameter β_T (see [Qin \(1998\)](#)) and an empirical estimate of the limiting variance. When the semiparametric assumption (5.1) fails, the test commonly used is the non parametric Wilcoxon rank-sum (see e.g. [Randles and Wolfe \(1979\)](#) and [Hollander and Wolfe \(1999\)](#)). We expect it not to be powerful, since it does not use the model (5.1). For the model (5.1), the empirical likelihood ratio statistic is not well defined under the null hypothesis $\mathcal{H}_0 : Q = P$ (see Section 1.3 below). This problem has been observed also by [Zou et al. \(2002\)](#) in the context of a semiparametric mixture models with known weights (see [Zou et al. \(2002\)](#) Theorem 1). We propose to use, instead of the empirical likelihood ratio statistic, its “dual” form, see (5.21), to perform a test of the null hypothesis $\mathcal{H}_0 : Q = P$, which is well defined regardless of the null hypothesis. Simulation results, presented in [Keziou and Leoni-Aubin \(2008\)](#) Section 4, show that the observed level of our test based on the statistic (5.21) converges (to the nominal level) better than the observed level of the test proposed by [Fokianos et al. \(2001\)](#). Using φ -divergences and “duality” technique, we give an interpretation for the proposed statistic (5.21), the dual form of the empirical likelihood ratio statistic; see (5.29). This interpretation allows us to give the asymptotic law of the proposed test statistic under the alternative hypothesis. We apply this result to give approximation of the power function of the test in a similar way to [Morales and Pardo \(2001\)](#), who gave some approximations of the power functions of φ -divergences tests in the context of finite discrete parametric models for distributions. Recall that a φ -divergence between two probability measures Q and P , defined on a measurable space $(\mathcal{X}, \mathcal{B})$, when Q is absolutely continuous with respect to P , is defined by

$$D_\varphi(Q, P) := \int_{\mathcal{X}} \varphi\left(\frac{dQ}{dP}(x)\right) dP(x), \quad (5.4)$$

where φ is a real nonnegative convex function satisfying $\varphi(1) = 0$. Note that $D_\varphi(Q, P)$ is nonnegative, $D_\varphi(Q, P) = 0$ when $Q = P$. Further, if φ is strictly convex on a neighborhood of one, then $D_\varphi(Q, P) = 0$ if and only if $Q = P$; we refer to [Liese and Vajda \(1987\)](#) for a systematic theory of φ -divergences.

5.1.2 Logistic model and multiplicative-intercept risk model

Consider the logistic model which has been widely used in statistical applications for the analysis of binary data, in particular in binary classification, see e.g. [Agresti \(1990\)](#), [Hosmer and Lemeshow \(1999\)](#) and [Hosmer and Lemeshow \(2000\)](#). Suppose that $Y \in \{0, 1\}$ is a binary response variable and that $X \in \mathbb{R}^m$, with here $m = d$, is the associate covariate vector. The logistic model has the form

$$\Pr(Y = 1|X = \mathbf{x}) = \frac{\exp(\gamma + \beta^\top \mathbf{x})}{1 + \exp(\gamma + \beta^\top \mathbf{x})}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}^d. \quad (5.5)$$

Note that the marginal distribution of X , denote it $P_X(\cdot)$, is left completely unspecified. One of the major reasons of the logistic regression model has seen such a wide use, especially in epidemiologic research, is the ease of obtaining adjusted odds ratios from the estimated slope coefficients when sampling is performed conditional on the outcome variables, as in a case-control study. In a case-control study the

binary outcome variable is fixed by stratification. In this type of study design, two random samples of sizes n_0 and n_1 are chosen from the two strata defined by the outcome variable, i.e., from the subsets of the population with $Y = 0$ and $Y = 1$, respectively. This is called retrospective sampling. Denote by $X_{0,1}, \dots, X_{0,n_0}$ the observed covariates from the control group, and by $X_{1,1}, \dots, X_{1,n_1}$ those from the case group. We aim to estimate the parameters γ and β using the two samples $X_{0,1}, \dots, X_{0,n_0}$ and $X_{1,1}, \dots, X_{1,n_1}$. We show that the logistic model (5.5) writes in the form of the model (5.1). So, let $f_X(\cdot)$ denotes the (density) function of the covariates X , and put

$$\pi := \Pr(Y = 1) = \int_{\mathbb{R}^d} \Pr(Y = 1|X = \mathbf{x}) f_X(\mathbf{x}) d\mathbf{x},$$

and assume that

$$f_i(\mathbf{x}) := f_X(\mathbf{x}|Y = i) = dP_X(\mathbf{x}|y = i)/d\mathbf{x}, \quad i = 0, 1,$$

exist and represent the conditional density functions of X given $Y = i$. It is not difficult to manipulate the case-control likelihood function to obtain a logistic regression model in which the dependent variable is the outcome variable of interest to the investigator. The key step in this development is a straightforward application of the Bayes theorem, that yields

$$f_1(\mathbf{x}) = \frac{\exp(\gamma + \beta^\top \mathbf{x})}{(1 + \exp(\gamma + \beta^\top \mathbf{x}))\pi} f_X(\mathbf{x}) \quad \text{and} \quad f_0(\mathbf{x}) = \frac{1}{(1 + \exp(\gamma + \beta^\top \mathbf{x}))(1 - \pi)} f_X(\mathbf{x}).$$

So,

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{1 - \pi}{\pi} \exp(\gamma + \beta^\top \mathbf{x}) = \exp \left\{ \gamma + \log \left(\frac{1 - \pi}{\pi} \right) + \beta^\top \mathbf{x} \right\} =: \exp(\alpha + \beta^\top \mathbf{x}),$$

where $\alpha := \gamma + \log \left(\frac{1 - \pi}{\pi} \right)$. Thus, model (5.5) is equivalent to the following two-sample semiparametric model

$$\begin{aligned} X_{0,1}, \dots, X_{0,n_0} &\sim f_X(\mathbf{x}|Y = 0) =: f_0(\mathbf{x}), \\ X_{1,1}, \dots, X_{1,n_1} &\sim f_X(\mathbf{x}|Y = 1) =: f_1(\mathbf{x}) = \exp(\alpha + \beta^\top \mathbf{x}) f_0(\mathbf{x}). \end{aligned} \quad (5.6)$$

More generally, we can consider the multiplicative-intercept risk model, i.e.,

$$\Pr(Y = 1|X = \mathbf{x}) = \frac{\exp(\gamma + \beta^\top \mathbf{b}(\mathbf{x}))}{1 + \exp(\gamma + \beta^\top \mathbf{b}(\mathbf{x}))}, \quad \gamma \in \mathbb{R}, \quad \beta \in \mathbb{R}^d,$$

where $\mathbf{b}(\cdot)$ is a given (basis) function from $\mathbb{R}^m$ onto $\mathbb{R}^d$. In this case, as above, we obtain the following two-sample semiparametric model

$$\begin{aligned} X_{0,1}, \dots, X_{0,n_0} &\sim f_0(\mathbf{x}) \\ X_{1,1}, \dots, X_{1,n_1} &\sim f_1(\mathbf{x}) = \exp(\alpha + \beta^\top \mathbf{b}(\mathbf{x})) f_0(\mathbf{x}). \end{aligned} \quad (5.7)$$

Hence, models (5.6) and (5.7) are particular cases of models (5.1) by taking $\frac{dP}{dx} = f_0$ and $\frac{dQ}{dx} = f_1$. Many other examples and motivation of density ratio models (5.1), as well as their rule in machine learning and binary classification, are given in the book by Sugiyama *et al.* (2012a). For models (5.1), when the two samples $X_{0,1}, \dots, X_{0,n_0}$ and $X_{1,1}, \dots, X_{1,n_1}$ are independent, Qin (1998) presented an estimation procedure of θ_T based on the empirical likelihood approach (see Owen (1988, 1990, 2001)), using the likelihood of the independent variables $X_{0,1}, \dots, X_{0,n_0}, X_{1,1}, \dots, X_{1,n_1}$. However, an important special case of the case-control study is the matched (or paired) study. In this design, subjects are stratified on the basis of variables believed to be associated with the outcome (an example of stratification variable is the age for each of the individuals in the survey). Within each stratum, samples of cases ($Y = 1$) and controls ($Y = 0$) are chosen; the most common matched design includes one case and one control per stratum and is thus referred as 1-1 matched study. The rest of the present Chapter is organized as follows: we end this Section recalling the estimation method proposed by Qin (1998). In Section 2, we show that the irregularity problem of the profile empirical likelihood can be adjusted in the context of model (5.1). We

next give a regularized version of the profile empirical likelihood using duality techniques. A statistical test, for the null hypothesis $\mathcal{H}_0 : Q = P$, is then proposed. An other point of view at the test statistic is given using the convex conjugate expressions of the φ -divergences. In Section 3, we study the asymptotic behavior of the proposed test statistic under the null and the alternative hypotheses with independent samples, and we give an approximation to the power function which leads to an approximation to the sample size $n = n_0 + n_1$ guaranteeing a desired power for a given alternative. Concluding remarks and possible developments are presented in Section 4. In the sequel, we sometimes write Pf instead of $\int f(x) dP(x)$ for any function f and any measure P .

5.1.3 The profile EL and its irregularity under the null hypothesis $\mathcal{H}_0 : Q = P$

In the present setting, the estimation method proposed by [Qin \(1998\)](#), which is based on the empirical likelihood approach (see [Owen \(1988, 1990, 2001\)](#)), can be summarized as follows. For any $\boldsymbol{\theta} \in \Theta$, the empirical likelihood of the two combined samples, $X_{0,1}, \dots, X_{0,n_0} \sim P$ and $X_{1,1}, \dots, X_{1,n_1} \sim Q$, if they are independent, is $L(\boldsymbol{\theta}, p, q) := \prod_{i=1}^{n_0} p(X_{0,i}) \prod_{i=1}^{n_1} q(X_{1,i})$. Since $q(\mathbf{x}) = r(\boldsymbol{\theta}, \mathbf{x})p(\mathbf{x})$, then $L(\boldsymbol{\theta}, p, q)$ can be written as $L(\boldsymbol{\theta}, p, q) = \prod_{k=0}^1 \prod_{i=1}^{n_k} p(X_{k,i}) \prod_{i=1}^{n_1} r(\boldsymbol{\theta}, X_{1,i})$. For brevity, we write $p_{k,i}$ instead of $p(X_{k,i})$, and put $\mathbf{p} := (p_{0,1}, \dots, p_{0,n_0}, p_{1,1}, \dots, p_{1,n_1})^\top \in \mathbb{R}^n$, where $n := n_0 + n_1$ is the total sample size. Hence, the log-likelihood is $l(\boldsymbol{\theta}, \mathbf{p}) := \sum_{k=0}^1 \sum_{i=1}^{n_k} \log p_{k,i} + \sum_{i=1}^{n_1} \log [r(\boldsymbol{\theta}, X_{1,i})]$. The profile log-likelihood (in $\boldsymbol{\theta}$) is then

$$\ell(\boldsymbol{\theta}) := \sup_{\mathbf{p} \in \mathcal{C}_\boldsymbol{\theta}} l(\boldsymbol{\theta}, \mathbf{p}), \quad (5.8)$$

where $\mathbf{p}$ is constrained to the set

$$\mathcal{C}_\boldsymbol{\theta} := \left\{ \mathbf{p} \in \mathbb{R}_+^{*n} \text{ such that } \sum_{k=0}^1 \sum_{i=1}^{n_k} p_{k,i} = 1 \text{ and } \sum_{k=0}^1 \sum_{i=1}^{n_k} p_{k,i} [r(\boldsymbol{\theta}, X_{k,i}) - 1] = 0 \right\}.$$

The EL estimate of $\boldsymbol{\theta}_T$, proposed by [Qin \(1998\)](#), is

$$\tilde{\boldsymbol{\theta}} := \arg \sup_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}). \quad (5.9)$$

[Qin \(1998\)](#) has proved that the estimate $\tilde{\boldsymbol{\theta}}$ is optimal, in the sense of [Godambe \(1960\)](#), in the class of all estimates obtained by unbiased estimating functions, when $r(\boldsymbol{\theta}, \mathbf{x})$ is of the exponential form $\exp\{\alpha + \boldsymbol{\beta}^\top \mathbf{b}(\mathbf{x})\}$ and α is unknown; see [Qin \(1998\)](#) Theorem 3. For a given $\boldsymbol{\theta} \in \Theta$, the profile log-likelihood $\ell(\boldsymbol{\theta})$ is well defined (and finite) if and only if

$$\text{there exists } \mathbf{p} \in \mathcal{C}_\boldsymbol{\theta} \text{ such that } |l(\boldsymbol{\theta}, \mathbf{p})| < \infty. \quad (5.10)$$

This condition means that 0 is inside the convex hull generated by the points $[r(\boldsymbol{\theta}, X_{k,i}) - 1]$, $k = 0, 1, i = 1, \dots, n_k$, i.e.,

$$\min_{k,i} [r(\boldsymbol{\theta}, X_{k,i}) - 1] < 0 < \max_{k,i} [r(\boldsymbol{\theta}, X_{k,i}) - 1]. \quad (5.11)$$

So, when $\boldsymbol{\beta}_T \neq 0$ and if P is not degenerate, using similar arguments to those in [Zou et al. \(2002\)](#) Theorem 1, we can show that there exists a neighborhood of $\boldsymbol{\theta}_T$, say $N(\boldsymbol{\theta}_T)$, such that for all $\boldsymbol{\theta} \in N(\boldsymbol{\theta}_T)$, the assumption (5.10) holds as $n_0 \rightarrow \infty$. Hence, $\boldsymbol{\theta} \in N(\boldsymbol{\theta}_T) \mapsto \ell(\boldsymbol{\theta})$ is well defined for n_0 sufficiently large. However, when $Q = P$ (i.e., when $\boldsymbol{\beta}_T = 0$), then obviously the set $\mathcal{C}_\boldsymbol{\theta}$ is empty for all $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^\top)^\top \in \Theta$ with $\alpha \neq 0$ and $\boldsymbol{\beta} = 0$. So, when $Q = P$ (i.e., when $\boldsymbol{\theta}_T = 0$), there exists no neighborhood $N(\boldsymbol{\theta}_T)$ of $\boldsymbol{\theta}_T$ such that the profile empirical log-likelihood function $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{\theta})$ is well defined on all $N(\boldsymbol{\theta}_T)$. Consequently the estimate $\tilde{\boldsymbol{\theta}}$ is not well defined also in this case. In the following Section, we will show, using some arguments of duality theory, that this problem can be adjusted in the context of the model (5.1).

5.2 Adjustment of the profile empirical likelihood

If the assumption (5.10) holds, then $\ell(\boldsymbol{\theta})$ is finite, and the unique “optimal solution”, say $\bar{\mathbf{p}}$ (i.e., the value of $\mathbf{p}$ which yields the supremum in (5.8)), as an explicit expression of (5.8) can be derived by a Lagrange multiplier argument and the Kuhn-Tucker Theorem (see e.g. Rockafellar (1970) Section 28). In fact, the “dual” problem associated to the “primal” one (i.e., the optimization problem (5.8)) can be written as follows

$$\inf_{\lambda_0, \lambda_1 \in \mathbb{R}} \left\{ -\lambda_0 - n - \sum_{k=0}^1 \sum_{i=1}^{n_k} \overline{\log}(-\lambda_0 - \lambda_1 [r(\boldsymbol{\theta}, X_{k,i}) - 1]) + \sum_{i=1}^{n_1} \log r(\boldsymbol{\theta}, X_{1,i}) \right\},$$

where $\overline{\log}(\cdot)$ is the function defined on $\mathbb{R}$ by $\overline{\log}(x) = \log x$ if $x > 0$ and $\overline{\log}(x) = -\infty$ elsewhere. So, by the Kuhn-Tucker Theorem, under condition (5.10), the infimum in (5.12) is attained, and the following equality

$$\sup_{\mathbf{p} \in \mathcal{C}_{\boldsymbol{\theta}}} l(\boldsymbol{\theta}, \mathbf{p}) = \inf_{\lambda_0, \lambda_1 \in \mathbb{R}} \left\{ -\lambda_0 - n - \sum_{k=0}^1 \sum_{i=1}^{n_k} \overline{\log}(-\lambda_0 - \lambda_1 [r(\boldsymbol{\theta}, X_{k,i}) - 1]) + \sum_{i=1}^{n_1} \log r(\boldsymbol{\theta}, X_{1,i}) \right\}, \quad (5.12)$$

holds. The “dual” optimal solution, say $(\bar{\lambda}_0, \bar{\lambda}_1)$, (i.e., the argument infimum in (5.12)) can be derived by differentiation. Furthermore, $(\bar{\lambda}_0, \bar{\lambda}_1)$ and $\bar{\mathbf{p}}$ satisfy

$$\bar{p}_{k,i} = \frac{1}{-\bar{\lambda}_0 - \bar{\lambda}_1 [r(\boldsymbol{\theta}, X_{k,i}) - 1]}, \quad k = 0, 1, i = 1, \dots, n_k.$$

Hence, using the fact that $\bar{\mathbf{p}}$ satisfies the constraints, we obtain $\bar{\lambda}_0 = -n$ and $\bar{\lambda}_1$ is the solution (in λ_1) of the equality

$$\sum_{k=0}^1 \sum_{i=1}^{n_k} \frac{r(\boldsymbol{\theta}, X_{k,i}) - 1}{n - \lambda_1 [r(\boldsymbol{\theta}, X_{k,i}) - 1]} = 0.$$

Finally, under condition (5.10), the equality

$$\ell(\boldsymbol{\theta}) := \sup_{\mathbf{p} \in \mathcal{C}_{\boldsymbol{\theta}}} l(\boldsymbol{\theta}, \mathbf{p}) = \inf_{\lambda \in \mathbb{R}} \left\{ -\sum_{k=0}^1 \sum_{i=1}^{n_k} \overline{\log} [n(1 + \lambda [r(\boldsymbol{\theta}, X_{k,i}) - 1])] + \sum_{i=1}^{n_1} \log r(\boldsymbol{\theta}, X_{1,i}) \right\} \quad (5.13)$$

holds with finite values, and the unique optimal solution $\bar{\mathbf{p}} := (\bar{p}_{k,i})$, $k = 0, 1$, $i = 1, \dots, n_k$, exists and it is given by

$$\bar{p}_{k,i} = \frac{1}{n} \frac{1}{1 + \bar{\lambda} [r(\boldsymbol{\theta}, X_{k,i}) - 1]}, \quad \forall k, \forall i, \quad (5.14)$$

where $\bar{\lambda}$ is the unique dual optimal solution in (5.13). It is the solution (in λ) of the equation

$$\sum_{k=0}^1 \sum_{i=1}^{n_k} \frac{1}{n} \frac{r(\boldsymbol{\theta}, X_{k,i}) - 1}{1 + \lambda [r(\boldsymbol{\theta}, X_{k,i}) - 1]} = 0. \quad (5.15)$$

Hence, the EL estimate $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_T$ can be written as follows

$$\tilde{\boldsymbol{\theta}} = \arg \sup_{\boldsymbol{\theta} \in \Theta} \inf_{\lambda \in \mathbb{R}} \left\{ -\sum_{k=0}^1 \sum_{i=1}^{n_k} \overline{\log} [n(1 + \lambda [r(\boldsymbol{\theta}, X_{k,i}) - 1])] + \sum_{i=1}^{n_1} \log r(\boldsymbol{\theta}, X_{1,i}) \right\}. \quad (5.16)$$

By differentiation with respect to α and λ , interestingly, one can see by simple calculus that the Lagrange multiplier $\bar{\lambda}$ in (5.16) has the explicit solution $\bar{\lambda}(\tilde{\boldsymbol{\theta}}) = \frac{n_1}{n}$, which does not depend on the data. Hence, the value of the log-likelihood (5.13) in $\tilde{\boldsymbol{\theta}}$ can be simply written under the following full parametric expression

$$\ell(\tilde{\boldsymbol{\theta}}) = -n \log n - \sum_{k=0}^1 \sum_{i=1}^{n_k} \log \left(1 + \frac{n_1}{n} [r(\tilde{\boldsymbol{\theta}}, X_{k,i}) - 1] \right) + \sum_{i=1}^{n_1} \log r(\tilde{\boldsymbol{\theta}}, X_{1,i}). \quad (5.17)$$

Therefore, the EL estimate $\tilde{\boldsymbol{\theta}}$ can be written as

$$\tilde{\boldsymbol{\theta}} = \arg \sup_{\boldsymbol{\theta} \in \Theta} \left\{ -n \log n - \sum_{k=0}^1 \sum_{i=1}^{n_k} \log \left(\hat{\rho}_0 + \hat{\rho}_1 r(\tilde{\boldsymbol{\theta}}, X_{k,i}) \right) + \sum_{i=1}^{n_1} \log r(\tilde{\boldsymbol{\theta}}, X_{1,i}) \right\}, \quad (5.18)$$

where $\hat{\rho}_0 := n_0/n$ and $\hat{\rho}_1 := n_1/n$. In summary, under the null hypothesis $\mathcal{H}_0 : Q = P$, i.e., when $\beta_T = 0$, the profile log-likelihood $\ell(\boldsymbol{\theta})$ is not defined for some $\boldsymbol{\theta}$ (see Section 1.3). So, in view of (5.17) and (5.18), we propose to consider, instead of $\ell(\boldsymbol{\theta})$, the “dual form”:

$$\ell_d(\boldsymbol{\theta}) := \left\{ -n \log n - \sum_{k=0}^1 \sum_{i=1}^{n_k} \log \left(\hat{\rho}_0 + \hat{\rho}_1 r(\tilde{\boldsymbol{\theta}}, X_{k,i}) \right) + \sum_{i=1}^{n_1} \log r(\tilde{\boldsymbol{\theta}}, X_{1,i}) \right\}, \quad (5.19)$$

which is well defined for all $\boldsymbol{\theta} \in \Theta$ regardless of the null hypothesis $\mathcal{H}_0 : Q = P$, and to redefine the EL estimate $\tilde{\boldsymbol{\theta}}$ as follows

$$\hat{\boldsymbol{\theta}} := \arg \sup_{\boldsymbol{\theta} \in \Theta} \ell_d(\boldsymbol{\theta}), \quad (5.20)$$

Note that, under condition (5.10), we have $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}$ and $\ell_d(\hat{\boldsymbol{\theta}}) = \ell(\tilde{\boldsymbol{\theta}})$. Now, we give an interpretation to the “dual form”

$$S_n := 2W_d(\hat{\boldsymbol{\theta}}) := 2 \left[\sup_{\boldsymbol{\theta} \in \Theta} \ell_d(\boldsymbol{\theta}) + n \log n \right] \quad (5.21)$$

of the empirical likelihood ratio statistic

$$2W(\tilde{\boldsymbol{\theta}}) := 2 \left[\sup_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}) + n \log n \right] \quad (5.22)$$

(associated to the null hypothesis $\mathcal{H}_0 : Q = P$). First, let Q_{n_1} and P_{n_0} be, respectively, the empirical measures associated to the samples $X_{1,1}, \dots, X_{1,n_1}$ and $X_{0,1}, \dots, X_{0,n_0}$, namely,

$$Q_{n_1}(\cdot) := \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{X_{1,i}}(\cdot) \quad \text{and} \quad P_{n_0}(\cdot) := \frac{1}{n_0} \sum_{i=1}^{n_0} \delta_{X_{0,i}}(\cdot),$$

with $\delta_{\mathbf{x}}(\cdot)$ denotes the Dirac measure at point $\mathbf{x}$, for all $\mathbf{x}$. By simple calculus, we can show that the proposed statistic (5.21) can be written under the form

$$S_n = 2n \sup_{\boldsymbol{\theta} \in \Theta} \left\{ \int_{\mathbb{R}^m} f(\boldsymbol{\theta}, \hat{\rho}, \mathbf{x}) dQ_{n_1}(\mathbf{x}) - \int_{\mathbb{R}^m} g(\boldsymbol{\theta}, \hat{\rho}, \mathbf{x}) dP_{n_0}(\mathbf{x}) \right\}, \quad (5.23)$$

where

$$f(\boldsymbol{\theta}, \hat{\rho}, \mathbf{x}) = \hat{\rho}_1 \log(\hat{\rho}_0 + \hat{\rho}_1 r(\boldsymbol{\theta}, \mathbf{x})) - \hat{\rho}_1 \log r(\boldsymbol{\theta}, \mathbf{x}), \quad (5.24)$$

$$g(\boldsymbol{\theta}, \hat{\rho}, \mathbf{x}) = \hat{\rho}_0 \log(\hat{\rho}_0 + \hat{\rho}_1 r(\boldsymbol{\theta}, \mathbf{x})), \quad (5.25)$$

with $\hat{\rho} := (\hat{\rho}_0, \hat{\rho}_1)^\top = (n_0/n, n_1/n)^\top$. Assume that, when the total sample size $n := n_0 + n_1 \rightarrow \infty$, then $\hat{\rho} \rightarrow \rho := (\rho_0, \rho_1)^\top$ for some positive real numbers ρ_0 and ρ_1 in $]0, 1[$. Obviously, we have $\rho_0 + \rho_1 = 1$. Therefore, in (5.23), $2n$ is a normalizing term, and the second term can be seen as the “empirical” version of

$$\sup_{\boldsymbol{\theta} \in \Theta} \left\{ \int_{\mathbb{R}^m} f(\boldsymbol{\theta}, \rho, \mathbf{x}) dQ(\mathbf{x}) - \int_{\mathbb{R}^m} g(\boldsymbol{\theta}, \rho, \mathbf{x}) dP(\mathbf{x}) \right\}. \quad (5.26)$$

On the other hand, applying the “dual representation of φ -divergences” (see Theorem 2.1 in Keziou (2003a) and Theorem 4.4 in Broniatowski and Keziou (2006)), for the class of functions

$\mathcal{F} := \{\mathbf{x} \in \mathbb{R}^m \mapsto \varphi'_\rho(r(\boldsymbol{\theta}, \mathbf{x})); \boldsymbol{\theta} \in \Theta\}$, we obtain the equality

$$\sup_{\boldsymbol{\theta} \in \Theta} \left\{ \int_{\mathbb{R}^m} f(\boldsymbol{\theta}, \rho, \mathbf{x}) dQ(\mathbf{x}) - \int_{\mathbb{R}^m} g(\boldsymbol{\theta}, \rho, \mathbf{x}) dP(\mathbf{x}) \right\} = \int_{\mathbb{R}^m} \varphi_\rho \left(\frac{dQ}{dP}(\mathbf{x}) \right) dP(\mathbf{x}) = D_{\varphi_\rho}(Q, P), \quad (5.27)$$

where $\varphi_\rho(\cdot)$ is the particular real strictly convex function, defined $\forall x \in \mathbb{R}_+$, by

$$\varphi_\rho(x) := \rho_1 \varphi(x) - \varphi(\rho_0 + \rho_1 x) \quad \text{with} \quad \varphi(x) = x \log x - x + 1. \quad (5.28)$$

Note that the divergence $D_{\varphi_\rho}(\cdot, \cdot)$, associated to the above function $\varphi_\rho(\cdot)$ in (5.28), is a member of the class of φ -divergences (5.4), since $\varphi_\rho(\cdot)$ is convex on $\mathbb{R}$ with values in $[0, +\infty]$ and satisfying $\varphi_\rho(1) = 0$. We propose to call it “weighted” Kullback-Leiber divergence. So, by (5.23), (5.26) and (5.27), $2n^{-1}W_d(\hat{\theta})$ can be seen as an empirical estimate (which we denote $\hat{D}_{\varphi_\rho}(Q, P)$) of the weighted Kullback-Leiber divergence $D_{\varphi_\rho}(Q, P)$, between Q and P . Since $D_{\varphi_\rho}(Q, P)$ is nonnegative and takes value 0 only when $Q = P$, it is reasonable to perform a test that rejects the null hypothesis $\mathcal{H}_0 : Q = P$ when the statistic

$$S_n = 2n\hat{D}_{\varphi_\rho}(Q, P) = 2n \sup_{\theta \in \Theta} \left\{ \int_{\mathbb{R}^m} f(\theta, \hat{\rho}, \mathbf{x}) dQ_{n_1}(\mathbf{x}) - \int_{\mathbb{R}^m} g(\theta, \hat{\rho}, \mathbf{x}) dP_{n_0}(\mathbf{x}) \right\}, \quad (5.29)$$

see (5.21) and (5.23), takes large values. The DEL estimate $\hat{\theta}$ of θ_T , see (5.20), can be written then as follows

$$\hat{\theta} = \arg \sup_{\theta \in \Theta} \left\{ \int_{\mathbb{R}^m} f(\theta, \hat{\rho}, \mathbf{x}) dQ_{n_1}(\mathbf{x}) - \int_{\mathbb{R}^m} g(\theta, \hat{\rho}, \mathbf{x}) dP_{n_0}(\mathbf{x}) \right\}. \quad (5.30)$$

On the other hand, by Theorem 2.1 in Keziou (2003a) and Theorem 4.4 in Broniatowski and Keziou (2006), we can prove that the supremum in (5.26) is unique and reached at $\theta = \theta_T$. This indicates that the estimate $\hat{\theta}$ of θ_T may converge (as M-estimate) to θ_T even when the samples are paired.

5.3 Asymptotic behavior of the estimate and test statistic under the null and the alternative hypotheses, and approximation of the power function

In this Section, for independent samples, we give the asymptotic properties of the estimate $\hat{\theta}$ (of the parameter θ_T) and the test statistic (5.29) both under the null and the alternative hypotheses. As an application, we obtain approximation of the power function for a given alternative. For brevity, in the following, we sometimes omit the subscript ρ in $f(\theta, \rho, \mathbf{x})$ and $g(\theta, \rho, \mathbf{x})$. In all the sequel, $f'(\theta, \mathbf{x})$ and $f''(\theta, \mathbf{x})$ denote, respectively, the gradient and the Hessian of f at the point θ , for all $\mathbf{x}$ and any function f . We use the notation $|\cdot|$ for the Euclidean norm. Let $\hat{\rho}_1 := n_1/n$ and $\hat{\rho}_0 := n_0/n$, and assume that $\hat{\rho}_1 \rightarrow \rho_1 > 0$ and $\hat{\rho}_0 \rightarrow \rho_0 > 0$ when $n := n_0 + n_1 \rightarrow \infty$. Denote also

$$\ell_\varphi(\theta) := 2n \left\{ \int_{\mathbb{R}^m} f(\theta, \hat{\rho}, \mathbf{x}) dQ_{n_1}(\mathbf{x}) - \int_{\mathbb{R}^m} g(\theta, \hat{\rho}, \mathbf{x}) dP_{n_0}(\mathbf{x}) \right\}.$$

We give our results under the following assumptions

- (A.1) There exists a neighborhood $N(\theta_T)$ of θ_T such that the third order partial derivative functions $\{\mathbf{x} \mapsto (\partial^3/\partial\theta_i\partial\theta_j\partial\theta_k)f(\theta, \mathbf{x}); \theta \in N(\theta_T)\}$ (resp. $\{\mathbf{x} \mapsto (\partial^3/\partial\theta_i\partial\theta_j\partial\theta_k)g(\theta, \mathbf{x}); \theta \in N(\theta_T)\}$) are dominated by some function Q -integrable (resp. some function P -integrable);
- (A.2) The integrals $Q|f'(\theta_T, \mathbf{x})|^2$, $P|g'(\theta_T, \mathbf{x})|^2$, $Q|f''(\theta_T, \mathbf{x})|$ and $P|g''(\theta_T, \mathbf{x})|$ are finite, and the matrix $U_\varphi(\theta_T) := -[Qf''(\theta_T, \mathbf{x}) - Pg''(\theta_T, \mathbf{x})]$ is non singular.

Theorem 5.1. *Assume that assumptions (A.1-2) hold.*

- (a) *Let $B(\theta_T, n^{-1/3}) := \{\theta \in \Theta; |\theta - \theta_T| \leq n^{-1/3}\}$. Then as $n \rightarrow \infty$, with probability one, $\ell_\varphi(\theta)$ attains its maximum value at some point $\hat{\theta}$ in the interior of the ball $B(\theta_T, n^{-1/3})$, and the estimate $\hat{\theta}$ satisfies $\ell'_\varphi(\hat{\theta}) = 0$.*

(b) $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T)$ converges in distribution to a centered multivariate normal random variable with covariance matrix $M_\varphi(\boldsymbol{\theta}_T) = U_\varphi(\boldsymbol{\theta}_T)^{-1} - D$, where $D = \begin{bmatrix} \rho_0^{-1} + \rho_1^{-1} & \mathbf{0}_d^\top \\ \mathbf{0}_d & \mathbf{0}_d \mathbf{0}_d^\top \end{bmatrix}$. If $Q = P$, then the limiting covariance matrix can be written as

$$M_\varphi(\mathbf{0}) = \rho_0^{-1} \rho_1^{-1} \begin{bmatrix} 1 & P\mathbf{b}(\mathbf{x})^\top \\ P\mathbf{b}(\mathbf{x}) & P(\mathbf{b}(\mathbf{x})\mathbf{b}(\mathbf{x})^\top) \end{bmatrix}^{-1} - \begin{bmatrix} \rho_0^{-1} + \rho_1^{-1} & \mathbf{0}_d^\top \\ \mathbf{0}_d & \mathbf{0}_d \mathbf{0}_d^\top \end{bmatrix}. \quad (5.31)$$

(c) Under the null hypothesis $\mathcal{H}_0 : Q = P$, the statistic S_n converges in distribution to a χ^2 -random variable with d degrees of freedom.

In order to give the asymptotic properties of the test statistic S_n under the alternative hypothesis $\mathcal{H}_1 : Q \neq P$, we need the following additional assumption pertaining to the functions f and g defined in (5.24) and (5.25).

(A.3) The integrals $Q(f(\boldsymbol{\theta}_T, \mathbf{x})^2)$ and $P(g(\boldsymbol{\theta}_T, \mathbf{x})^2)$ are finite.

Theorem 5.2. Assume that assumptions (A.1-3) hold. Then, under the alternative hypothesis $\mathcal{H}_1 : Q \neq P$, we have $\sqrt{n}[(2n)^{-1}S_n - D_{\varphi_\rho}(Q, P)]$ converges in distribution to a centered normal random variable with variance $\sigma^2(\boldsymbol{\theta}_T) := \rho_1^{-1}V_Q(f(\boldsymbol{\theta}_T, \mathbf{x})) + \rho_0^{-1}V_P(g(\boldsymbol{\theta}_T, \mathbf{x}))$.

Remark 5.1. Using Theorem 5.1 part (c), we propose to reject the null hypothesis $\mathcal{H}_0 : Q = P$ if $S_n > \chi_d^2(1 - \alpha)$, where $\chi_d^2(1 - \alpha)$ is the $(1 - \alpha)$ -quantile of the χ^2 distribution with d degrees of freedom. This leads to a test asymptotically of level α . The asymptotic result in Theorem 5.2 allows to give approximation of the power function for a given alternative: for a given $\boldsymbol{\beta}_T \neq \mathbf{0}$, we obtain for the power function $\beta(\cdot) : \boldsymbol{\theta}_T \mapsto \beta(\boldsymbol{\theta}_T) := P_{\boldsymbol{\theta}_T}\{S_n > \chi_d^2(1 - \alpha)\}$ the following approximation $P_{\boldsymbol{\theta}_T}\{S_n > \chi_d^2(1 - \alpha)\} \approx 1 - F_N\left(\frac{\sqrt{n}}{\hat{\sigma}(\boldsymbol{\theta}_T)}[(2n)^{-1}\chi_d^2(1 - \alpha) - H_n(\boldsymbol{\theta}_T)]\right)$, where $F_N(\cdot)$ is the cumulative distribution function of a normal random variable with mean zero and variance one, $\hat{\sigma}(\boldsymbol{\theta}_T)^2 := \hat{\rho}_1^{-1}[Q_{n_1}(f(\boldsymbol{\theta}_T)^2) - (Q_{n_1}f(\boldsymbol{\theta}_T))^2] + \hat{\rho}_0^{-1}[P_{n_0}(g(\boldsymbol{\theta}_T)^2) - (P_{n_0}g(\boldsymbol{\theta}_T))^2]$, and $H_n(\boldsymbol{\theta}_T) := Q_{n_1}f(\boldsymbol{\theta}_T) - P_{n_0}g(\boldsymbol{\theta}_T)$. Note also that the power $\beta(\boldsymbol{\theta}_T)$, by the asymptotic result in Theorem 5.2, tends to one, as $n \rightarrow \infty$, under the alternative hypothesis $\mathcal{H}_1 : Q \neq P$.

5.4 Discussion and directions for future research

We have addressed the problems of estimation and test of homogeneity in semiparametric two-sample density ratio models. The profile EL poses an irregularity problem under the null hypothesis $\mathcal{H}_0$ that the two laws of the two samples are equal. We have showed that the dual form of the profile EL is well defined even under the null hypothesis, then we have proposed a test of homogeneity based on the dual form of the EL ratio statistic. We have showed, using the dual representation of φ -divergences, that the test statistic can be seen, up to some normalization term, as an estimate of a particular φ -divergence, that we have called weighted Kullback-Leibler divergence, between the two laws, and that the dual EL estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_T$ can be seen as the dual optimal solution in the dual representation of this divergence. The advantage of this new point of view is

- It allows to obtain the limiting distribution of the test statistic under the alternative hypothesis which we use to give approximation to the power function of the test;
- It suggests to generalize the test and the estimate of the parameter to a class of tests and to a class of estimates using the weighted φ_ρ -divergences associated to the subclass of real convex functions $\varphi_\rho(\cdot)$ of the form $\varphi_\rho(x) = \rho_1\varphi(x) - \varphi(\rho_0 + \rho_1x)$, where $\varphi(\cdot)$ is any convex l.s.c. function defined on $\mathbb{R}$ with values in $[0, +\infty]$ such that $\varphi(1) = 0$. The estimates of the divergence $D_{\varphi_\rho}(Q, P)$ and the parameter $\boldsymbol{\theta}_T$ will be defined as in (5.29) and (5.30), where $\rho := (\rho_0, \rho_1)^\top$ will be replaced naturally by $\hat{\rho} := (n/n_0, n/n_1)^\top$. We obtain then a large class of estimates and tests that contains

the dual empirical likelihood ones. Using the result of [Qin \(1998\)](#) Theorem 3, one can show that the DEL estimate is the optimal one (in the sense that the limiting variance is minimal) among all estimates associated to weighted φ_ρ -divergences, if the ratio $r(\cdot)$ is of the exponential form (5.1). But it is not clear if this interesting result holds if the ratio takes other form than the exponential one. It would be interesting in this case to give how to choose the divergence D_{φ_ρ} which leads to an “optimal” (in some sense) estimate or test in term of efficiency/robustness, for different models $r(\theta, x)$, not necessary of the exponential form (5.1). Note that this class of estimates is different from that proposed in [Keziou \(2003b\)](#) Chapter 2, where we have used divergences associated to convex function φ not depending on the weights of the two samples. Therefore the class of estimates introduced in [Keziou \(2003b\)](#) Chapter 2 does not contain the DEL one. [Nguyen et al. \(2010\)](#) considered also the problem of estimating $D_\varphi(Q, P)$ and the ratio $r(\cdot)$ using the same method as in [Keziou \(2003b\)](#) Chapter 2, and extend the results to the case where the ratio $r(\cdot)$ may lay in infinite-dimensional function space, but this class of estimates also does not contains the DEL one.

- In the important case of paired samples, the asymptotic results presented in Section 4 hold with some modifications for the limiting variance expressions. It would be worthwhile also to involve the problem of Bartlett correctability of the test statistics S_n , as well as model selection for the ratio $r(\theta, x)$ using the proposed estimate of the weighted Kullback-Leibler divergence. We aim also to investigate our proposed estimates using D_{φ_ρ} -divergences in the case where the ratio $r(\cdot)$ lies in infinite-dimensional space.
- Other promising problem, is the use, in a retrospective sampling context, of the proposed estimates of the ratio, to define binary classifiers and to compare their efficiency with existing methods. These developments will be reported in future communications.

Chapter 6

Statistical inference for semiparametric multi-sample density ratio models

We use our dual representations of multivariate ψ -divergences, presented in [Keziou \(2015\)](#), see also Chapter 2, to define a new family of estimates and test statistics with multiple samples ($1 + K$ samples) under semiparametric density ratio models. We show that the dual empirical likelihood estimate and test correspond to a special “weighted” ψ_ρ -multivariate divergence, depending on the parameters $\rho := (\rho_0, \dots, \rho_K)^\top$ which represent the weights of the samples. Moreover, the present approach allows obtaining the asymptotic properties of the estimates and test statistics both under the null and the alternative hypotheses. This leads to accurate approximations of the power function for any used criterion, including the dual empirical likelihood one, which is of its own interest. Moreover, the multivariate divergences context leads to a new definition of Information criterion, and by product a new model selection procedure with multiple samples under density ratio models, as well as a new procedure of multi-group classification based on the proposed estimated ratios. For more detail and proofs, we refer to [Keziou \(2015, 2016\)](#).

6.1 Introduction, notation and motivation

In this Chapter, we investigate statistical inference for multiple density ratio models (MDRM). Suppose we have $1 + K$ random samples,

$$(X_{0,1}, \dots, X_{0,n_0}), (X_{1,1}, \dots, X_{1,n_1}), \dots, (X_{K,1}, \dots, X_{K,n_K}), \quad (6.1)$$

of $1 + K$ random variables (or vectors) $X_0, X_1, \dots, X_K$, defined on a probability space $(\Omega, \mathcal{A}, P)$ with values in $\mathcal{X} = \mathbb{R}^m, m \geq 1$, and distributions $P_0, Q_1, \dots, Q_K$, respectively. The MDRM postulates that

$$\frac{dQ_k}{dP_0}(x) = r_k(\boldsymbol{\theta}_k, x), \quad \text{for all } k = 1, \dots, K, \quad (6.2)$$

for some known function $r_k(\boldsymbol{\theta}_k, x)$, of $x \in \mathbb{R}^m$, and unknown vector-valued parameters $\boldsymbol{\theta}_k \in \Theta_k \subset \mathbb{R}^{1+d_k}$. We require, for all $k = 1, \dots, K$, the identifiability of the parameter $\boldsymbol{\theta}_k$, in the sense that

$$\text{if } r_k(\boldsymbol{\theta}_k, x) = r_k(\boldsymbol{\theta}'_k, x) \quad \forall x - P_0\text{-a.s.}, \quad \text{then } \boldsymbol{\theta}_k = \boldsymbol{\theta}'_k. \quad (6.3)$$

We denote then by $\boldsymbol{\theta}_T := (\boldsymbol{\theta}_{1T}^\top, \dots, \boldsymbol{\theta}_{KT}^\top)^\top$ the unique true value of the parameter $\boldsymbol{\theta} := (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_K^\top)^\top \in \Theta_1 \times \dots \times \Theta_K$. Typical examples of such MDRM models are the following exponential ones

$$r_k(\boldsymbol{\theta}_k, x) = \exp \left\{ \boldsymbol{\theta}_k^\top \underline{\mathbf{b}}_k(x) \right\}, \quad k = 1, \dots, K, \quad (6.4)$$

for some known vector-valued (basis) function

$$\underline{\mathbf{b}}_k(x) := \left(b_{k,0}(x), \mathbf{b}_k(x)^\top \right)^\top = (b_{k,0}(x), b_{k,1}(x), \dots, b_{k,d_k}(x))^\top \in \mathbb{R}^{1+d_k}, \quad x \in \mathbb{R}^m,$$

with values in $\mathbb{R}^{1+d_k}$, and corresponding unknown vector-valued parameters $\boldsymbol{\theta}_k := (\alpha_k, \boldsymbol{\beta}_k^\top)^\top \in \Theta_k \subset \mathbb{R}^{1+d_k}$. When the exponential MDRM (6.4) is considered, for technical reasons, we will represent the vector parameter using a different parametrization as follows $\boldsymbol{\theta}^\top := (\alpha_1, \dots, \alpha_K, \boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_K^\top)$, and will denote $\boldsymbol{\theta}_T^\top := (\alpha_{1T}, \dots, \alpha_{KT}, \boldsymbol{\beta}_{1T}^\top, \dots, \boldsymbol{\beta}_{KT}^\top)$ for the true value of the parameter; Θ denotes the new domain of the parameter $\boldsymbol{\theta}$. Note that the parameter space Θ is a subset of $\mathbb{R}^{K+d}$ with $d := \sum_{k=1}^K d_k$. We require also the first component of $\mathbf{b}_k(\cdot)$, for each k , to be the constant function one, i.e., $b_{k,0}(\cdot) := \mathbb{1}_{\mathbb{R}^m}(\cdot)$, the indicator function of $\mathbb{R}^m$. This makes the corresponding first real true value α_{kT} , of the vector $\boldsymbol{\theta}_{kT} := (\alpha_{kT}, \boldsymbol{\beta}_{kT}^\top)^\top$, a normalization parameter, for all $k = 1, \dots, K$. In particular, we have $\alpha_{kT} = 0$, whenever $\boldsymbol{\beta}_{kT} = \mathbf{0}$. However, the value $\alpha_k \neq \alpha_{kT}$ is assumed to be free, for all $k = 1, \dots, K$. We will assume then, for all $k = 1, \dots, K$, that the parameter space Θ_k is such that its interior $\text{int}(\Theta_k)$ in $\mathbb{R}^{1+d_k}$ is not void. The basis function vector $\mathbf{b}_k(\cdot)$ in (6.4) may depend on k , with different dimensions $d_1, \dots, d_K$. Observe also that, the distribution P_0 is unspecified. Only the density ratios are specified through parametric functions. The components of $\mathbf{b}_k(\cdot)$ in (6.4) may be chosen, for instance, to be the first $1 + d_k$ elements of basis functions. Many other interesting examples of density ratio models, are given in the book by Sugiyama *et al.* (2012a), for the two-sample case, in particular, in large dimension (i.e., when the dimension m of the variable space is large). Consider the test problem of the null hypothesis

$$\mathcal{H}_0 : Q_1 = \dots = Q_K = P_0 \quad \text{against the alternative} \quad \mathcal{H}_1 : \exists Q_k \neq P_0, \quad (6.5)$$

which is equivalent, in the context of the exponential MDRM, to testing

$$\mathcal{H}_0 : \boldsymbol{\beta}_{1T} = \dots = \boldsymbol{\beta}_{KT} = \mathbf{0} \quad \text{against the alternative} \quad \mathcal{H}_1 : \exists \boldsymbol{\beta}_{kT} \neq \mathbf{0}, \quad (6.6)$$

since $\boldsymbol{\beta}_{kT} = \mathbf{0}$ implies $\alpha_{kT} = 0$, for all $k = 1, \dots, K$. In view of the basic property (2.3), in order to perform a statistical test of the above hypotheses, we can use an estimate of $D_\psi(Q_1, \dots, Q_K; P_0)$, and reject the null hypothesis $\mathcal{H}_0$ when the obtained estimate takes large value. The plug-in estimate of $D_\psi(Q_1, \dots, Q_K; P_0)$ is given by

$$\tilde{D}_\psi(Q_1, \dots, Q_K; P_0) := D_\psi(\hat{Q}_1, \dots, \hat{Q}_K; \hat{P}_0), \quad (6.7)$$

obtained from the definition (2.2) of the ψ -divergence by replacing the distribution $P_0, Q_1, \dots, Q_K$ by their corresponding empirical measures, namely,

$$\hat{P}_0(\cdot) := \frac{1}{n_0} \sum_{i=1}^{n_0} \delta_{X_{0,i}}(\cdot), \hat{Q}_1(\cdot) := \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{X_{1,i}}(\cdot), \dots, \hat{Q}_K(\cdot) := \frac{1}{n_K} \sum_{i=1}^{n_K} \delta_{X_{K,i}}(\cdot), \quad (6.8)$$

associated to the multiple samples given in (6.1). Unfortunately, the above estimate is not generally well defined, since, the absolute continuity condition is not satisfied (even for large sample sizes, if the distributions are continuous, or discrete with unbounded support). Moreover, the estimator does not make use of the semiparametric form of the MDRM. In the following section, we define, through our convex conjugate representation of multivariate ψ -divergences (2.25), a new family of estimates and test statistics in the MDRM context.

6.2 Inference with multiple samples in MDRM through conjugate representation of ψ -divergences

Let us introduce the following notations for the measure and the density ratio vectors

$$\mathbf{Q} := (Q_1, \dots, Q_K) \quad \text{and} \quad \mathbf{r}(\boldsymbol{\theta}, x) := (r_1(\boldsymbol{\theta}, x), \dots, r_K(\boldsymbol{\theta}, x))^\top. \quad (6.9)$$

Here, we consider the Borel σ -field $(\mathcal{X}, \mathcal{B}) = (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$. Taking into account the form of the MDRM (6.2), and choosing the class of functions

$$\mathcal{F} := \prod_{k=1}^K \mathcal{F}_k := \prod_{k=1}^K \{x \in \mathbb{R}^m \mapsto \psi'_{x_k}(\mathbf{r}(\boldsymbol{\theta}, x)); \boldsymbol{\theta} \in \Theta\}, \quad (6.10)$$

we obtain from Theorem 2.3, the following dual formula

$$D_\psi(\mathbf{Q}; P_0) = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{k=1}^K \int_{\mathbb{R}^m} \psi'_{x_k}(\mathbf{r}(\boldsymbol{\theta}, x)) dQ_k(x) - \int_{\mathbb{R}^m} \psi^*(\psi'_{x_1}(\mathbf{r}(\boldsymbol{\theta}, x)), \dots, \psi'_{x_K}(\mathbf{r}(\boldsymbol{\theta}, x))) dP_0(x) \right\}. \quad (6.11)$$

Furthermore, the supremum in the above display is unique and achieved in $\boldsymbol{\theta}_T$, the true value of the parameter $\boldsymbol{\theta}$, i.e.,

$$\boldsymbol{\theta}_T = \operatorname{argsup}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{k=1}^K \int_{\mathbb{R}^m} \psi'_{x_k}(\mathbf{r}(\boldsymbol{\theta}, x)) dQ_k(x) - \int_{\mathbb{R}^m} \psi^*(\psi'_{x_1}(\mathbf{r}(\boldsymbol{\theta}, x)), \dots, \psi'_{x_K}(\mathbf{r}(\boldsymbol{\theta}, x))) dP_0(x) \right\}. \quad (6.12)$$

The uniqueness of the dual solution $\boldsymbol{\theta}_T$ follows from the uniqueness of the dual solution $\bar{\mathbf{f}}(\cdot)$ in Theorem 2.3 and the identifiability assumption (6.3). Furthermore, we have, for the function in the second integral in (6.11) and (6.12), the explicit expression

$$\psi^*(\psi'_{x_1}(\mathbf{r}(\boldsymbol{\theta}, x)), \dots, \psi'_{x_K}(\mathbf{r}(\boldsymbol{\theta}, x))) = \sum_{k=1}^K r_k(\boldsymbol{\theta}_k, x) \psi'_{x_k}(\mathbf{r}(\boldsymbol{\theta}, x)) - \psi(\mathbf{r}(\boldsymbol{\theta}, x)). \quad (6.13)$$

For brevity, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, denote by $f_k(\boldsymbol{\theta}, \cdot)$ and $g(\boldsymbol{\theta}, \cdot)$, respectively, the real valued functions

$$f_k(\boldsymbol{\theta}, \cdot) : x \in \mathbb{R}^m \mapsto f_k(\boldsymbol{\theta}, x) := \psi'_{x_k}(\mathbf{r}(\boldsymbol{\theta}, x)), \quad k = 1, \dots, K, \quad (6.14)$$

and

$$g(\boldsymbol{\theta}, \cdot) : x \in \mathbb{R}^m \mapsto g(\boldsymbol{\theta}, x) := \sum_{k=1}^K r_k(\boldsymbol{\theta}_k, x) \psi'_{x_k}(\mathbf{r}(\boldsymbol{\theta}, x)) - \psi(\mathbf{r}(\boldsymbol{\theta}, x)). \quad (6.15)$$

We may write also $Qf(\boldsymbol{\theta}, x)$, instead of $\int_{\mathcal{X}} f(\boldsymbol{\theta}, x) dQ(x)$, for any function $f(\boldsymbol{\theta}, x)$ and any measure Q . Whence, the above formulas (6.11) and (6.12) can be written as follows

$$\begin{aligned} D_\psi(\mathbf{Q}; P_0) &= \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{k=1}^K Q_k f_k(\boldsymbol{\theta}, x) - P_0 g(\boldsymbol{\theta}, x) \right\} \\ &= \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{k=1}^K E_{Q_k}(f_k(\boldsymbol{\theta}, X_k)) - E_{P_0}(g(\boldsymbol{\theta}, X_0)) \right\} =: \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell_\psi(\boldsymbol{\theta}) \end{aligned} \quad (6.16)$$

and

$$\boldsymbol{\theta}_T = \operatorname{argsup}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{k=1}^K Q_k f_k(\boldsymbol{\theta}, x) - P_0 g(\boldsymbol{\theta}, x) \right\} =: \operatorname{argsup}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell_\psi(\boldsymbol{\theta}), \quad (6.17)$$

where

$$\ell_\psi(\boldsymbol{\theta}) := \sum_{k=1}^K Q_k f_k(\boldsymbol{\theta}, x) - P_0 g(\boldsymbol{\theta}, x). \quad (6.18)$$

Therefore, by replacing, in the above dual formula, the distributions $P_0, Q_1, \dots, Q_K$ by their empirical ones (6.8), we obtain the following “dual” plug-in estimate of the ψ -divergence $D_\psi(\mathbf{Q}; P_0)$

$$\begin{aligned} \widehat{D}_\psi(\mathbf{Q}; P_0) &:= \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{k=1}^K \widehat{Q}_k f_k(\boldsymbol{\theta}, x) - \widehat{P}_0 g(\boldsymbol{\theta}, x) \right\} \\ &= \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{k=1}^K \frac{1}{n_k} \sum_{i=1}^{n_k} f_k(\boldsymbol{\theta}, X_{k,i}) - \frac{1}{n_0} \sum_{i=1}^{n_0} g(\boldsymbol{\theta}, X_{0,i}) \right\} =: \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \widehat{\ell}_\psi(\boldsymbol{\theta}) \end{aligned} \quad (6.19)$$

and the following “dual maximum” ψ -divergence estimates of the parameter $\boldsymbol{\theta}_T$

$$\widehat{\boldsymbol{\theta}}_\psi := \operatorname{argsup}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{k=1}^K \widehat{Q}_k f_k(\boldsymbol{\theta}, x) - \widehat{P}_0 g(\boldsymbol{\theta}, x) \right\} =: \operatorname{argsup}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \widehat{\ell}_\psi(\boldsymbol{\theta}). \quad (6.20)$$

6.3 EL and ψ -divergences under the exponential MDRM

We give a brief introduction of empirical likelihood (EL) approach of Owen (1991) and Qin and Lawless (1994) in the MDRM context; for more information we can refer to Qin and Zhang (1997), Qin (1998), Fokianos *et al.* (2001), Keziou and Leoni-Aubin (2008), Chen and Liu (2013) and Cai *et al.* (2013). For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the EL associated to the independent samples $(X_{k,i}, i = 1, \dots, n_k), k = 0, 1, \dots, K$, under the exponential MDRM (6.4), is $L(\boldsymbol{\theta}, p) = \prod_{k=0}^K \prod_{i=1}^{n_k} p(X_{k,i}) r_k(\boldsymbol{\theta}_k, X_{k,i})$; For $k = 0$, we set $\boldsymbol{\theta}_0 := \mathbf{0}$, so that $r_0(\boldsymbol{\theta}_0, x) = 1, \forall x \in \mathbb{R}^m$. In the EL approach, we consider the probability distribution P as if it is discrete, and $p(X_{k,i})$ is the mass induced by P on the observation $X_{k,i}$, for all $k = 1, \dots, K$, and all $i = 1, \dots, n_k$. Hence, p can be seen as a vector in $\mathbb{R}_+^{*n}$, where $n := n_0 + n_1 + \dots + n_K$ is the total sample size, and such that the sum of its components is equal to one. The log-EL is

$$\log L(\boldsymbol{\theta}, p) = \log \prod_{k=0}^K \prod_{i=1}^{n_k} p(X_{k,i}) r_k(\boldsymbol{\theta}_k, X_{k,i}) = \sum_{k=0}^K \sum_{i=1}^{n_k} (\log p(X_{k,i}) + \log r_k(\boldsymbol{\theta}_k, X_{k,i})).$$

The model assumption (6.4) implies that

$$\int_{\mathbb{R}^m} r_l(\boldsymbol{\theta}_l, x) dP(x) = 1, \quad \forall l = 0, 1, \dots, K. \quad (6.21)$$

Thus,

$$\sum_{k=0}^K \sum_{i=1}^{n_k} p(X_{k,i}) r_l(\boldsymbol{\theta}_l, X_{k,i}) = 1, \quad \forall l = 0, 1, \dots, K, \quad (6.22)$$

for the empirical version. The profil log-EL (in $\boldsymbol{\theta}$) is then defined by

$$\ell(\boldsymbol{\theta}) := \sup_{p \in \mathcal{C}_{\boldsymbol{\theta}}} \log L(\boldsymbol{\theta}, p) = \sup_{p \in \mathcal{C}_{\boldsymbol{\theta}}} \sum_{k=0}^K \sum_{i=1}^{n_k} (\log p(X_{k,i}) + \log r_k(\boldsymbol{\theta}_k, X_{k,i})), \quad (6.23)$$

where p , due to (6.22), is constrained to the set

$$\mathcal{C}_{\boldsymbol{\theta}} := \left\{ p \in \mathbb{R}_+^{*n} : \sum_{k=0}^K \sum_{i=1}^{n_k} p(X_{k,i}) = 1, \sum_{k=0}^K \sum_{i=1}^{n_k} p(X_{k,i}) r_l(\boldsymbol{\theta}_l, X_{k,i}) = 1, \forall l = 1, \dots, K \right\}. \quad (6.24)$$

The EL estimate $\hat{\boldsymbol{\theta}}_{EL}$ of the true value of the parameter $\boldsymbol{\theta}_T$ is then

$$\hat{\boldsymbol{\theta}}_{EL} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argsup}} \ell(\boldsymbol{\theta}). \quad (6.25)$$

For a given $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, if the condition

$$\exists p \in \mathcal{C}_{\boldsymbol{\theta}} \quad \text{such that} \quad |\log L(\boldsymbol{\theta}, p)| < \infty \quad (6.26)$$

is satisfied, then by Lagrange multipliers theorem, we can see that the supremum over p , in (6.23), is attained when

$$p(X_{k,i}) = \frac{1}{n \left\{ 1 + \sum_{l=1}^K \hat{\rho}_l [r_l(\boldsymbol{\theta}_l, X_{k,i}) - 1] \right\}}, \quad \forall k = 0, 1, \dots, K, \forall i = 1, \dots, n_k,$$

where $(\hat{\rho}_1, \dots, \hat{\rho}_K)$ is the solution, which depends on $\boldsymbol{\theta}$ and on the data, of the system of equations

$$\sum_{k=0}^K \sum_{i=1}^{n_k} p(X_{k,i}) r_l(\boldsymbol{\theta}_l, X_{k,i}) = 1, \quad \forall l = 1, \dots, K. \quad (6.27)$$

Given $\widehat{\boldsymbol{\theta}}_{EL}$. Interestingly, using the explicit form the exponential MDRM (6.4), by differentiating $\ell(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ at the point $\widehat{\boldsymbol{\theta}}_{EL}$, and solving the system of equations

$$(\partial/\partial\alpha_k)\ell(\widehat{\boldsymbol{\theta}}_{EL}) = 0, \quad (\partial/\partial\beta_k)\ell(\widehat{\boldsymbol{\theta}}_{EL}) = \mathbf{0}, \quad \forall k = 1, \dots, K,$$

one can show, as in the case of two-sample problem, see [Keziou and Leoni-Aubin \(2008\)](#), that the corresponding Lagrange multipliers $\widehat{\rho}_1, \dots, \widehat{\rho}_K$ depend only on the sample sizes, and we always have $\widehat{\rho}_l = n_l/n$, $\forall l = 1, \dots, K$. We obtain then, $\widehat{\boldsymbol{\theta}}_{EL} := \operatorname{argsup}_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}) = \operatorname{argsup}_{\boldsymbol{\theta} \in \Theta} \sup_{p \in \mathcal{C}_\theta} \log L(\boldsymbol{\theta}, p) = \operatorname{argsup}_{\boldsymbol{\theta} \in \Theta} \ell_d(\boldsymbol{\theta})$, where

$$\ell_d(\boldsymbol{\theta}) := -n \log n - \sum_{k=0}^K \sum_{i=1}^{n_k} \log \left(\widehat{\rho}_0 + \sum_{l=1}^K \widehat{\rho}_l r_l(\boldsymbol{\theta}_l, X_{k,i}) \right) + \sum_{k=1}^K \sum_{i=1}^{n_k} \log r_k(\boldsymbol{\theta}_k, X_{k,i}), \quad (6.28)$$

with $\widehat{\rho}_l = n_l/n$, $\forall l = 0, 1, \dots, K$. The function $\ell_d(\cdot)$ has the same maximum value and point as the profile log-EL $\ell(\cdot)$. Observe also that the function $\ell_d(\cdot)$ are well defined for any $\boldsymbol{\theta}$, unlike the profile log-EL $\ell(\cdot)$. In fact, in the particular case of two-sample problem, under the alternative hypothesis $\mathcal{H}_1$, that is when the distributions are different, i.e., when $\boldsymbol{\beta}_T \neq \mathbf{0}$, [Zou et al. \(2002\)](#) showed that if P_0 is not degenerate, there exists a neighborhood $N(\boldsymbol{\theta}_T)$ on which condition (6.26) is satisfied for n_0 sufficiently large, and then the profile log-EL $\ell(\cdot) : \boldsymbol{\theta} \in N(\boldsymbol{\theta}_T) \mapsto \ell(\boldsymbol{\theta})$ is well defined and finite; furthermore, they proved that there is no such neighborhood under the null hypothesis that the laws are equal, and proposed to use a partial profile empirical likelihood instead of $\ell(\cdot)$. [Keziou and Leoni-Aubin \(2008\)](#) showed that the dual version $\ell_d(\cdot)$ is well defined regardless the null hypothesis, and proposed to use it instead of the profile log-EL $\ell(\cdot)$, since $\ell_d(\cdot)$ is well defined regardless the null hypothesis that the laws are equal, and coincides with the log-EL $\ell(\cdot)$ -profile under the assumption that the laws are not equal. Moreover, they proposed a two-sample homogeneity test based on $\ell_d(\cdot)$ -profile. Consider now, in the context of the exponential MDRM, the test problem of the null hypothesis

$$\mathcal{H}_0 : Q_1 = \dots = Q_K = P_0 \quad \text{against the alternative} \quad \mathcal{H}_1 : \exists Q_k \neq P_0.$$

or equivalently

$$\mathcal{H}_0 : \boldsymbol{\beta}_{1T} = \dots = \boldsymbol{\beta}_{KT} = \mathbf{0} \quad \text{against the alternative} \quad \mathcal{H}_1 : \exists \boldsymbol{\beta}_{kT} \neq \mathbf{0},$$

since $\boldsymbol{\beta}_{kT} = \mathbf{0}$ implies $\alpha_k = 0$, for all $k = 1, \dots, K$. The corresponding EL ratio statistic can be written as follows

$$\begin{aligned} S_n &:= 2 \log \frac{\sup_{\boldsymbol{\theta} \in \Theta} \sup_{p \in \mathcal{C}_\theta} \ell(\boldsymbol{\theta}, p)}{\sup_{p \in \mathcal{C}_0} \prod_{k=0}^K \prod_{i=1}^{n_k} p(X_{k,i})} = 2 \left(\sup_{\boldsymbol{\theta} \in \Theta} \ell_d(\boldsymbol{\theta}) + n \log n \right) \\ &= 2 \sup_{\boldsymbol{\theta} \in \Theta} \left\{ - \sum_{k=0}^K \sum_{i=1}^{n_k} \log(\widehat{\rho}_0 + \sum_{l=1}^K \widehat{\rho}_l r_l(\boldsymbol{\theta}_l, X_{k,i})) + \sum_{k=1}^K \sum_{i=1}^{n_k} \log r_k(\boldsymbol{\theta}_k, X_{k,i}) \right\} \\ &= 2 \sup_{\boldsymbol{\theta} \in \Theta} \left\{ \sum_{k=1}^K \sum_{i=1}^{n_k} \log r_k(\boldsymbol{\theta}_k, X_{k,i}) - \sum_{k=1}^K \sum_{i=1}^{n_k} \log(\widehat{\rho}_0 + \sum_{l=1}^K \widehat{\rho}_l r_l(\boldsymbol{\theta}_l, X_{k,i})) \right. \\ &\quad \left. - \sum_{i=1}^{n_0} \log(\widehat{\rho}_0 + \sum_{l=1}^K \widehat{\rho}_l r_l(\boldsymbol{\theta}_l, X_{0,i})) \right\} \\ &= 2n \sup_{\boldsymbol{\theta} \in \Theta} \left\{ \sum_{k=1}^K \int_{\mathbb{R}^m} \widehat{\rho}_k \log r_k(\boldsymbol{\theta}_k, x) - \widehat{\rho}_k \log(\widehat{\rho}_0 + \sum_{l=1}^K \widehat{\rho}_l r_l(\boldsymbol{\theta}_l, x)) d\widehat{Q}_k(x) \right. \\ &\quad \left. - \int_{\mathbb{R}^m} \widehat{\rho}_0 \log(\widehat{\rho}_0 + \sum_{l=1}^K \widehat{\rho}_l r_l(\boldsymbol{\theta}_l, x)) d\widehat{P}_0(x) \right\} = 2n \widehat{D}_{\psi_{EL}}(\mathbf{Q}; P_0), \end{aligned} \quad (6.29)$$

where $\psi_{EL}(\cdot)$ is the particular “weighted” nonnegative proper closed convex function on $\mathbb{R}^K$ defined by

$$\psi_{EL}(x_1, \dots, x_K) := \psi_{\hat{\rho}}(x_1, \dots, x_K) := \left(\sum_{l=1}^K \hat{\rho}_l \varphi_1(x_l) \right) - \varphi_1 \left(\hat{\rho}_0 + \sum_{l=1}^K \hat{\rho}_l x_l \right), \quad (6.30)$$

where φ_1 is the nonnegative proper closed convex function on $\mathbb{R}$ defined by

$$\varphi_1(x) := (x \log x - x + 1) \mathbb{1}_{\mathbb{R}_+}(x) + (+\infty) \mathbb{1}_{]-\infty, 0[}(x). \quad (6.31)$$

Moreover, we have $\psi_{EL}(\mathbf{1}_K) = 0$, and it is straightforward to see that $\psi_{EL}(\cdot)$ is a member of the ψ -functions defined in (2.1). Whence, the dual empirical likelihood ratio statistic S_n is equal to $2n$ times the dual plug-in estimate of the particular ψ_ρ -divergence defined by the convex function (6.30). Moreover, the EL estimate $\hat{\theta}_{EL}$ is equal to the corresponding dual maximum ψ_ρ -divergence estimate, namely,

$$\begin{aligned} \hat{\theta}_{EL} = \operatorname{argsup}_{\theta \in \Theta} \left\{ \sum_{k=1}^K \int_{\mathbb{R}^m} \hat{\rho}_k \log r_k(\theta_k, x) - \hat{\rho}_k \log(\hat{\rho}_0 + \sum_{l=1}^K \hat{\rho}_l r_l(\theta_l, x)) d\hat{Q}_k(x) \right. \\ \left. - \int_{\mathbb{R}^m} \hat{\rho}_0 \log(\hat{\rho}_0 + \sum_{l=1}^K \hat{\rho}_l r_l(\theta_l, x)) d\hat{P}_0(x) \right\}. \end{aligned}$$

6.4 Asymptotic behavior of the estimates and test statistics

In this section, we give, in the context of exponential MDRM (6.4), the asymptotic properties of the dual maximum ψ_ρ -divergence estimates (6.20) of the parameter θ_T , and the dual plug-in estimates (6.19) of the ψ_ρ -divergences $D_{\psi_\rho}(Q_1, \dots, Q_K; P_0)$, when ψ_ρ is of the particular form

$$\psi_\rho : \mathbf{x} \in \mathbb{R}^K \mapsto \psi_\rho(\mathbf{x}) = \sum_{l=1}^K \rho_l \varphi(x_l) - \varphi \left(\rho_0 + \sum_{l=1}^K \rho_l x_l \right) \in [0, \infty], \quad (6.32)$$

where

$$\varphi : x \in \mathbb{R} \mapsto \varphi(x) \in [0, +\infty] \quad (6.33)$$

is any nonnegative proper closed strictly convex function, satisfying $\varphi(1) = 0$, and such that 1 is an interior point of its domain $\operatorname{dom}_\varphi := \{x \in \mathbb{R} \text{ s.t. } |\varphi(x)| < \infty\}$. Note that all functions in (6.32) are members of the ψ -functions defined in (2.1). We assume also that $\varphi(\cdot)$ is three times continuously differentiable on $\operatorname{int}(\operatorname{dom}_\varphi)$, $\varphi'(1) = 0$, and, without loss of generality, that $\varphi''(1) = 1$. Recall that the particular ψ_{EL} -function is obtained for the particular φ -function $\varphi(\cdot) = \varphi_1(\cdot)$ given in (6.31). Whence, the dual EL estimate and test statistic will be treated in this section as particular cases. In all the sequel, for any real valued function $f(\theta, x)$, we will denote by $f'(\theta, x)$ and $f''(\theta, x)$, respectively, the gradient and Hessian of $f(\theta, x)$, w.r.t the vector parameter θ , at the point θ , for all $x \in \mathbb{R}^m$. $f'''(\theta, x)$ denotes the matrix of the third order partial derivatives of $f(\theta, x)$ w.r.t. θ . Recall that the vector θ is of dimension $(K + d)$ with $d := \sum_{k=1}^K d_k$. Whence, f' is a column vector with dimension $(K + d)$, f'' and f''' are matrices, with dimension $(K + d)^2$ and $(K + d)^3$, respectively. We will denote by $|\cdot|$ for the Euclidean norm. Assume that the $(1+K)$ sample sizes, n_k , $k = 0, 1, \dots, K$, satisfy the condition

$$\hat{\rho}_k := \frac{n_k}{n} \rightarrow \rho_k \quad \text{for some real } \rho_k \in]0, 1[,$$

when the total sample size $n := \sum_{k=0}^K n_k \rightarrow +\infty$. We will use the following assumptions.

(A.1) θ_T is an interior point of Θ , and there exists a neighborhood $N(\theta_T)$ of θ_T such that

- (a) for all $k = 1, \dots, K$, the third order partial derivative functions $\{x \in \mathbb{R}^m \mapsto (f_k'''(\theta, x))_{i,j,k}; \theta \in N(\theta_T)\}$ are dominated by some Q_k -integrable function;

(b) the third order partial derivative functions $\{x \in \mathbb{R}^m \mapsto (g'''(\boldsymbol{\theta}, x))_{i,j,k}; \boldsymbol{\theta} \in N(\boldsymbol{\theta}_T)\}$ are dominated by some P_0 -integrable function;

(A.2) The integrals $\int_{\mathbb{R}^m} |f'_k(\boldsymbol{\theta}_T, x)|^2 dQ_k(x)$, $\int_{\mathbb{R}^m} |f''_k(\boldsymbol{\theta}_T, x)| dQ_k(x)$, $k = 1, \dots, K$, $\int_{\mathbb{R}^m} |g'(\boldsymbol{\theta}_T, x)|^2 dP_0(x)$, $\int_{\mathbb{R}^m} |g''(\boldsymbol{\theta}_T, x)| dP_0(x)$ are finite, and the matrix $U_\psi(\boldsymbol{\theta}_T) := -\sum_{k=1}^K Q_k f''_k(\boldsymbol{\theta}_T, x) + P_0 g''(\boldsymbol{\theta}_T, x)$ is nonsingular.

(A.3) The integrals $Q_k(f_k(\boldsymbol{\theta}_T, x)^2)$, $k = 1, \dots, K$, and $P_0(g(\boldsymbol{\theta}_T, x)^2)$ are finite.

Theorem 6.1. Assume that assumptions (A.1) and (A.2) hold.

- (1) Let $B(\boldsymbol{\theta}_T, n^{-1/3}) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_T| \leq n^{-1/3}\}$. Then, as $n \rightarrow \infty$, with probability one, $\widehat{\ell}_{\psi_\rho}(\boldsymbol{\theta})$ attains its maximum value at some point $\widehat{\boldsymbol{\theta}}_{\psi_\rho}$, in the interior of the ball $B(\boldsymbol{\theta}_T, n^{-1/3})$, satisfying $\widehat{\ell}'_{\psi_\rho}(\widehat{\boldsymbol{\theta}}_{\psi_\rho}) = \mathbf{0}$.
- (2) $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\psi_\rho} - \boldsymbol{\theta}_T)$ converges in distribution to a centered normal random vector with covariance matrix

$$M_{\psi_\rho}(\boldsymbol{\theta}_T) = U_{\psi_\rho}(\boldsymbol{\theta}_T)^{-1} V_{\psi_\rho}(\boldsymbol{\theta}_T) U_{\psi_\rho}(\boldsymbol{\theta}_T)^{-1}, \quad (6.34)$$

where $V_{\psi_\rho}(\boldsymbol{\theta}_T) = \sum_{k=1}^K \rho_k^{-1} V_{Q_k}(f'_k(\boldsymbol{\theta}_T, x)) + \rho_0^{-1} V_{P_0}(g'(\boldsymbol{\theta}_T, x))$.

- (3) Under the null hypothesis $\mathcal{H}_0$, the statistic $S_n^{\psi_\rho} := 2n \widehat{D}_{\psi_\rho}(\mathbf{Q}; P_0) := 2n \widehat{\ell}_{\psi_\rho}(\widehat{\boldsymbol{\theta}}_{\psi_\rho})$ converges in distribution to a χ^2 -random variable with $d = \sum_{k=1}^K d_k$ degrees of freedom;
- (4) Assume that the additional condition (A.3) is satisfied. Then, under the alternative hypothesis $\mathcal{H}_1$, $\sqrt{n} \left(\frac{S_n^{\psi_\rho}}{2n} - D_{\psi_\rho}(\mathbf{Q}; P_0) \right)$ converges in distribution to a centered normal random variable with variance $\sigma_{\psi_\rho}(\boldsymbol{\theta}_T)^2 = \sum_{k=1}^K \rho_k^{-1} V_{Q_k}(f_k(\boldsymbol{\theta}_T, x)) + \rho_0^{-1} V_{P_0}(g(\boldsymbol{\theta}_T, x))$.

Remark 6.1. (a) For the particular dual empirical likelihood case, i.e., when $\psi = \psi_{EL}$, by some algebra computations, one can show that the matrix $V_{\psi_{EL}}(\boldsymbol{\theta}_T)$ can be written as

$$V_{\psi_{EL}}(\boldsymbol{\theta}_T) = U_{\psi_{EL}}(\boldsymbol{\theta}_T) - U_{\psi_{EL}}(\boldsymbol{\theta}_T) D U_{\psi_{EL}}(\boldsymbol{\theta}_T), \quad (6.35)$$

where D is the $K + d$ by $K + d$ matrix given by

$$D = \begin{bmatrix} D_{1,1} & D_{1,2} \\ D_{2,1} & D_{2,2} \end{bmatrix} \quad (6.36)$$

with $D_{1,1} = \rho_0^{-1} \mathbf{1}_K \mathbf{1}_K^\top + \text{diag}\{\rho_1^{-1}, \dots, \rho_K^{-1}\}$, $D_{1,2} = D_{2,1}^\top = \mathbf{0}_K \mathbf{0}_d^\top$ and $D_{2,2} = \mathbf{0}_d \mathbf{0}_d^\top$. Hence, the limiting variance of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\psi_{EL}} - \boldsymbol{\theta}_T)$ can be written as $M_{\psi_{EL}}(\boldsymbol{\theta}_T) = U_{\psi_{EL}}(\boldsymbol{\theta}_T)^{-1} - D$.

- (b) Generally, the limiting variance $M_{\psi_\rho}(\boldsymbol{\theta}_T)$ of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\psi_\rho} - \boldsymbol{\theta}_T)$ depends on ψ_ρ , and we can extend the results of [Qin \(1998\)](#) Theorem 3 to the present context of multisample problem, to prove that, when the ratios are of the exponential form (6.4), then the dual empirical likelihood estimate $\widehat{\boldsymbol{\theta}}_{EL}$ is optimal in the sense that $M_{\psi_{EL}}(\boldsymbol{\theta}_T) - M_{\psi_\rho}(\boldsymbol{\theta}_T)$ is positive semidefinite for any ψ_ρ , under assumption (A.1) and (A.2). However, for more general MDRM (6.2), it is not clear which estimate $\widehat{\boldsymbol{\theta}}_{\psi_\rho}$ is optimal.
- (c) Under the null hypothesis that all distributions are equal, the limiting matrix variance M_{ψ_ρ} does not depend on ψ_ρ , and some algebra computation shows that $M_{\psi_\rho} = M = U^{-1} - D$ with

$$U = \begin{bmatrix} A & E_{P_0}(B(x)) \\ E_{P_0}(B(x))^\top & E_{P_0}(C(x)) \end{bmatrix},$$

$$A = \psi''_{\rho}(\mathbf{1}_K) \quad (6.37)$$

$$= \begin{bmatrix} \rho_1 - \rho_1^2 & -\rho_1\rho_2 & \cdots & -\rho_1\rho_K \\ -\rho_2\rho_1 & \rho_2 - \rho_2^2 & \cdots & -\rho_2\rho_K \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_K\rho_1 & \cdots & \cdots & \rho_K - \rho_K^2 \end{bmatrix} \quad (6.38)$$

$$B(x) = \left(A \begin{bmatrix} \mathbf{1}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{0}_{d_K}^\top \\ \mathbf{0}_{d_1}^\top & \mathbf{1}_{d_2}^\top & \cdots & \mathbf{0}_{d_K}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{1}_{d_K}^\top \end{bmatrix} \right) \circ \begin{bmatrix} \mathbf{b}(x)^\top \\ \vdots \\ \mathbf{b}(x)^\top \end{bmatrix} \quad (6.39)$$

$$= \left(A \begin{bmatrix} \mathbf{1}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{0}_{d_K}^\top \\ \mathbf{0}_{d_1}^\top & \mathbf{1}_{d_2}^\top & \cdots & \mathbf{0}_{d_K}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{1}_{d_K}^\top \end{bmatrix} \right) \circ [\mathbf{b}(x)^\top \otimes \mathbf{1}_K] \quad (6.40)$$

and

$$C(x) = \left(\begin{bmatrix} \mathbf{1}_{d_1} & \mathbf{0}_{d_1} & \cdots & \mathbf{0}_{d_1} \\ \mathbf{0}_{d_2} & \mathbf{1}_{d_2} & \cdots & \mathbf{0}_{d_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{d_K} & \mathbf{0}_{d_K} & \cdots & \mathbf{1}_{d_K} \end{bmatrix} A \begin{bmatrix} \mathbf{1}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{0}_{d_K}^\top \\ \mathbf{0}_{d_1}^\top & \mathbf{1}_{d_2}^\top & \cdots & \mathbf{0}_{d_K}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{1}_{d_K}^\top \end{bmatrix} \right) \circ [\mathbf{b}(x) \mathbf{b}(x)^\top].$$

Remark 6.2. Consider the problem of testing the null hypothesis $\mathcal{H}_0 : Q_1 = \cdots = Q_K = P_0$ against the alternative $\mathcal{H}_1 : \exists Q_k \neq P_0$. In view of the above theorem 6.1 part (3), we propose the following decision rule: we reject $\mathcal{H}_0$ iff the statistic $S_n^{\psi_\rho} > \chi_d^2(1-\alpha)$, where $\chi_d^2(1-\alpha)$ is the $(1-\alpha)$ -quantile of the χ^2 -distribution with d degrees of freedom. This defines a test asymptotically of level $\alpha \in]0, 1[$. The critical region is then of the form $CR_{\psi_\rho} := \{S_n^{\psi_\rho} > \chi_d^2(1-\alpha)\}$. Using part (4) of the above Theorem, we obtain for the power function $\boldsymbol{\theta}_T \in \boldsymbol{\Theta}/\{\mathbf{0}\} \mapsto \beta_{\psi_\rho}(\boldsymbol{\theta}_T) := P_{\mathcal{H}_1}(S_n^{\psi_\rho} > \chi_d^2(1-\alpha))$, the following approximation $\tilde{\beta}_{\psi_\rho}(\boldsymbol{\theta}_T) := 1 - F_{\mathcal{N}}\left(\frac{\sqrt{n}}{\hat{\sigma}_{\psi_\rho}(\boldsymbol{\theta}_T)} \left[\frac{\chi_d^2(1-\alpha)}{2n} - \hat{\ell}_{\psi_\rho}(\boldsymbol{\theta}_T)\right]\right)$, where $\hat{\ell}_{\psi_\rho}(\boldsymbol{\theta}_T) = \sum_{k=1}^K \hat{Q}_k f_k(\boldsymbol{\theta}_T, x) - \hat{P}_0 g(\boldsymbol{\theta}_T, x)$ and $\hat{\sigma}_{\psi_\rho}(\boldsymbol{\theta}_T)^2 = \sum_{k=1}^K \hat{\rho}_k^{-1} V_{\hat{Q}_k}(f_k(\boldsymbol{\theta}_T, x)) + \hat{\rho}_0^{-1} V_{\hat{P}_0}(g(\boldsymbol{\theta}_T, x))$. So, if $\mathcal{H}_0$ is rejected, the estimate $\hat{\boldsymbol{\theta}}_{\psi_\rho}$ is significantly different from $\mathbf{0}$, and the power of the proposed test can be approximated by $\tilde{\beta}_{\psi_\rho}(\hat{\boldsymbol{\theta}}_{\psi_\rho})$.

6.5 Discussion and directions for future research

We have introduced new estimates and test for multi-sample semiparametric density ratio models, through our dual representations of multivariate divergences introduced in Keziou (2015) and presented in Chapter 2. Our approach recovers the dual empirical likelihood one and allows to obtain the asymptotic properties of the proposed estimates and test statistics, including the empirical likelihood ones, both under the model and under misspecification. In testing context, the obtained asymptotic distributions under alternative hypotheses lead to accurate approximation of the power functions. We aim at extending the present approach to the case where the ratios belong to infinite-dimensional function spaces. Moreover, the proposed dual estimates of the introduced multivariate divergences can be used also as criteria for model selection, in multi-sample density ratio models, and for multi-group classification, that we will investigate in future works.

Chapter 7

Semiparametric estimation of mutual information and related criteria : Optimal test of independence

This Chapter presents our papers [Keziou and Regnault \(2015\)](#) and [Keziou and Regnault \(2016\)](#). We derive independence tests by means of dependence measures thresholding in a semiparametric context. Precisely, estimates of φ -mutual informations, associated to φ -divergences between a joint distribution and the product distribution of its marginals, are derived through the dual representation of φ -divergences. The asymptotic properties of the proposed estimates are established, including consistency, asymptotic distributions and large deviations principle. The obtained tests of independence are compared via their relative asymptotic Bahadur efficiency and numerical simulations. It follows that the proposed semiparametric Kullback-Leibler Mutual information test is the optimal one. On the other hand, the proposed approach provides a new method for estimating the Kullback-Leibler mutual information in a semiparametric setting, as well as a model selection procedure in large class of dependency models including semiparametric copulas.

7.1 Introduction and notations

Measuring the dependence between random variables has been a central aim of probability theory since its earliest developments. Classical examples of dependence measures are correlation measures of Pearson, Kendall or Spearman. While the first one focuses on linear relationship between real random variables, the two second ones measure the monotonic relationship between variables taking values in ordered sets. Pure-independence measures, between variables X and Y taking values in general measurable spaces $(\mathcal{X}, \mathcal{A}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{A}_\mathcal{Y})$, can be defined by considering any divergence between the joint distribution $\mathbb{P}$ of (X, Y) and the product distribution of its marginals $\mathbb{P}^\otimes := \mathbb{P}_1 \otimes \mathbb{P}_2$, where $\mathbb{P}_1$ and $\mathbb{P}_2$ are, respectively, the marginal distributions of X and Y . The most outstanding and widely used example of such dependence measures is the χ^2 -divergence between $\mathbb{P}$ and $\mathbb{P}^\otimes$ defined by

$$\chi^2(\mathbb{P}, \mathbb{P}^\otimes) := \frac{1}{2} \int_{\mathcal{X} \times \mathcal{Y}} \left(\frac{d\mathbb{P}}{d\mathbb{P}^\otimes}(x, y) - 1 \right)^2 d\mathbb{P}^\otimes(x, y), \quad (7.1)$$

where $\frac{d\mathbb{P}}{d\mathbb{P}^\otimes}$ denotes the density of $\mathbb{P}$ with respect to (w.r.t.) $\mathbb{P}^\otimes$. Note that, if $\mathbb{P}$ is a discrete distribution, i.e., if its support $\mathcal{X} \times \mathcal{Y} := \text{supp}(\mathbb{P})$ is discrete (finite or countably infinite) set, then the above divergence writes

$$\chi^2(\mathbb{P}, \mathbb{P}^\otimes) = \frac{1}{2} \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{(p_{x,y} - p_x p_y)^2}{p_x p_y},$$

where $\mathbb{P} := (p_{x,y})_{(x,y)}$, $\mathbb{P}^\otimes = (p_x p_y)_{(x,y)}$, with $p_x := \sum_y p_{x,y}$ and $p_y := \sum_x p_{x,y}$. Another classical example, associated to the Kullback-Leibler (KL) divergence between $\mathbb{P}$ and $\mathbb{P}^\otimes$, is the well-known mutual information (MI) defined by (see e.g. [Cover and Thomas \(2006\)](#))

$$I_{KL}(\mathbb{P}) := \mathbb{K}(\mathbb{P}, \mathbb{P}^\otimes) := \int_{\mathcal{X} \times \mathcal{Y}} \frac{d\mathbb{P}}{d\mathbb{P}^\otimes}(x, y) \log \frac{d\mathbb{P}}{d\mathbb{P}^\otimes}(x, y) d\mathbb{P}^\otimes(x, y), \quad (7.2)$$

which, in the case of discrete distributions, can be written under the form

$$I_{KL}(\mathbb{P}) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p_{x,y} \log \frac{p_{x,y}}{p_x p_y}.$$

We will call the above classical measures of dependence (7.1) and (7.2), respectively, χ^2 -mutual information (χ^2 -MI) and KL-mutual information (KL-MI). When dealing with i.i.d. observations $(X_1, Y_1), \dots, (X_n, Y_n)$, of two random variables (X, Y) , we may test the null hypothesis, that the variables X and Y are independent, by means of estimating such dependence measure and deciding to reject the null hypothesis of independence if the estimate is sufficiently far from zero; the classical χ^2 -independence test is such a procedure : the corresponding test statistic (in the discrete-distribution case) is

$$2n \chi^2(\hat{\mathbb{P}}, \hat{\mathbb{P}}^\otimes) = n \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{(\hat{p}_{x,y} - \hat{p}_x \hat{p}_y)^2}{\hat{p}_x \hat{p}_y}, \quad (7.3)$$

where $\hat{\mathbb{P}} := (\hat{p}_{x,y})_{(x,y)}$ and $\hat{\mathbb{P}}^\otimes := (\hat{p}_x \hat{p}_y)_{(x,y)}$ are, respectively, the empirical versions of $\mathbb{P} = (p_{x,y})_{(x,y)}$ and $\mathbb{P}^\otimes = (p_x p_y)_{(x,y)}$. Likewise, to test the independence, we can consider as dependence measure the KL-MI and use the test statistic

$$2n I_{KL}(\hat{\mathbb{P}}) = 2n \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \hat{p}_{x,y} \log \frac{\hat{p}_{x,y}}{\hat{p}_x \hat{p}_y}. \quad (7.4)$$

The dependence measure can also be any other φ -divergence between $\mathbb{P}$ and $\mathbb{P}^\otimes$. The asymptotic properties of the tests based on such dependence measures, including the χ^2 -MI and KL-MI ones, have been extensively studied in the case of finite-discrete distributions; see e.g. [Pardo \(2006\)](#) Chapter 8, and the references therein. [Kontoyiannis and Skoularidou \(2015\)](#) consider the estimating problem of mutual information by plug-in for Markov chains, and perform a hypothesis test for the presence of causal influence between the two processes. The challenging nonasymptotic setting, particularly, for entropy and mutual information estimation, still in the case of discrete distributions, was recently investigated by [Valiant and Valiant \(2011b,a\)](#), [Jiao et al. \(2015\)](#) and [Wu and Yang \(2016\)](#). See also [Chan et al. \(2014\)](#) for a nonasymptotic study of testing closeness of discrete distributions, with minimax setting. When dealing with continuous distributions (or continuous random variables), obviously, the above direct plug-in estimates (7.3) and (7.4), of the dependence measures (7.1) and (7.2), are not well defined. Moreover, for countably-infinite discrete distributions, although the above estimates (7.3) and (7.4) remain well defined, they suffer from loss of efficiency, even in the case of finite-discrete distributions, if the support size is large. Therefore, in the case of non finite-discrete distributions, particularly, for the widely used KL-MI, other kind of estimates have been proposed and studied in the literature; see e.g. [Moon et al. \(1995\)](#) for a kernel density estimate, [Kraskov et al. \(2004\)](#) for a k -nearest-neighbor estimate extending those of Shannon entropy in one dimension based on m -spacing; see e.g. [Tsybakov and van der Meulen \(1996\)](#), [Dudewicz and van der Meulen \(1981\)](#) and [Beirlant et al. \(1997\)](#) among others. [Van Hulle \(2005\)](#) derive an estimate using Edgeworth approximation of Shannon entropy. [Darbellay and Vajda \(1999\)](#), [Wang et al. \(2005\)](#) and [Cellucci et al. \(2005\)](#) propose estimates based on adaptative partitioning of $\mathcal{X} \times \mathcal{Y}$. See also [Khan et al. \(2007\)](#) for an overview and numerical comparisons of these estimates. Based on the Kullback-Leibler importance estimation procedure, see [Sugiyama et al. \(2008\)](#), [Suzuki et al. \(2008\)](#) obtain an estimate of KL-MI called maximum likelihood mutual information, see also [Sugiyama et al. \(2012a\)](#) Chapter 11. Unfortunately, their (asymptotic) distributions are generally not available. Hence, testing independence from these estimates requires Monte-Carlo or Bootstrap approximations of the related p -values. On the other hand, the above nonparametric estimates suffer from loss of efficiency, due

to smoothing or partitioning, and suffer also from the difficulty of conveniently choosing the classes, the number of classes or the smoothing parameters (the bandwidths and the kernels). The present Chapter introduces new efficient semiparametric estimates of φ -mutual information (φ -MI), i.e., dependence measures associated to φ -divergence functionals, including the well known KL-MI and χ^2 -MI. These estimates are obtained by making use of a dual representation of φ -MI, presented in Section 7.2, without using any smoothing nor partitioning. The obtained estimates are defined in the same way for both finite-discrete or non-discrete distributions, and coincide with the direct plug-in ones in the case of finite-discrete distributions, with a specific choice of the model ratio; see Example 7.6. Their asymptotic properties are presented in Section 7.3. Particularly, the consistency is stated for a large variety of semiparametric models for $d\mathbb{P}/d\mathbb{P}^\otimes$; the asymptotic distribution is obtained for the KL-MI estimate in a special setting. The present approach leads to new independence tests, whose Bahadur efficiency are compared in Section 7.4; the most efficient test is shown to be the one based on the proposed estimate of the particular KL-MI criterion. It can be used also in order to build a large variety of dependence models, through for instance a cross validation-type model selection procedure based on the proposed estimate of φ -MI measure of dependence; see Section 7.2.4. The powers of φ -MI based tests are compared numerically to classical noncorrelation tests in Keziou and Regnault (2015). The results in the present Chapter have the advantage (unlike the classical noncorrelation tests) to remain valid in the case of multisample problem (estimating φ -mutual informations of a multidimensional random variable as well as testing simultaneous independence of its components), but for simplicity, the results will be presented only for the two-sample case. The same results hold for the multisample problem.

7.2 φ -mutual informations, dual representations and estimation strategy

Given an i.i.d. sample, $(X_1, Y_1), \dots, (X_n, Y_n)$, of a random vector (X, Y) taking values in a measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}_\mathcal{X} \otimes \mathcal{A}_\mathcal{Y})$, we aim at testing the null hypothesis $\mathcal{H}_0$ of independence of the marginal X and Y ; formally

$$\mathcal{H}_0 : X \text{ and } Y \text{ are independent, against } \mathcal{H}_1 : X \text{ and } Y \text{ are dependent.} \quad (7.5)$$

We derive such tests by estimating and thresholding φ -mutual informations between X and Y in a semiparametric context. Sections 7.2.1, 7.2.2 and 7.2.3 to follow, respectively, define φ -mutual informations, present the semiparametric model under study, and introduce estimates of φ -MI used as test statistics for the test problem (7.5). Section 7.2.4 defines a cross-validation procedure for model selection among L candidate models for the ratio $d\mathbb{P}/d\mathbb{P}^\otimes$, using the proposed estimate of φ -MI.

7.2.1 Introducing φ -mutual informations

Denote $M^1(\mathcal{X} \times \mathcal{Y})$ the set of all distributions on the product measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}_\mathcal{X} \otimes \mathcal{A}_\mathcal{Y})$. Let $\varphi : \mathbb{R} \rightarrow [0, +\infty]$ be some nonnegative closed proper convex function such that its domain $\text{dom}_\varphi := \{x \in \mathbb{R}; \varphi(x) < \infty\} =: (a_\varphi, b_\varphi)$ is an interval, with endpoints $a_\varphi < 1 < b_\varphi$, and $\varphi(1) = 0$. The interval (a_φ, b_φ) may be bounded or unbounded, open or not. The φ -divergence between any probability distributions $Q, P \in M^1(\mathcal{X} \times \mathcal{Y})$, if Q is absolutely continuous with respect to (a.c.w.r.t.) P , is defined by $D_\varphi(Q, P) := \int_{\mathcal{X} \times \mathcal{Y}} \varphi\left(\frac{dQ}{dP}(x, y)\right) dP(x, y)$. If Q is not a.c.w.r.t. P , we set $D_\varphi(Q, P) = +\infty$. Note that $D_\varphi(Q, P) \geq 0$, for any Q and P . Moreover, if φ is strictly convex on some neighborhood of 1, we have the fundamental property $D_\varphi(Q, P) \geq 0$, with equality if and only if $Q = P$. In the following, we assume that the function φ is strictly convex and two times continuously differentiable on the interior of its domain (a_φ, b_φ) . We have then $\varphi'(1) = 0$, and without loss of generality, we can assume that $\varphi''(1) = 1$. The well-known Kullback-Leibler divergence $\mathbb{K}(\cdot, \cdot)$ is obtained for $\varphi(x) = \varphi_1(x) := x \log x - x + 1$, the “modified” Kullback-Leibler divergence $\mathbb{K}_m(\cdot, \cdot)$ is obtained for $\varphi(x) = \varphi_0(x) := -\log x + x - 1$. The χ^2 and modified- χ^2 divergences, denoted $\chi^2(\cdot, \cdot)$ and $\chi_m^2(\cdot, \cdot)$, are associated, respectively, to the convex functions $\varphi(x) = \varphi_2(x) := (x - 1)^2/2$ and $\varphi(x) = \varphi_{-1}(x) := (x - 1)^2/(2x)$. The so-called Hellinger

distance $H(\cdot, \cdot)$ is obtained for $\varphi(x) = \varphi_{1/2}(x) := 2(\sqrt{x} - 1)^2$; see Table 1.1. All these divergences are members of the so-called “power-divergences” $D_{\varphi_\gamma}(\cdot, \cdot)$ associated to the convex functions $\varphi_\gamma(\cdot)$ defined by

$$\varphi_\gamma(\cdot) : x \in \mathbb{R}_+^* \mapsto \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)} \quad (7.6)$$

if $\gamma \in \mathbb{R} \setminus \{0, 1\}$, $\varphi_0(x) := -\log x + x - 1$ and $\varphi_1(x) := x \log x - x + 1$. The standard divergences $\mathbb{K}(\cdot, \cdot)$, $\mathbb{K}_m(\cdot, \cdot)$, $\chi^2(\cdot, \cdot)$, $\chi_m^2(\cdot, \cdot)$ and $H(\cdot, \cdot)$ are then associated, respectively, to the real convex functions $\varphi_1(\cdot)$, $\varphi_0(\cdot)$, $\varphi_2(\cdot)$, $\varphi_{-1}(\cdot)$ and $\varphi_{1/2}(\cdot)$. Note that the divergences are generally not symmetric; particularly, we have for any $Q, P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, $\mathbb{K}_m(Q, P) = \mathbb{K}(P, Q)$ and $\chi_m^2(Q, P) = \chi^2(P, Q)$. For more details and proofs, we can refer to [Liese and Vajda \(1987\)](#) and [Broniatowski and Keziou \(2006\)](#). For any probability distribution $P \in M^1(\mathcal{X} \times \mathcal{Y})$, let $P^\otimes$ denotes the product distribution $P^\otimes := P_1 \otimes P_2$ of the marginals P_1 and P_2 of P . The φ -mutual information of P , associated to the divergence $D_\varphi(\cdot, \cdot)$, is defined as

$$I_\varphi(P) := D_\varphi(P, P^\otimes).$$

For any random vector (X, Y) taking its values in $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}_\mathcal{X} \otimes \mathcal{A}_\mathcal{Y})$, with joint distribution $\mathbb{P} \in M^1(\mathcal{X} \times \mathcal{Y})$, the φ -mutual information (φ -MI) of (X, Y) is defined to be

$$I_\varphi(X, Y) := I_\varphi(\mathbb{P}) = D_\varphi(\mathbb{P}, \mathbb{P}^\otimes) = \int_{\mathcal{X} \times \mathcal{Y}} \varphi \left(\frac{d\mathbb{P}}{d\mathbb{P}^\otimes}(x, y) \right) d\mathbb{P}^\otimes(x, y). \quad (7.7)$$

Since $D_\varphi(\mathbb{P}, \mathbb{P}^\otimes) \geq 0$, with equality if and only if $\mathbb{P} = \mathbb{P}^\otimes$, i.e., if and only if X and Y are independent, φ -MI measures then the dependence between the random variables X and Y . In contrast to the correlation coefficients of Pearson, Kendall or Spearman, the φ -MI does not focus on the linear or monotonic relationship between random variables; it constitutes a proper dependency measure. Note that I_{φ_1} and I_{φ_2} , with φ_1 and φ_2 given in Table 1.1, are, respectively, the KL-MI and χ^2 -MI, given by (7.2) and (7.1). Thus, the test problem (7.5) is equivalent, in the context of I_φ criteria, to testing $I_\varphi(\mathbb{P}) = 0$ against $I_\varphi(\mathbb{P}) > 0$. Hence, we can use as test statistic an estimate of $I_\varphi(\mathbb{P})$, and reject the null hypothesis $\mathcal{H}_0$ when the estimate takes large values. A natural attempt to estimate the φ -MI of (X, Y) consists in considering the plug-in estimate of $I_\varphi(\mathbb{P})$ obtained by replacing $\mathbb{P}(\cdot)$ by its empirical counterpart

$$\hat{\mathbb{P}}(\cdot) = \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, Y_i)}(\cdot), \quad (7.8)$$

associated to the i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of (X, Y) . Here, $\delta_{(x, y)}(\cdot)$ denotes the Dirac measure at (x, y) for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Unfortunately, by doing so, we only measure dependence of the contingency table associated to the sample. When dealing with variables X and Y absolutely continuous with respect to Lebesgue measure, the contingency table is almost surely an $n \times n$ table with all coefficients except diagonal ones equal to zero ; particularly, variables X and Y appear (misleadingly) purely dependent, yielding to reject systematically the null hypothesis. A second, less crude, approach consists in gathering the values X_i and Y_i into classes and testing independence between the induced finite-discrete variables $\tilde{X}$ and $\tilde{Y}$, by empirically estimating the φ -MI of $(\tilde{X}, \tilde{Y})$. This widespread approach suffers from the difficulty of conveniently choosing the classes. Moreover, an important amount of information carried by the sample is lost during this process, yielding to poor efficiency – or power – of these tests. An other approach, is to use kernel nonparametric estimates of the joint density and the marginal ones, but as it is well known this provides less efficient estimates and leads to the difficulty of choosing the optimal smoothing parameters. As an alternative, we propose in the present work semiparametric modeling of the ratio $d\mathbb{P}/d\mathbb{P}^\otimes$, and the use of duality to obtain well-defined estimates of φ -MI without smoothing nor partitioning. The present approach applies for both continuous or discrete distributions, or mixtures of continuous and discrete distributions.

7.2.2 Semiparametric modeling of the ratio $d\mathbb{P}/d\mathbb{P}^\otimes$

Assume that the joint distribution $\mathbb{P}$ of the random vector (X, Y) belongs to the semiparametric model

$$\mathcal{M}_\Theta := \left\{ P \in M^1(\mathcal{X} \times \mathcal{Y}) \text{ such that } \frac{dP}{dP^\otimes}(\cdot, \cdot) =: h_\theta(\cdot, \cdot); \theta \in \Theta \right\}, \quad (7.9)$$

where $\Theta \subset \mathbb{R}^{1+d}$ is the parameter space, and $h_\theta(\cdot, \cdot) : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto h_\theta(x, y) \in \mathbb{R}$ is some specified real-valued function, indexed by the parameter θ . In the sequel, we will consider the following assumptions on the model $\mathcal{M}_\Theta$.

(A.1) $(h_\theta(\cdot, \cdot) = h_{\theta'}(\cdot, \cdot), \mathbb{P}^\otimes\text{-a.s.}) \Rightarrow (\theta = \theta')$ (identifiability);

(A.2) there exists (a unique) $\theta_0 \in \text{int}(\Theta)$ satisfying $h_{\theta_0}(\cdot, \cdot) = 1$, $\mathbb{P}^\otimes\text{-a.s.}$

Assumption (A.1) is a natural identifiability condition for $dP/dP^\otimes$. Assumption (A.2) ensures independence is covered by the model $\mathcal{M}_\Theta$. The uniqueness of θ_0 follows from Assumption (A.1). Denote by θ_T the “true” unknown value of the parameter, namely, the unique value satisfying

$$\frac{d\mathbb{P}}{d\mathbb{P}^\otimes}(x, y) = h_{\theta_T}(x, y), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y},$$

which is assumed to be an interior point of Θ . Then, we have $\theta_T = \theta_0$ if and only if X and Y are independent. Below are listed some relevant examples of the model (7.9).

Example 7.1. Let $(X, Y) \in \mathbb{R}^2$ be a centered Gaussian random vector with correlation coefficient $\rho \in]-1, 1[$ and centered normal marginals with the same variance $\sigma^2 > 0$. A straightforward computation shows that the ratio $d\mathbb{P}/d\mathbb{P}^\otimes$ can be written under the form of the model (7.9) where

$$h_\theta(x, y) = \exp \{ \alpha + \beta_1(x^2 + y^2) + \beta_2xy \}, \quad (7.10)$$

$\theta := (\alpha, \beta_1, \beta_2)^\top \in \mathbb{R}^3$, with $\alpha = -\log(1 - \rho^2)/2$, $\beta_1 = -\rho^2/(2\sigma^2(1 - \rho^2))$ and $\beta_2 = \rho/(\sigma^2(1 - \rho^2))$. Note that the parameter value, corresponding to the independence hypothesis, is $\theta_0 = (0, 0, 0)^\top$. Moreover, if the distribution of (X, Y) is Gaussian with unknown mean $\mu := (\mu_1, \mu_2)^\top$ and unknown variance matrix Γ , then we can show that the ratio $d\mathbb{P}/d\mathbb{P}^\otimes$ can be written under the form of the model (7.9) with

$$h_\theta(x, y) = \exp \{ \alpha + \beta_1x + \beta_2y + \beta_3x^2 + \beta_4y^2 + \beta_5xy \}, \quad (7.11)$$

and $\theta := (\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\top$. Note that the number of free parameters in θ_T is $d = 5$, and that α_T is considered as a normalizing parameter due to the constraint $\int_{\mathcal{X} \times \mathcal{Y}} h_{\theta_T}(x, y) d\mathbb{P}^\otimes(x, y) = \int_{\mathcal{X} \times \mathcal{Y}} d\mathbb{P}(x, y) = 1$ since $\mathbb{P}$ is a probability distribution. Moreover, we have $\theta_0 = \mathbf{0} \in \mathbb{R}^6$.

Example 7.2. Let $\psi_0(\cdot, \cdot) := \mathbb{1}_{\mathcal{X} \times \mathcal{Y}}(\cdot, \cdot), \psi_1(\cdot, \cdot), \psi_2(\cdot, \cdot), \dots$, be some basis functions of the space $L^2(\mathcal{X} \times \mathcal{Y}, \mathbb{P}^\otimes)$, and assume that $\log(d\mathbb{P}/d\mathbb{P}^\otimes(\cdot, \cdot)) \in L^2(\mathcal{X} \times \mathcal{Y}, \mathbb{P}^\otimes)$. We can then build increasing models of the form (7.9) developing the function $(x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto \log \frac{d\mathbb{P}}{d\mathbb{P}^\otimes}(x, y)$ according to the above basis functions. Using for instance the first $(1 + d)$ -basis functions, we obtain the following model for $d\mathbb{P}/d\mathbb{P}^\otimes(\cdot, \cdot)$ $h_\theta : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto h_\theta(x, y) = \exp(\alpha + \beta_1\psi_1(x, y) + \dots + \beta_d\psi_d(x, y))$, where $\theta = (\alpha, \beta_1, \dots, \beta_d)^\top \in \Theta \subset \mathbb{R}^{1+d}$. Then, the independence parameter value is $\theta_0 = \mathbf{0} \in \mathbb{R}^{1+d}$.

Example 7.3. Assume that the support of $\mathbb{P}$, $\text{supp}(\mathbb{P}) =: \mathcal{X} \times \mathcal{Y}$, is a known finite-discrete set of size K_1K_2 ; denote by $(\mathbb{P}(x, y))_{(x, y) \in \mathcal{X} \times \mathcal{Y}} := (p_{x, y})_{(x, y) \in \mathcal{X} \times \mathcal{Y}}$ the density of $\mathbb{P}$ with respect to the counting measure on $\mathcal{X} \times \mathcal{Y}$. Then we have

$$\frac{d\mathbb{P}}{d\mathbb{P}^\otimes}(x, y) = \exp \left(\sum_{(a, b) \in \mathcal{X} \times \mathcal{Y}} \theta_{a, b} \mathbb{1}_{\{a\}}(x) \mathbb{1}_{\{b\}}(y) \right), \quad (7.12)$$

where $\theta_{a, b} = \log \frac{p_{a, b}}{p_a p_b}$, $(a, b) \in \mathcal{X} \times \mathcal{Y}$. If we denote, for instance the elements of $\mathcal{X}$ and $\mathcal{Y}$, as $\mathcal{X} := \{a_1, \dots, a_{K_1}\}$ and $\mathcal{Y} := \{b_1, \dots, b_{K_2}\}$, then we can see that $\mathbb{P}$ belongs to the model (7.9) taking

$$h_\theta(x, y) = \exp \left(\alpha + \sum_{(i, j) \neq (1, 1)} \beta_{i, j} \mathbb{1}_{\{a_i\}}(x) \mathbb{1}_{\{b_j\}}(y) \right), \quad (7.13)$$

with the parametrization $\theta = (\alpha, \beta^\top)^\top \in \mathbb{R}^{K_1K_2}$, where α is a scalar and $\beta = (\beta_{i, j})_{(i, j) \neq (1, 1)}$ is the $(K_1K_2 - 1)$ -dimensional vector obtained from the $K_1 \times K_2$ -matrix of real entries $(\beta_{i, j})$ removing the first entry $\beta_{1, 1}$. Moreover, we have for the true value θ_T , $\alpha_T = \log \frac{p_{a_1, b_1}}{p_{a_1} p_{b_1}}$, and $\beta_{i, jT} = \log \frac{p_{a_i, b_j}}{p_{a_i} p_{b_j}} - \log \frac{p_{a_1, b_1}}{p_{a_1} p_{b_1}}$, for all $(i, j) \in \{1, \dots, K_1\} \times \{1, \dots, K_2\} \setminus \{(1, 1)\}$, and that the number of free-parameters in θ_T is equal to $(K_1 - 1)(K_2 - 1)$. Moreover, we have $\theta_0 = (0, \dots, 0)^\top \in \mathbb{R}^{K_1K_2}$.

Example 7.4. Assume that the distribution $\mathbb{P}$ of the random vector $(X, Y) \in \mathbb{R}^2$ is of continuous marginals. The copula $C(\cdot, \cdot)$ of the vector (X, Y) , see e.g. [Nelsen \(2006\)](#), is defined, $\forall(u, v) \in]0, 1[^2$, by $C(u, v) := F(F_1^{-1}(u), F_2^{-1}(v))$, where $F(\cdot, \cdot)$ is the cumulative distribution function of the vector (X, Y) , and F_1 and F_2 are the (marginal) cumulative distribution functions of X and Y , respectively. The copula $C(\cdot, \cdot)$ is in itself a distribution function on $]0, 1[^2$. If $F(\cdot, \cdot)$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$, then we have the relation $\frac{d\mathbb{P}}{d\mathbb{P}^\otimes}(x, y) = \frac{f(x, y)}{f_1(x)f_2(y)} = c(F_1(x), F_2(y))$, where $f(\cdot, \cdot)$ is the joint density of (X, Y) , f_1 and f_2 are the marginal densities of X and Y , and $c(\cdot, \cdot)$ the copula density. Numerous parametric examples of the model (7.9) can then be obtained taking the function

$$h_\theta(x, y) = c_\beta(F_{1,\gamma_1}(x), F_{2,\gamma_2}(y)) \quad (7.14)$$

where $\{c_\beta(\cdot, \cdot); \beta \in D \subset \mathbb{R}^m\}$ is some parametric copula density model, see e.g. [Nelsen \(2006\)](#) or [Joe \(1997\)](#) for examples of such models, and $\{F_{1,\gamma_1}; \gamma_1 \in \Gamma_1\}$ and $\{F_{2,\gamma_2}; \gamma_2 \in \Gamma_2\}$ are some parametric models for the marginal distribution functions. Here, the parameter of interest is $\theta := (\gamma_1, \gamma_2, \beta) \in \Theta := \Gamma_1 \times \Gamma_2 \times D$. Note that the assumption (A.2) is generally not satisfied for this particular model. In fact, if we denote β_0 the particular value corresponding to the copula of independence, then we have $h_{(\gamma_1, \gamma_2, \beta_0)}(\cdot, \cdot) = 1$ for any $(\gamma_1, \gamma_2) \in \Gamma_1 \times \Gamma_2$. Although assumption (A.2) is generally not satisfied, models (7.14) can be used in estimating φ -MI under the assumption that the marginals are dependent.

Example 7.5. We can also deal with semiparametric models induced by semiparametric models of copula densities, with nonparametric unknown continuous marginal distribution functions $F_1(\cdot)$ and $F_2(\cdot)$, taking $h_\theta(x, y) = c_\theta(F_1(x), F_2(y)); \theta \in \Theta \subset \mathbb{R}^d$.

7.2.3 Dual representation and dual estimation of φ -MI

We define estimates of φ -MI by taking advantage of the modeling (7.9) and the dual representation of φ -divergences obtained in [Keziou \(2003a\)](#) and [Broniatowski and Keziou \(2006\)](#). Denote $\varphi^*(\cdot)$ the convex conjugate of the convex function $\varphi(\cdot)$, namely, the function defined by $\varphi^* : t \in \mathbb{R} \mapsto \varphi^*(t) := \sup_{x \in \mathbb{R}} \{tx - \varphi(x)\} \in \mathbb{R} \cup \{+\infty\}$. Note that $\varphi^*(\cdot)$ is, in turn, a proper closed convex function, in particular, $\varphi^*(0) = 0$. Assume that $\varphi(\cdot)$ is essentially smooth, i.e., differentiable on $]a_\varphi, b_\varphi[$ with $\lim_{x \downarrow a_\varphi} \varphi'(x) = -\infty$ if $a_\varphi > -\infty$ and $\lim_{x \uparrow b_\varphi} \varphi'(x) = +\infty$ if $b_\varphi < +\infty$. This is equivalent to the condition that $\varphi^*(\cdot)$ is strictly convex on its domain. Provided that

(A.3) the φ -mutual information $I_\varphi(\mathbb{P}) < \infty$,

see its definition (7.7), it can be rewritten under the form

$$I_\varphi(\mathbb{P}) = \sup_{f \in \mathcal{F}} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} f(x, y) d\mathbb{P}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} \varphi^*(f(x, y)) d\mathbb{P}^\otimes(x, y) \right\}, \quad (7.15)$$

where $\mathcal{F}$ is any class, of measurable real-valued functions $f : \mathcal{X} \times \mathcal{Y} \rightarrow \overline{\mathbb{R}}$, that contains the particular function $\varphi'(d\mathbb{P}/d\mathbb{P}^\otimes)$ and satisfies the condition $\int_{\mathcal{X} \times \mathcal{Y}} |f| d\mathbb{P} < \infty$, for all $f \in \mathcal{F}$. Note that, for all $x \in (a_\varphi, b_\varphi)$, we have $\varphi^*(\varphi'(x)) = x\varphi'(x) - \varphi(x)$. We refer to Table 1.1 for explicit formulas of convex conjugates of some standard divergences. From (7.15), taking into account the model (7.9) by specifying $\mathcal{F} = \{\varphi'(h_\theta); \theta \in \Theta\}$, and assuming in addition that

(A.4) for all $\theta \in \Theta$, we have $\int_{\mathcal{X} \times \mathcal{Y}} |\varphi'(h_\theta(x, y))| d\mathbb{P}(x, y) < \infty$,

we obtain

$$I_\varphi(\mathbb{P}) = \sup_{\theta \in \Theta} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} \varphi'(h_\theta(x, y)) d\mathbb{P}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} \varphi^*(\varphi'(h_\theta(x, y))) d\mathbb{P}^\otimes(x, y) \right\}. \quad (7.16)$$

Moreover, the supremum is unique and achieved in $\theta = \theta_T$. The uniqueness of the supremum θ_T follows from the strict convexity of $\varphi^*(\cdot)$ and the identifiability assumption (A.1). We propose then the following

“dual” estimate of $I_\varphi(\mathbb{P})$

$$\begin{aligned}\widehat{I}_\varphi &:= \sup_{\theta \in \Theta} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} \varphi'(h_\theta(x, y)) \, d\widehat{\mathbb{P}}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} \varphi^*(\varphi'(h_\theta(x, y))) \, d\widehat{\mathbb{P}}_1 \otimes \widehat{\mathbb{P}}_2(x, y) \right\} \\ &= \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi'(h_\theta(X_i, Y_i)) - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \varphi^*(\varphi'(h_\theta(X_i, Y_j))) \right\},\end{aligned}\quad (7.17)$$

and the following “dual” estimate of the parameter θ_T

$$\begin{aligned}\widehat{\theta}_\varphi &:= \arg \sup_{\theta \in \Theta} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} \varphi'(h_\theta(x, y)) \, d\widehat{\mathbb{P}}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} \varphi^*(\varphi'(h_\theta(x, y))) \, d\widehat{\mathbb{P}}_1 \otimes \widehat{\mathbb{P}}_2(x, y) \right\} \\ &= \arg \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi'(h_\theta(X_i, Y_i)) - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \varphi^*(\varphi'(h_\theta(X_i, Y_j))) \right\},\end{aligned}\quad (7.18)$$

where $\widehat{\mathbb{P}}(\cdot)$ is the empirical distribution, associated to the sample, given by (7.8). For ease of presentation, define, $\forall \theta \in \Theta$ and $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$, the functions

$$f_\theta(x, y) := \varphi'(h_\theta(x, y)), \quad (7.19)$$

$$g_\theta(x, y) := \varphi^*(\varphi'(h_\theta(x, y))) = h_\theta(x, y) \varphi'(h_\theta(x, y)) - \varphi(h_\theta(x, y)), \quad (7.20)$$

which we assume to be continuous, in θ , on the set Θ ,

$$M : \theta \in \Theta \mapsto M(\theta) := \int_{\mathcal{X} \times \mathcal{Y}} f_\theta(x, y) \, d\mathbb{P}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} g_\theta(x, y) \, d\mathbb{P}_1 \otimes \mathbb{P}_2(x, y) \quad (7.21)$$

and its empirical version

$$M_n : \theta \in \Theta \mapsto M_n(\theta) := \int_{\mathcal{X} \times \mathcal{Y}} f_\theta(x, y) \, d\widehat{\mathbb{P}}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} g_\theta(x, y) \, d\widehat{\mathbb{P}}_1 \otimes \widehat{\mathbb{P}}_2(x, y). \quad (7.22)$$

Therefore, the formula (7.16) becomes

$$I_\varphi(\mathbb{P}) = \sup_{\theta \in \Theta} M(\theta) = M(\theta_T), \quad \text{and} \quad \theta_T = \arg \sup_{\theta \in \Theta} M(\theta). \quad (7.23)$$

The estimates (7.17) and (7.18), in turn, can be written as

$$\widehat{I}_\varphi = \sup_{\theta \in \Theta} M_n(\theta) = M_n(\widehat{\theta}_\varphi) \quad (7.24)$$

and

$$\widehat{\theta}_\varphi = \arg \sup_{\theta \in \Theta} M_n(\theta). \quad (7.25)$$

Note that the functions $f_\theta(\cdot, \cdot)$, $g_\theta(\cdot, \cdot)$, $M(\cdot)$ and $M_n(\cdot)$ all depend on $\varphi(\cdot)$, but the subscript φ is omitted for simplicity.

For the two-sample problem, with independent samples, say $X_{1,1}, \dots, X_{1,n_1}$ drawn i.i.d. from Q , and $X_{0,1}, \dots, X_{0,n_0}$ drawn i.i.d. from P , [Nguyen et al. \(2010\)](#) considered the problem of estimating $D_\varphi(Q, P)$ and the ratio $r(\cdot) := \frac{dQ}{dP}(\cdot)$, through the dual representation of φ -divergences, using the same method as that proposed in [Keziou \(2003b\)](#) Chapter 2, and extended the approach to the case where the ratio $r(\cdot)$ may lay in infinite-dimensional function space; The estimates are obtained by replacing, in the dual representation, the distributions Q and P , respectively, by $\widehat{Q}(\cdot) := \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{X_{1,i}}(\cdot)$ and $\widehat{P}(\cdot) := \frac{1}{n_0} \sum_{i=1}^{n_0} \delta_{X_{0,i}}(\cdot)$, the empirical measures, associated to the above two samples. For other works on the subject, in the same framework of two-sample problem, particularly, for homogeneity test, one can also

refer to Keziou and Leoni-Aubin (2005), Keziou and Leoni-Aubin (2008), Kanamori *et al.* (2012) and Sugiyama *et al.* (2012b). The main difference of our proposed estimates $\hat{I}_\varphi$, with the above ones, lies in the fact that they are functionals of the joint empirical measure $\hat{\mathbb{P}}(\cdot, \cdot)$ and the product empirical measure $\hat{\mathbb{P}}_1 \otimes \hat{\mathbb{P}}_2$, and that the study of their asymptotic properties is quite different from the above ones, due to the loss of independency between the terms of the first and second sums in (7.17). On the other hand, in the present work, we are more concerned by the independent test problem, and by the power optimality of the proposed test statistics in Bahadur sense.

Example 7.6. *In the context of finite-discrete distributions, using the exponential model described in Example 7.3, we show that the proposed dual estimate (7.17) of $I_\varphi(\mathbb{P})$, obtained by the above “duality” technique, equals the direct plug-in one*

$$\hat{I}_\varphi^{\text{emp}} := I_\varphi(\hat{\mathbb{P}}) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \varphi \left(\frac{\hat{p}_{x,y}}{\hat{p}_x \hat{p}_y} \right) \hat{p}_x \hat{p}_y. \quad (7.26)$$

Indeed, we have by its proper definition

$$\hat{I}_\varphi = \sup_{\theta \in \Theta} M_n(\theta), \text{ where } M_n(\theta) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left[\varphi'(e^{\theta_{x,y}}) \hat{p}_{x,y} - e^{\theta_{x,y}} \varphi'(e^{\theta_{x,y}}) \hat{p}_x \hat{p}_y + \varphi(e^{\theta_{x,y}}) \hat{p}_x \hat{p}_y \right]. \quad (7.27)$$

Differentiating (7.27) with respect to $\theta_{x,y}$ for $(x,y) \in \mathcal{X} \times \mathcal{Y}$ yields

$$\frac{\partial}{\partial \theta_{x,y}} M_n(\theta) = \varphi''(e^{\theta_{x,y}}) \left(e^{\theta_{x,y}} \hat{p}_{x,y} - e^{2\theta_{x,y}} \hat{p}_x \hat{p}_y \right).$$

Canceling derivatives $\frac{\partial}{\partial \theta_{x,y}} M_n(\theta)$ yields $\hat{\theta}_{x,y} = \log \frac{\hat{p}_{x,y}}{\hat{p}_x \hat{p}_y}$, $(x,y) \in \mathcal{X} \times \mathcal{Y}$, which is independent from the choice of φ for this particular model. Finally, straightforward simplifications yield $\hat{I}_\varphi = M_n(\hat{\theta}) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \varphi \left(\frac{\hat{p}_{x,y}}{\hat{p}_x \hat{p}_y} \right) \hat{p}_x \hat{p}_y = \hat{I}_\varphi^{\text{emp}}$. Particularly, for $\varphi(x) = \varphi_2(x) := (x-1)^2/2$, the estimate $\hat{I}_{\varphi_2}$ of the χ^2 -mutual information – or χ^2 measure of independence – obtained by the duality technique is shown to equal (up to the factor $2n$) the classical χ^2 statistics. Hence, in the context of finite-discrete distributions, using the exponential model described in Example 7.3, we see that the proposed approach, via duality technique, recovers the classical direct plug-in one, in particular, the well-known classical χ^2 -independence test.

Remark 7.1. *For finite discrete distributions (with known support, of size say K , see Example 7.3), as in plug-in estimation of Shannon entropy (see e.g. Chao and Shen (2003)), the direct plug-in estimates $\hat{I}_\varphi^{\text{emp}}$ are valid with small bias if the sample size $n \gg K$. If the sample size n is not sufficiently large compared to the space size K , models $h_\theta(\cdot)$ other than (7.12) should be used (through e.g. the model selection procedure described in Section 7.2.4), with small parameter dimension, and the corresponding dual estimate $\hat{I}_\varphi$, if the model $h_\theta(\cdot)$ is correctly specified, could be more promising than the direct plug-in one $\hat{I}_\varphi^{\text{emp}}$.*

Example 7.7. *Note that when dealing with semiparametric copula models $h_\theta(x,y) = c_\theta(F_1(x), F_2(y))$, with unknown nonparametric cumulative distribution functions F_1 and F_2 , it is necessary to estimate them, using for example their empirical counterparts. Denote by $\hat{F}_1(\cdot)$ and $\hat{F}_2(\cdot)$ the empirical cumulative distribution functions associated, respectively, to the samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$, i.e., $\hat{F}_1(x) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{]-\infty, x]}(X_i)$ and $\hat{F}_2(y) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{]-\infty, y]}(Y_i)$. So that $\hat{I}_\varphi$ and $\hat{\theta}_\varphi$ become*

$$\begin{aligned} \hat{I}_\varphi &= \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi' \left(c_\theta \left(\hat{F}_1(X_i), \hat{F}_2(Y_i) \right) \right) - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \varphi^* \left(\varphi' \left(c_\theta \left(\hat{F}_1(X_i), \hat{F}_2(Y_j) \right) \right) \right) \right\} \\ \hat{\theta}_\varphi &= \arg \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi' \left(c_\theta \left(\hat{F}_1(X_i), \hat{F}_2(Y_i) \right) \right) - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \varphi^* \left(\varphi' \left(c_\theta \left(\hat{F}_1(X_i), \hat{F}_2(Y_j) \right) \right) \right) \right\}. \end{aligned}$$

Note that $n\hat{F}_1(X_i)$ is the rank of X_i in the sample $X_1, \dots, X_n$ and $n\hat{F}_2(X_j)$ is the rank of Y_j in the sample $Y_1, \dots, Y_n$. For some copula models, the copula density $c_\theta(u_1, u_2)$ may be unbounded when either u_1 or u_2 tends to 1; see e.g. [Genest et al. \(1995\)](#). In this case, to avoid this difficulty, the “rescaled” empirical cumulative distribution functions $\tilde{F}_1(\cdot) := \frac{n}{n+1}\hat{F}_1(\cdot)$, and $\tilde{F}_2(\cdot) := \frac{n}{n+1}\hat{F}_2(\cdot)$ should be used instead of the standard ones $\hat{F}_1(\cdot)$ and $\hat{F}_2(\cdot)$.

7.2.4 A model selection procedure for the ratio $d\mathbb{P}/d\mathbb{P}^\otimes$ through φ -MI criterion

Let $\mathcal{M}_{\Theta_1} := \{h_{\theta_{1,1}}(\cdot, \cdot); \theta_1 \in \Theta_1 \subset \mathbb{R}^{d_1}\}, \dots, \mathcal{M}_{\Theta_L} := \{h_{\theta_{L,L}}(\cdot, \cdot); \theta_L \in \Theta_L \subset \mathbb{R}^{d_L}\}$ be L candidate models for the ratio $d\mathbb{P}/d\mathbb{P}^\otimes$. For any model $\mathcal{M}_{\Theta_\ell}$, denote by $\hat{\theta}_\ell$ the estimate of $\theta_{\ell T}$ given by $\hat{\theta}_\ell := \arg \sup_{\theta_\ell \in \Theta_\ell} M_n(\theta_\ell)$. The corresponding “expected” criterion is

$M(\hat{\theta}_\ell) = \int_{\mathcal{X} \times \mathcal{Y}} f_{\hat{\theta}_\ell}(x, y) d\mathbb{P}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} g_{\hat{\theta}_\ell}(x, y) d\mathbb{P}(x, y)^\otimes$. From the representation (7.23), we can see that the larger the expected criterion $M(\hat{\theta}_\ell)$ of the model is, the closer the model is to the true one. We propose then the following k -fold cross-validation procedure for model selection using the proposed estimate (7.24) of φ -MI.

- (1) Partition the sample $(X_1, Y_1), \dots, (X_n, Y_n)$ into k equal size (n_k) subsamples. (Denote the i -th subsample $(X_{(i-1)n_k+1}, Y_{(i-1)n_k+1}), \dots, (X_{in_k}, Y_{in_k})$, for all $i = 1, \dots, k$);
- (2) Consider a candidate model $\mathcal{M}_{\Theta_\ell}$;
- (3) From the sample $(X_1, Y_1), \dots, (X_n, Y_n)$ remove the i -th subsample; compute the estimate $\hat{\theta}_\ell^{(-i)}$ given by (7.25) using the remaining $n - n_k$ observations, i.e., $\hat{\theta}_\ell^{(-i)} = \arg \sup_{\theta_\ell \in \Theta_\ell} M_{n-n_k}(\theta_\ell)$;
- (4) Repeat steps (2) and (3) for all $i = 1, \dots, k$, and obtain the following “estimate”

$$C_V(\mathcal{M}_{\Theta_\ell}) := \frac{1}{k} \sum_{i=1}^k \left(\frac{1}{n_k} \sum_{j=(i-1)n_k+1}^{in_k} f_{\hat{\theta}_\ell^{(-i)}}(X_j, Y_j) - \frac{1}{n_k^2} \sum_{j,m=(i-1)n_k+1}^{in_k} g_{\hat{\theta}_\ell^{(-i)}}(X_j, Y_m) \right)$$

of the expected criterion $M(\hat{\theta}_\ell)$, i.e., $M(\hat{\theta}_\ell) = \int_{\mathcal{X} \times \mathcal{Y}} f_{\hat{\theta}_\ell}(x, y) d\mathbb{P}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} g_{\hat{\theta}_\ell}(x, y) d\mathbb{P}(x, y)^\otimes$;

- (5) Repeat steps (2-4) for all $\ell = 1, \dots, L$, and select the “optimal” model $\mathcal{M}_{\Theta_{\ell^*}}$ that maximizes $C_V(\mathcal{M}_{\Theta_\ell})$ over $\ell = 1, \dots, L$, i.e., the model $\mathcal{M}_{\Theta_{\ell^*}}$ with $\ell^* := \arg \sup_{\ell \in \{1, \dots, L\}} C_V(\mathcal{M}_{\Theta_\ell})$.

Other model selection-type procedures can be investigated, through e.g. correcting the bias of $M_n(\hat{\theta}_\ell)$ as an estimate of the expected criterion $M(\hat{\theta}_\ell)$, and selecting the model that maximizes the obtained information criterion corrected from bias. The correction can be made e.g. by asymptotic evaluation of the bias as in classical AIC criterion, or using bootstrap; see e.g. [Konishi and Kitagawa \(2008\)](#) and [Shao and Tu \(1995\)](#).

7.3 Asymptotic properties of the estimates

We state in Section 7.3.1 the consistency of both estimates $\hat{I}_\varphi$ and $\hat{\theta}_\varphi$, of the φ -MI and the parameter θ_T . Section 7.3.2 gives, under the null hypothesis of independence, the limiting distribution of the estimate $\hat{I}_{\varphi_1}$ of the KL-MI, as well as the corresponding estimate $\hat{\theta}_{\varphi_1}$ of the parameter θ_T , for some specific forms of the model $\{h_\theta(\cdot, \cdot); \theta \in \Theta\}$. Section 7.3.3 provides bootstrap calibration of the critical value of any $\hat{I}_\varphi$ -based test statistic for general forms of the model $\{h_\theta(\cdot, \cdot); \theta \in \Theta\}$.

7.3.1 Consistency

In this section, we state consistency of the estimate $\hat{I}_\varphi$, of the φ -MI, defined by (7.17), as well as the consistency of the estimates $\hat{\theta}_\varphi$ of θ_T . We will use classical techniques from M-estimation theory. We will make use of the following conditions.

(A.5) The parameter space Θ is a compact subset of $\mathbb{R} \times \mathbb{R}^d$;

(A.6) $\int_{\mathcal{X} \times \mathcal{Y}} \sup_{\theta \in \Theta} |f_\theta(x, y)| \, d\mathbb{P}(x, y) < \infty$;

(A.7) $\int_{\mathcal{X} \times \mathcal{Y}} \sup_{\theta \in \Theta} g_\theta(x, y)^2 \, d\mathbb{P}^\otimes(x, y) < \infty$,

where f_θ and g_θ are defined, respectively, by (7.19) and (7.20). Note that assumptions (A.6-7) imply (A.3-4).

Proposition 7.1. *Assume that conditions (A.1, 5-7) hold. Then, the estimates $\hat{I}_\varphi$ of $I_\varphi(\mathbb{P})$ defined by (7.17) and the estimates $\hat{\theta}_\varphi$ of θ_T defined by (7.18) are consistent. Precisely, as $n \rightarrow \infty$, the following convergences in probability hold $\hat{I}_\varphi \rightarrow I_\varphi(\mathbb{P})$ and $\hat{\theta}_\varphi \rightarrow \theta_T$.*

Remark 7.2. *Since in practice, all models are generally “misspecified”, the true parameter value θ_T may not exist, it can however be replaced by the “pseudo-true” value $\theta_T^* := \arg \sup_{\theta \in \Theta} M(\theta)$, and the results of consistency in the above proposition remain valid.*

7.3.2 The limiting distribution of the estimate $\hat{I}_{\varphi_1}$ of KL-MI

We will give now the limiting distribution of the particular statistical test based on the estimate $\hat{I}_{\varphi_1}$ of classical KL-MI, for specific forms of the model $h_\theta(\cdot, \cdot)$, under the null hypothesis of independence $\mathcal{H}_0 : \mathbb{P} = \mathbb{P}^\otimes$. Consider the following specific form of the model $h_\theta(\cdot, \cdot)$

$$h_\theta(x, y) = \exp(\alpha + m_\beta(x, y)) \quad \text{with} \quad m_\beta(x, y) := \sum_{k=1}^d \beta_k \xi_k(x) \zeta_k(y), \quad (7.28)$$

for some specified measurable real valued functions ξ_k and ζ_k , $k = 1, \dots, d$, defined, respectively, on $\mathcal{X}$ and $\mathcal{Y}$. The parameter θ is the vector $\theta := (\alpha, \beta_1, \dots, \beta_d)^\top \in \Theta \subset \mathbb{R} \times \mathbb{R}^d$. In this case, (7.19) and (7.20) become $f_\theta(x, y) = \alpha + \sum_{k=1}^d \beta_k \xi_k(x) \zeta_k(y)$ and $g_\theta(x, y) = \exp\left(\alpha + \sum_{k=1}^d \beta_k \xi_k(x) \zeta_k(y)\right) - 1$. The value θ_0 , corresponding to the independence, here is $\theta_0 = \mathbf{0} \in \mathbb{R}^{1+d}$. We will give the limiting distributions of $\hat{\theta}_{\varphi_1}$ and $\hat{I}_{\varphi_1}$, under the null hypothesis of independence $\mathbb{P} = \mathbb{P}^\otimes$, i.e., when $\theta_T = \theta_0 = \mathbf{0}$. We will consider the following assumptions.

(A.8) There exists a neighborhood $N(\theta_T)$ of θ_T such that the third order partial derivative functions $\{(x, y) \mapsto (\partial^3 / \partial^3 \theta) f_\theta(x, y); \theta \in N(\theta_T)\}$ (resp. $\{(x, y) \mapsto (\partial^3 / \partial^3 \theta) g_\theta(x, y); \theta \in N(\theta_T)\}$) are dominated by some functions $\mathbb{P}$ -integrable (resp. some function $\mathbb{P}^\otimes$ -square-integrable);

(A.9) The integrals $\mathbb{P} \|f'_{\theta_T}\|^2$, $\mathbb{P}^\otimes \|g'_{\theta_T}\|^2$, $\mathbb{P} \|f''_{\theta_T}\|$, $\mathbb{P}^\otimes \|g''_{\theta_T}\|^2$ exist, and the matrix

$$\Sigma_1 := -(\mathbb{P} f''_{\theta_T} - \mathbb{P}^\otimes g''_{\theta_T}) \quad (7.29)$$

is nonsingular.

Theorem 7.1. *Assume that conditions (A.1-2, 5-9) hold and that $\mathbb{P} = \mathbb{P}^\otimes$ (i.e., $\theta_T = \mathbf{0}$). Then,*

(a) $\sqrt{n} \hat{\theta}_{\varphi_1}$ converges in distribution to a centered multivariate normal random variable with covariance matrix $\Sigma = \Sigma_1^{-1} \Sigma_2 \Sigma_1^{-1}$, where Σ_1 and Σ_2 are given respectively by (7.29) and $\Sigma_2 := \psi'(\mu) \Sigma \psi'(\mu)^\top$, where $\psi(\cdot)$ is the continuously derivable function defined on $\mathbb{R}^{1+3d}$ into $\mathbb{R}^{1+d}$ by

$$\psi(x_0, x_1, \dots, x_d, y_1, \dots, y_d, z_1, \dots, z_d) = (0, x_1 y_1 - z_1, \dots, x_d y_d - z_d)^\top,$$

and

$$\mu := \mathbb{E}(V(X, Y)) = (1, \mathbb{P}_1 \xi_1, \dots, \mathbb{P}_1 \xi_d, \mathbb{P}_2 \zeta_1, \dots, \mathbb{P}_2 \zeta_d, \mathbb{P}_1 \xi_1 \mathbb{P}_2 \zeta_1, \dots, \mathbb{P}_1 \xi_d \mathbb{P}_2 \zeta_d)^\top.$$

(b) $2n \hat{I}_{\varphi_1}$ converges in distribution to the random variable $Z^\top Z$, where Z is a centered multivariate normal random variable with covariance matrix $C = \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}$.

Remark 7.3. *For the finite-discrete case, using the modeling (7.13) in Example 7.3, we can see that the corresponding matrix Σ_2 is of rank $(K_1 - 1)(K_2 - 1)$ and that the limiting distribution of $2n \hat{I}_\varphi = 2n \hat{I}_\varphi^{\text{emp}}$ is a χ^2 -distribution with $(K_1 - 1)(K_2 - 1)$ degrees of freedom, in particular, we recover the classical χ^2 -independence test theorem (for the case of finite-discrete distributions).*

7.3.3 Bootstrap calibration

In the general context of model (7.9), for a given φ -MI, we propose the following bootstrap procedure to calibrate the critical value of the corresponding test statistic. The critical value, denote it b_α , is the upper α -quantile of the distribution of the test statistic $S_n := 2n\hat{I}_\varphi$, under the null hypothesis $\mathcal{H}_0$ of independence.

1. Generate bootstrap sample $(X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$ from the product empirical distribution $\hat{\mathbb{P}}^\otimes = \hat{\mathbb{P}}_1 \otimes \hat{\mathbb{P}}_2$ of the original sample $(X_1, Y_1), \dots, (X_n, Y_n)$;
2. Compute the value of the statistic $S_n^* := 2n\hat{I}_\varphi^*$ from the bootstrap sample;
3. Repeat steps (1) and (2) $B = 1000$ times, independently, to obtain the realizations $\{S_{n,1}^*, S_{n,2}^*, \dots, S_{n,B}^*\}$;
4. Estimate b_α by $\tilde{b}_\alpha :=$ the $(1 - \alpha)$ th quantile of the sequence $\{S_{n,1}^*, S_{n,2}^*, \dots, S_{n,B}^*\}$.

7.4 Large deviations principle and Bahadur asymptotic efficiency

In this Section, we compare Bahadur asymptotic efficiency of φ -MI based independence tests and show that the test based on classical Kullback-Leibler mutual information is the most efficient. Given $(\hat{I}_{\varphi_1})_n$ and $(\hat{I}_{\varphi_2})_n$ two sequences of statistics, for the test problem (7.5), numbers $\alpha \in (0, 1)$, $\gamma \in (0, 1)$ and an alternative hypothesis $\mathbb{P} \neq \mathbb{P}^\otimes$, we define $n_i(\alpha, \gamma, \mathbb{P})$, for $i \in \{1, 2\}$, respectively, as the minimal number of observations needed for the test based on $\hat{I}_{\varphi_i}$ to have signification level α and power level γ . Then, Bahadur asymptotic relative efficiency of $(\hat{I}_{\varphi_1})_n$ with respect to $(\hat{I}_{\varphi_2})_n$ is defined as (if the limit exists) $\lim_{\alpha \rightarrow 0} \frac{n_2(\alpha, \gamma, \mathbb{P})}{n_1(\alpha, \gamma, \mathbb{P})}$. It is well known, see for example Nikitin (1995) and van der Vaart (1998) Chapter 14, that if both sequences $(\hat{I}_{\varphi_1})_n$ and $(\hat{I}_{\varphi_2})_n$ satisfy a large deviation principle under the null hypothesis (with good rate functions $e_{\varphi_1}(\cdot)$ and $e_{\varphi_2}(\cdot)$) and also a law of large number under a given alternative hypothesis $\mathcal{H}_1 : \mathbb{P} \neq \mathbb{P}^\otimes$, with asymptotic means $\mu_{\varphi_1}(\mathbb{P})$ and $\mu_{\varphi_2}(\mathbb{P})$, respectively, then the Bahadur asymptotic relative efficiency equals $e_{\varphi_1}(\mu_{\varphi_1}(\mathbb{P}))/e_{\varphi_2}(\mu_{\varphi_2}(\mathbb{P}))$. Particularly, the most efficient test maximizes Bahadur slope $e_\varphi(\mu_\varphi(\mathbb{P}))$. A law of large number under the alternative hypothesis is given for the sequence $(\hat{I}_\varphi)_n$ in Proposition 7.1 above; the expected value $\mu_\varphi(\mathbb{P})$ being $\mu_\varphi(\mathbb{P}) = I_\varphi(\mathbb{P}) = D_\varphi(\mathbb{P}, \mathbb{P}^\otimes)$. The following theorem establishes a large deviation principle under the null hypothesis of independence. It relies on some generalization due to Eichelsbacher and Schmock (2002) of classical Sanov theorem to finer topologies and the contraction principle. Let $\mathcal{G}$ be the set of measurable functions, from $\mathcal{X} \times \mathcal{Y}$ into $\mathbb{R}$, given by $\mathcal{G} := \mathcal{F}_b \cup \{\varphi'(h_\theta); \theta \in \Theta\} \cup \{\varphi^*(\varphi'(h_\theta)); \theta \in \Theta\}$, where $\mathcal{F}_b$ is the set of all measurable bounded functions from $\mathcal{X} \times \mathcal{Y}$ into $\mathbb{R}$. Recall that $M^1 := M^1(\mathcal{X} \times \mathcal{Y})$ is the set of all probability measures on $\mathcal{X} \times \mathcal{Y}$, and let us introduce the subset

$$\mathcal{M}_\mathcal{G} := \mathcal{M}_\mathcal{G}(\mathcal{X} \times \mathcal{Y}) := \left\{ P \in M^1 : \int_{\mathcal{X} \times \mathcal{Y}} |\varphi'(h_\theta)| dP < \infty, \int_{\mathcal{X} \times \mathcal{Y}} |\varphi^*(\varphi'(h_\theta))| dP^\otimes < \infty, \forall \theta \in \Theta \right\}.$$

Define on $\mathcal{M}_\mathcal{G}$ the $\tau_\mathcal{G}$ -topology as the coarsest one that makes applications $P \in \mathcal{M}_\mathcal{G} \mapsto \int_{\mathcal{X} \times \mathcal{Y}} \varphi'(h_\theta) dP$, $P \in \mathcal{M}_\mathcal{G} \mapsto \int_{\mathcal{X} \times \mathcal{Y}} f dP$, $P \in \mathcal{M}_\mathcal{G} \mapsto \int_{\mathcal{X} \times \mathcal{Y}} \varphi^*(\varphi'(h_\theta)) dP^\otimes$ and $P \in \mathcal{M}_\mathcal{G} \mapsto \int_{\mathcal{X} \times \mathcal{Y}} f dP^\otimes$ continuous, for all $\theta \in \Theta$ and all $f \in \mathcal{F}_b$. Finally, define, for all $Q \in \mathcal{M}_\mathcal{G}$, the “pseudo-divergence”

$$\mathcal{D}_\varphi(Q, Q^\otimes) := \sup_{\theta \in \Theta} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} \varphi'(h_\theta(x, y)) dQ(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} \varphi^*(\varphi'(h_\theta(x, y))) dQ^\otimes(x, y) \right\}.$$

Obviously, $\mathcal{D}_\varphi(Q, Q^\otimes) \leq D_\varphi(Q, Q^\otimes) =: I_\varphi(Q)$ with equality for probability distributions such that $dQ/dQ^\otimes = h_\theta$ for some $\theta \in \Theta$. Note also that $Q \in \mathcal{M}_\mathcal{G} \mapsto \mathcal{D}_\varphi(Q, Q^\otimes)$ is continuous with respect to the $\tau_\mathcal{G}$ -topology as the supremum over the compact set Θ of continuous functions. The large deviation principle for the sequence $(\hat{\mathbb{P}}(\cdot))_n$ of empirical measures defined by (7.8), established by Eichelsbacher and Schmock (2002), requires the existence of exponential moments; in the context of the model (7.9), we thus assume

$$(A.10) \text{ for all } f \in \mathcal{G}, \text{ for all } a > 0, \quad \int_{\mathcal{X} \times \mathcal{Y}} \exp(a|f(x, y)|) d\mathbb{P}(x, y) < \infty.$$

Note that the strong assumption (A.10) implies (A.3-4) if $\mathbb{P} = \mathbb{P}^\otimes$. In the context of the models described in Examples 7.1 to 7.5, assumption (A.10) may not be satisfied for some φ -divergences ; particularly, it does not generally hold for power-divergences (except for finite-discrete distribution models described in Example 7.3). A sufficient condition for (A.10) is

$$(A.11) \text{ there exist real numbers } m, M \in (a_\varphi, b_\varphi) \text{ such that } m < h_\theta(\cdot, \cdot) < M, \quad \mathbb{P}^\otimes - a.s., \quad \forall \theta \in \Theta.$$

Indeed, for all $a > 0$, the functions $\exp(a|\varphi'(h_\theta)|)$ and $\exp(a|\varphi^*(\varphi'(h_\theta))|)$ are bounded and therefore integrable with respect to both $\mathbb{P}$ and $\mathbb{P}^\otimes$. Again, (A.11) is not generally satisfied for models described in the previous examples for power-divergences, but it may be artificially verified by truncating the distributions in the models. Let us also point out that Theorems 7.2 and 7.3 below may remain true with some alternative assumptions on the distribution queues, lighter than (A.10). Particularly, simulations performed for bivariate Gaussian distributions tend to show that Theorem 7.3 holds for the Gaussian model described in Example 7.1. For getting a closed form for the LDP of $(\hat{I}_\varphi)_n$, we will establish the right-continuity of the rate function, making use of one of the following assumptions:

$$(A.12.a) \text{ } (X, Y) \text{ is finite-discrete, supported by } \mathcal{X} \times \mathcal{Y};$$

$$(A.12.b) \text{ The model } \{h_\theta(\cdot, \cdot); \theta = (\alpha, \beta^\top)^\top \in \Theta\} \text{ is of the form } h_\theta(x, y) = \exp(\alpha + m_\beta(x, y)) \text{ with the condition that, for any constant } c \text{ and any } \beta, \text{ we have } \mathbb{P}^\otimes(m_\beta(X, Y) = c) \neq 0 \text{ iff } \beta = (0, \dots, 0)^\top \text{ and } c = 0.$$

Theorem 7.2. *Let (X, Y) be a couple of independent random variables with joint distribution $\mathbb{P} = \mathbb{P}^\otimes \in \mathcal{M}_\Theta \cap \mathcal{M}_\mathcal{G}$.*

- (1) *Suppose that conditions (A.1-2, 5-7, 10 and 12.b) are satisfied. Then, the sequence $(\hat{I}_\varphi)_n$ of estimates, of $I_\varphi(\mathbb{P}) = 0$, given by (7.17), satisfies the following large deviation principle*

$$\frac{1}{n} \log \mathbb{P}^\otimes(\hat{I}_\varphi > d) \xrightarrow{n \rightarrow \infty} -e_\varphi(d), \quad d > 0, \quad (7.30)$$

where the good rate function $e_\varphi(\cdot)$ is

$$e_\varphi(d) := \inf_{Q \in \Omega_d} \mathbb{K}(Q, \mathbb{P}^\otimes) \quad \text{with } \Omega_d := \{Q \in \mathcal{M}_\mathcal{G} \text{ such that } \mathcal{D}_\varphi(Q, Q^\otimes) \geq d\}. \quad (7.31)$$

- (2) *Assume that conditions (A.1-2, 5 and 12.a) are satisfied. Then the above statement holds if $\mathcal{M}_\mathcal{G}$ is replaced by the set of all discrete-finite distributions with the same finite support $\mathcal{X} \times \mathcal{Y}$.*

In view of Proposition 7.1 and Theorem 7.2 above, the Bahadur slope of the independence test based on $\hat{I}_\varphi$, for any φ , is then $s_\varphi := e_\varphi(I_\varphi(\mathbb{P})) = \inf\{\mathbb{K}(Q, \mathbb{P}^\otimes) : \mathcal{D}_\varphi(Q, Q^\otimes) \geq D_\varphi(\mathbb{P}, \mathbb{P}^\otimes)\}$. Since $\mathcal{D}_\varphi(\mathbb{P}, \mathbb{P}^\otimes) = D_\varphi(\mathbb{P}, \mathbb{P}^\otimes)$, we have $\mathbb{P} \in \{Q : \mathcal{D}_\varphi(Q, Q^\otimes) \geq D_\varphi(\mathbb{P}, \mathbb{P}^\otimes)\}$, so that, for any φ ,

$$s_\varphi \leq \mathbb{K}(\mathbb{P}, \mathbb{P}^\otimes) = I_{KL}(\mathbb{P}) = I_{\varphi_1}(\mathbb{P}). \quad (7.32)$$

Equality is achieved in (7.32) for the divergence $D_\varphi = \mathbb{K}$. Indeed, $s_{KL} = \inf\{\mathbb{K}(Q, \mathbb{P}^\otimes) : \mathcal{D}_{KL}(Q, Q^\otimes) \geq \mathbb{K}(\mathbb{P}, \mathbb{P}^\otimes)\}$. Straightforward computations yield $\mathbb{K}(Q, \mathbb{P}^\otimes) = \mathbb{K}(Q, Q^\otimes) + \mathbb{K}(Q_1, \mathbb{P}_1) + \mathbb{K}(Q_2, \mathbb{P}_2)$, for any $Q \in \mathcal{M}_\mathcal{G}$. Particularly, for any Q such that $\mathcal{D}_{KL}(Q, Q^\otimes) \geq \mathbb{K}(\mathbb{P}, \mathbb{P}^\otimes)$, we have $\mathbb{K}(Q, Q^\otimes) \geq \mathcal{D}_{KL}(Q, Q^\otimes) \geq \mathbb{K}(\mathbb{P}, \mathbb{P}^\otimes)$, hence, $\mathbb{K}(Q, \mathbb{P}^\otimes) \geq \mathbb{K}(Q, Q^\otimes) \geq \mathbb{K}(\mathbb{P}, \mathbb{P}^\otimes)$, so that

$$s_{KL} \geq \mathbb{K}(\mathbb{P}, \mathbb{P}^\otimes). \quad (7.33)$$

Combining (7.32) and (7.33), we obtain

Theorem 7.3. *Let (X, Y) be a couple of random variables with joint distribution $\mathbb{P} \in \mathcal{M}_\Theta \cap \mathcal{M}_\mathcal{G}$. Suppose that either conditions (A.1-2, 5-7, 10 and 12.b) or (A.1-2, 5 and 12.a) are satisfied. For the test problem (7.5), the test based on the estimate $\hat{I}_{\varphi_1}$, see (7.17), of the Kullback-Leibler mutual information, is uniformly (i.e., whatever be the alternative $\mathbb{P} \neq \mathbb{P}^\otimes$) the most efficient test, in Bahadur sense, among all $\hat{I}_\varphi$ -based tests, including the classical χ^2 -independence one.*

Remark 7.4. *Assume that $\mathbb{P}$ is a finite-discrete distribution. We obtain then that KL-MI based independence test is more efficient than the classical χ^2 independence one. This result was already stated, in goodness-of-fit testing for finite-discrete distributions, see e.g. [van der Vaart \(1998\)](#) Chapter 17 Section 17.6. The above theorem extends it to testing independence, for more general probability distributions, not necessarily finite-discrete.*

7.5 Concluding remarks and directions for future research

In this Chapter, we have defined and studied estimates of φ -mutual informations, based on the dual representation of φ -divergences and a semiparametric modeling of the density ratio between the joint distribution of the couple and the product distribution of its marginals. The consistency of these estimates – named dual-estimates – has been established assuming some classical regularity conditions on the model; the asymptotic normality has been established for classical Kullback-Leibler mutual information and specific models by means of classical M-estimation theory arguments. The asymptotic normality of other φ -mutual information dual-estimates may be derived similarly, for specific models depending on the considered φ -divergence. For example, when dealing with the power divergence associated to φ_γ functions given by (7.6), the asymptotic normality of the corresponding φ_γ -mutual-information dual-estimates may be derived in a similar way when focusing on the so-called γ -exponential semiparametric model

$$\mathbb{P} \in \left\{ P \in M^1(\mathcal{X} \times \mathcal{Y}) \text{ s.t. } \frac{dP}{dP^\otimes}(x, y) = \exp_\gamma \left(\sum_{k=0}^d \theta_k \xi_k(x) \zeta_k(y) \right), \theta = (\theta_0, \dots, \theta_d) \in \Theta \right\},$$

where $\exp_\gamma(t) := ((\gamma - 1)t + 1)_+^{\frac{1}{\gamma-1}}$, with $(\cdot)_+ = \max(0, \cdot)$. Our semiparametric approach for estimating mutual informations constitutes a promising alternative to classical nonparametric procedures based on kernel density estimation or adaptive partitioning. No parameters such as bandwidth or kernel type has to be adjusted. The asymptotic normality of dual-estimates is also of significative importance, particularly, for hypothesis-testing purpose. For the sake of both completeness and accessibility, we are developing a package for the R software providing user-ready procedures, including the k -fold cross validation procedure described in Section 7.2.4, for selecting the model that best matches the data. We also aim at comparing the dual-estimates of mutual informations to nonparametric estimates, and at extending our approach to the case where the ratio $\frac{dP}{dP^\otimes}$ lies in infinite-dimensional space. As an application of dual-estimation of mutual informations, we have derived a class of independence tests, recovering as a particular case, the classical χ^2 -independence test. For a large variety of situations including finite-discrete random couples, the most efficient test is based on the KL-MI estimates, outperforming the classical χ^2 -independence one. Motivated by the simulation experiments, we guess that the optimality of MI independence test can be extended to a larger family of models.

Chapter 8

A semiparametric likelihood-type test of independence in copula models

The purpose of this Chapter is to provide limit laws for semiparametric estimators of copulas. Some statistical tests of independence are introduced as a consequence of this methodology. We are primarily concerned with the case where the parameter lies on the boundary of the admissible domain. The results are published in [Bouzebda and Keziou \(2008\)](#).

8.1 Introduction and motivations

It is well known since the work of [Sklar \(1959\)](#) that the joint behavior of a bivariate vector (X_1, X_2) with d.f. $F(x_1, x_2) := P(X_1 \leq x_1, X_2 \leq x_2)$, and continuous marginal d.f.'s $F_i(x_i) := P(X_i \leq x_i)$, $i = 1, 2$, is characterized by the copula (or dependence function) C associated with $F(\cdot, \cdot)$. The copula function is defined, for all $(u_1, u_2) \in (0, 1)^2$, through the identity $C(u_1, u_2) := P\{F_1(X_1) \leq u_1, F_2(X_2) \leq u_2\}$. Many useful multivariate models for dependence between X_1 and X_2 turn out to be generated by *parametric* families of copulas of the form $\{C_\theta; \theta \in \Theta\}$, typically indexed by a vector valued parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ (see e.g. [Kimeldorf and Sampson \(1975b\)](#), [Kimeldorf and Sampson \(1975a\)](#), [Genest and MacKay \(1986\)](#), and [Joe \(1993\)](#)). The nonparametric approach to copula estimation has been initiated by [Deheuvels \(1979b\)](#) who introduced and investigated the *empirical copula process*. In addition, [Deheuvels \(1980, 1981\)](#) described the limiting behavior of this empirical process (see also [Fermanian et al. \(2004\)](#) and the references therein). In this Chapter, we consider semiparametric copula models with unknown general margins. In order to estimate the unknown *true* value of the parameter $\theta \in \Theta$, which we denote, throughout the sequel, by $\theta_T \in \Theta$, some semiparametric estimation procedures, based on the maximization, on the parameter space Θ , of properly chosen *pseudo-likelihood* criterion, have been studied by [Genest et al. \(1995\)](#), [Shih and Louis \(1995\)](#), [Wang and Ding \(2000\)](#) and [Tsukahara \(2005\)](#), among others. In each of these papers, some asymptotic normality properties are established for $\sqrt{n}(\hat{\theta}_n - \theta_T)$, where $\hat{\theta}_n$ denotes a properly chosen estimator of θ_T . This is achieved, provided that θ_T lies in the *interior*, denoted by $\overset{\circ}{\Theta}$, of the parameter space $\Theta \subseteq \mathbb{R}^d$. On the other hand, the case where $\theta_T \in \partial\Theta := \overline{\Theta} - \overset{\circ}{\Theta}$ is a *boundary value* of Θ , has not been studied in a systematical way until present. Moreover, it turns out that, for the above-mentioned estimators, the asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta_T)$, may fail to hold for $\theta_T \in \partial\Theta$. To illustrate this phenomenon, we cite below some examples of parametric copulas, for which marginal independence is verified for some specific values of the parameter θ , on the boundary $\partial\Theta$ of the admissible parameter set Θ . We start with example for which θ varies within subsets of $\mathbb{R}$. Such is the case for the [Gumbel \(1960\)](#) family of copulas, which is one of the most popular model used to model bivariate extreme values. This family of copulas is defined, for $0 < u_1, u_2 < 1$, by

$$C_\theta(u_1, u_2) := \exp \left\{ - \left[(-\log u_1)^\theta + (-\log u_2)^\theta \right]^{1/\theta} \right\}, \quad \text{for } \theta \in [1, \infty); \quad (8.1)$$

the Joe (1993)'one is

$$C_\theta(u_1, u_2) = 1 - \left[(1 - u_1)^\theta + (1 - u_2)^\theta - (1 - u_1)^\theta (1 - u_2)^\theta \right]^{1/\theta}, \quad \theta \in [1, \infty); \quad (8.2)$$

the Galambos (1975)'one is

$$C_\theta(u_1, u_2) = u_1 u_2 \exp \left\{ \left[(-\log u_1)^{-\theta} + (-\log u_2)^{-\theta} \right]^{-1/\theta} \right\}, \quad \theta \in [0, \infty); \quad (8.3)$$

the Hüsler and Reiss (1989)'one is

$$C_\theta(u_1, u_2) = u_2^{\Phi(\theta^{-1} + 0.5\theta \log[(-\log u_2)/(-\log u_1)])} u_1^{\Phi(\theta^{-1} + 0.5\theta \log[(-\log u_1)/(-\log u_2)])}, \quad (8.4)$$

$\theta \in [0, \infty)$, where $\Phi(\cdot)$ denotes the standard normal distribution function. For all these examples, the independence case $C_{\theta_T}(u_1, u_2) = u_1 u_2$ occurs at the boundary of the parameter space Θ , i.e., when $\theta_T = 1$ for models (9.1) and (9.4), and $\theta_T = 0$ for models (8.3) and (9.3). In the sequel, we will denote by θ_0 the value of the parameter (when it exists), corresponding to the independence of the margins, i.e., the value of the parameter for which we have $C_{\theta_0}(u_1, u_2) := u_1 u_2$, for all $(u_1, u_2) \in (0, 1)^2$. We refer to Joe (1993) for more examples in the case when the parameter space Θ is univariate or multivariate. In contrast with the preceding examples, where $\theta_0 \in \partial\Theta$ is a boundary value of Θ , the case where θ_0 is an interior point of Θ may, at times, occur, but is more seldom. An example, where $\theta_0 \in \overset{\circ}{\Theta}$, is given by the Farlie-Gumbel-Morgenstern (FGM) copula, defined by

$$C_\theta(u_1, u_2) := u_1 u_2 + \theta u_1 u_2 (1 - u_1)(1 - u_2), \quad \theta \in \Theta := [-1, 1], \quad (8.5)$$

and for which $\theta_0 = 0 \in \overset{\circ}{\Theta} =]-1, 1[$. In the present Chapter, we will treat, in turn, parametric estimation of θ_T , and tests of the independence assumption $\mathcal{H}_0 : \theta_T = \theta_0$. We consider, successively, the case where $\theta_0 \in \overset{\circ}{\Theta}$ is an interior point of $\Theta \subseteq \mathbb{R}$, and the case where $\theta_0 \in \partial\Theta$ is a boundary value of $\Theta \subseteq \mathbb{R}$. A number of tests of independence have been introduced in a nonparametric setup. A standard procedure of the kind is based on the Kendall τ statistic and the Spearman ρ statistic. We further mention the nonparametric Kolmogorov-Smirnov-type tests of independence introduced by Deheuvels (1979a) and Deheuvels (1981). Another, among the most commonly used tests of independence, is due to Pearson, and based on the correlation coefficient. We refer to Genest and Verret (2005), for additional details and references on this problem. In the remainder of the present Chapter, we will limit ourselves to semiparametric procedures, to test $\mathcal{H}_0$. Typical examples of the kind are given by the, so-called, *score-type* tests. In the usual parametric framework, it is well known that the generalized likelihood ratio (GLR) test, is preferable to score type tests (see, e.g., Nikitin (1995)). This motivates our interest, in the present study, for the generalized “pseudo-likelihood” ratio (GPLR) test of independence (for semiparametric copula models). We will consider tests of the kind, when θ_0 is, either, an interior point, or a boundary point of the admissible parameter space Θ . We conjecture that the GPLR test of independence, should be more powerful, in the semiparametric setup, than score type tests and most classical nonparametric tests of independence (such as the Kendall, Spearman and Kolmogorov-Smirnov-type tests.) In the sequel, we will study the asymptotic behavior of the maximum “pseudo-likelihood” (MPL) estimator introduced by Oakes (1994) and studied by Genest et al. (1995). We will also investigate the properties of the GPLR statistic, to test the null hypothesis $\mathcal{H}_0 : \theta_T = \theta_0$ against the alternative $\mathcal{H}_1 : \theta_T \neq \theta_0$, when $\theta_0 \in \partial\Theta$ is a boundary point of Θ . We will show that, in general, the asymptotic normality of the MPL estimate $\tilde{\theta}_n$, proposed by Genest et al. (1995), does not hold when $\theta_T \in \partial\Theta$ is a boundary value of the admissible parameter space Θ . Likewise, the asymptotic distribution of the GPLR statistic is, in general, distinct from the χ^2 distribution, when $\theta_T \in \partial\Theta$. In the special case where the parameter is univariate, with $\Theta \subseteq \mathbb{R}$, we will provide the limiting asymptotic distributions, under the null hypothesis $\mathcal{H}_0$ of independence, of both, the MPL estimator and the GPLR statistic. This will allow us to introduce a new test of independence based on this GPLR statistic. The remainder of the present Chapter is organized as follows. In §8.2, we recall the principles underlying the MPL estimation procedure, and study, in the univariate case, the

asymptotic properties of the corresponding estimate $\tilde{\theta} = \tilde{\theta}_n$ of θ_T , when $\theta_T \in \partial\Theta$ is a boundary value of the parameter space Θ . It will become clear that the classical property of asymptotic normality does not always hold in this case. The limiting asymptotic distribution of the test statistic under the alternative hypothesis is used to derive an approximation of the power function. An application of the forthcoming results will allow us to evaluate the sample size necessary to guarantee a pre-assigned power level, with respect to a specified alternative. To establish our results, we use similar arguments as those developed by [Genest et al. \(1995\)](#) and [Tsukahara \(2005\)](#) in connection with the instrumental statements on rank statistics established by [Ruymgaart et al. \(1972\)](#) and [Rüschendorf \(1976\)](#) among others, combined with a new technique (based on the law of iterated logarithm given in Lemma 1 in [Bouzebda and Keziou \(2008\)](#)) to show the existence of the MPL estimator.

8.2 The maximum pseudo-likelihood procedure

8.2.1 Estimation

We recall the principles of the maximum pseudo-likelihood procedure introduced by [Oakes \(1994\)](#) and studied by [Genest et al. \(1995\)](#). Consider a random sample $\{(X_{1k}, X_{2k}); k = 1, \dots, n\}$ from the distribution of (X_1, X_2) , denoted by $F_{\theta_T}(x_1, x_2) := C_{\theta_T}(F_1(x_1), F_2(x_2)) = P(X_1 \leq x_1, X_2 \leq x_2)$. The corresponding log-likelihood is then defined by

$$\theta \in \Theta \mapsto L^*(\theta) := \sum_{k=1}^n \log \left[c_{\theta} \left(F_1(X_{1k}), F_2(X_{2k}) \right) \right], \quad (8.6)$$

where $c_{\theta}(\cdot, \cdot) := \frac{\partial^2}{\partial u_1 \partial u_2} C_{\theta}(\cdot, \cdot)$ denotes the density of $C_{\theta}(\cdot, \cdot)$ on $(0, 1)^2$, which we assume to exist, and to be continuous on $(0, 1)^2$ for all $\theta \in \Theta$. Since the margins, F_1 and F_2 , are assumed to be unknown, [Oakes \(1994\)](#) propose to maximize, on the parameter space Θ , the log pseudo-likelihood

$$\theta \in \Theta \mapsto L(\theta) := \sum_{k=1}^n \log \left[c_{\theta} \left(\hat{F}_{1n}(X_{1k}), \hat{F}_{2n}(X_{2k}) \right) \right], \quad (8.7)$$

where, for $j = 1, 2$, $\hat{F}_{jn}(\cdot)$ stands for $n/(n+1)$ times the marginal empirical distribution function of the j -th variable X_j . We have, namely, for $t \in \mathbb{R}$,

$$\hat{F}_{jn}(t) := \frac{n}{n+1} F_{jn}(t), \quad j = 1, 2, \quad (8.8)$$

where

$$F_{jn}(t) := \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{]-\infty, t]}(X_{jk}), \quad j = 1, 2, \quad (8.9)$$

and where $\mathbb{1}_A$ stands for the indicator function of the set A . The rescaling by the factor $n/(n+1)$, avoids difficulties arising from potential unboundedness of $\log [c_{\theta}(u_1, u_2)]$, when either u_1 or u_2 tends to 1. The estimate $\tilde{\theta}$, obtained by maximization of $L(\theta)$, is known to be consistent when $\theta_T \in \mathring{\Theta}$ is an interior point of Θ , with $\sqrt{n}(\tilde{\theta} - \theta_T)$ being, then, asymptotically normal (see, e.g., Proposition 2.1 in [Genest et al. \(1995\)](#)). In the present Chapter, we will establish, under appropriate conditions of regularity, the asymptotic distribution of $\sqrt{n}(\tilde{\theta} - \theta_T)$, when $\theta_T \in \partial\Theta$ is a boundary value of Θ . The following notation will be needed, following the lines of [Genest et al. \(1995\)](#). Set

$$\theta \in \Theta \mapsto \ell(\theta, u_1, u_2) := \log [c_{\theta}(u_1, u_2)]. \quad (8.10)$$

The *modified empirical copula* is defined by

$$\bar{C}_n(u_1, u_2) := \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{F_{1n}(X_{1k}) \leq u_1\}} \mathbb{1}_{\{F_{2n}(X_{2k}) \leq u_2\}}, \quad (u_1, u_2) \in (0, 1)^2, \quad (8.11)$$

where F_{1n} and F_{2n} are the empirical distribution functions, defined as in (8.9). The *modified empirical copula* is slightly different from the *empirical copula* $\mathbb{C}_n(u_1, u_2)$, introduced by Deheuvels (1979a), and defined by

$$\mathbb{C}_n(u_1, u_2) = F_n\left(F_{1n}^{-1}(u_1), F_{2n}^{-1}(u_2)\right) \quad \text{for } (u_1, u_2) \in (0, 1)^2, \quad (8.12)$$

where $F_{1n}^{-1}(\cdot)$ and $F_{2n}^{-1}(\cdot)$ denote the empirical quantile functions, associated with $F_{1n}(x_1) = F_n(x_1, \infty)$ and $F_{2n}(x_2) = F_n(\infty, x_2)$, respectively, and defined by

$$F_{jn}^{-1}(t) := \inf\{x \in \mathbb{R} \mid F_{jn}(x) \geq t\}, \quad j = 1, 2. \quad (8.13)$$

Here, $F_n(\cdot, \cdot)$ denotes the joint empirical distribution function, associated with the sample $\{(X_{1k}, X_{2k}); k = 1, \dots, n\}$, defined by

$$F_n(x_1, x_2) = \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{X_{1k} \leq x_1\}} \mathbb{1}_{\{X_{2k} \leq x_2\}}, \quad -\infty < x_1, x_2 < \infty. \quad (8.14)$$

Let C_n be the *rescaled modified empirical copula* (not to be confused with $\mathbb{C}_n$ in (8.12)), defined by

$$C_n(u_1, u_2) := \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{\widehat{F}_{1n}(X_{1k}) \leq u_1\}} \mathbb{1}_{\{\widehat{F}_{2n}(X_{2k}) \leq u_2\}}, \quad (u_1, u_2) \in (0, 1)^2, \quad (8.15)$$

where the $\widehat{F}_{jn}(\cdot)$, $j = 1, 2$, are as in (8.8). It is not too difficult to see that the log pseudo-likelihood (8.7) can be rewritten into the following integral form

$$L(\theta) = n \int_I \ell(\theta, u_1, u_2) dC_n(u_1, u_2), \quad (8.16)$$

where $I := (0, 1)^2$ and $\ell(\theta, u_1, u_2)$ is as in (8.10). We recall here that the random vectors (X_{1k}, X_{2k}) , $k = 1, 2, \dots$, are i.i.d. replicæ of (X_1, X_2) . Statistics of the form $\Psi_n := \int_I \psi(u_1, u_2) dC_n(u_1, u_2)$, belong to the general class of *multivariate rank statistics*. Their asymptotic properties have been investigated at length by a number of authors, among which we may cite Ruymgaart *et al.* (1972) and Rüschendorf (1976). In particular, the previous authors have provided regularity conditions, imposed on $\psi(\cdot, \cdot)$, which imply the asymptotic normality of Ψ_n . The corresponding arguments have been modified by Genest *et al.* (1995), as to establish almost sure convergence of the estimators they consider (see e.g. Genest *et al.* (1995) Proposition A.1). In the same spirit, the limiting behavior, as $n \rightarrow \infty$, of the estimators and test statistics which we will introduce later on, will make an instrumental use of the general theory of multivariate rank statistics, and rely, in particular, on Proposition A.1 in Genest *et al.* (1995). The existence and consistency of our estimators will be established through an application of the law of iterated logarithm for empirical copula processes, in combination with general arguments from multivariate rank statistics theory (we refer to Deheuvels (1979a), Fermanian *et al.* (2004) and the references therein). In the sequel, we shall limit ourselves to the following setup. In the case where θ_0 is a boundary value of Θ , we assume that the parameter space Θ is of the form $[\theta_0, \infty[$, with θ_0 , denoting the value of the parameter which corresponds to the hypothesis $\mathcal{H}_0$ of marginal independence. We recall that $\theta_T \in \Theta$ denotes the true value of the parameter of interest θ . These assumptions entail that $\theta_0 \in \partial\Theta$ is a boundary value of Θ . We will use a parametric family of copulas denoted by $\{C_\theta(u_1, u_2), \theta \in \Theta = [\theta_0, \infty[, (u_1, u_2) \in I = (0, 1)^2\}$. We assume, for each $\theta \in \Theta$, the existence of a continuous density $c_\theta(\cdot, \cdot)$ on I . It will become obvious later on that our arguments extend to more general parameter spaces. We will make use of the following definitions.

Definition 8.1. (i) Let $\mathcal{Q}$ be the set of continuous functions q on $[0, 1]$ which are positive on $(0, 1)$, symmetric about $1/2$, increasing on $[0, 1/2]$ and satisfy $\int_0^1 \{q(t)\}^{-2} dt < \infty$.

(ii) A function $r : (0, 1) \rightarrow (0, \infty)$ is called *u-shaped* if it is symmetric about $1/2$ and increasing on $(0, 1/2]$.

(iii) For $0 < \beta < 1$ and u -shaped function r , we define

$$r_\beta(t) := \begin{cases} r(\beta t) & \text{if } 0 < t \leq 1/2; \\ r\{1 - \beta(1 - t)\} & \text{if } 1/2 < t \leq 1. \end{cases}$$

If for $\beta > 0$ in a neighborhood of 0, there exists a constant M_β , such that $r_\beta \leq M_\beta r$ on $(0, 1)$, then r is called a reproducing u -shaped function. We denote by $\mathcal{R}$ the set of reproducing u -shaped functions.

Typical examples of elements, in $\mathcal{Q}$ and $\mathcal{R}$, are given by $q(t) = [t(1 - t)]^\zeta$, $0 < \zeta < 1/2$, and $r(t) = \varrho[t(1 - t)]^{-\varsigma}$, $\varsigma \geq 0$, $\varrho \geq 0$. We give our results under the following assumptions.

- (C.1) There exist functions $r_{1,k}, r_{2,k} \in \mathcal{R}$ such that $\left| \frac{\partial^k}{\partial \theta^k} \ell(\theta_T, u_1, u_2) \right| \leq r_{1,k}(u_1)r_{2,k}(u_2)$, for $k = 1, 2$, and $\left| \frac{\partial^3}{\partial \theta^3} \ell(\theta, u_1, u_2) \right| \leq r_{1,3}(u_1)r_{2,3}(u_2)$ on a neighborhood $N(\theta_T)$ of θ_T , where $\int_I \{r_{1,k}(u_1)r_{2,k}(u_2)\}^2 dC_{\theta_T}(u_1, u_2) < \infty$ for $k = 1, 2, 3$;
- (C.2) The function $(u_1, u_2) \in I \mapsto \frac{\partial}{\partial \theta} \ell(\theta_T, u_1, u_2)$ is of bounded variation on I ;
- (C.3) For each θ , the function $\frac{\partial}{\partial \theta} \ell(\theta, u_1, u_2) : I \rightarrow \mathbb{R}$ is continuously differentiable, and there exist functions $r_i \in \mathcal{R}$, $\tilde{r}_i \in \mathcal{R}$ and $q_i \in \mathcal{Q}$ for $i = 1, 2$, such that $\left| \frac{\partial^2}{\partial \theta \partial u_i} \ell(\theta, u_1, u_2) \right| \leq \tilde{r}_i(u_i)r_j(u_j)$, $i, j = 1, 2$ and $i \neq j$, and $\int_I \{q_i(u_i)\tilde{r}_i(u_i)r_j(u_j)\} dC_{\theta_T}(u_1, u_2) < \infty$, $i, j = 1, 2$ and $i \neq j$;
- (C.4) There exists $\eta > 0$, such that the function $\theta \in \Theta \mapsto \ell(\theta, u_1, u_2)$ is concave on $[\theta_0, \theta_0 + \eta]$, almost everywhere, with respect to $(u_1, u_2) \in I$.

The following additional notation will be needed. Letting, for $i = 1, 2$, $W_i(\theta_T, X_i) := \int_I \mathbb{1}_{\{F_i(X_i) \leq u_i\}} \frac{\partial^2}{\partial \theta \partial u_i} \ell(\theta_T, u_1, u_2) c_{\theta_T}(u_1, u_2) du_1 du_2$, set

$$\beta := -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \ell(\theta_T, F_1(X_1), F_2(X_2)) \right], \quad (8.17)$$

and

$$\sigma^2 := \text{Var} \left[\frac{\partial}{\partial \theta} \ell(\theta_T, F_1(X_1), F_2(X_2)) + W_1(\theta_T, X_1) + W_2(\theta_T, X_2) \right]. \quad (8.18)$$

It is readily checked that β and $W_i(X_i)$ can be rewritten into $\beta = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \ell(\theta_T, F_1(X_1), F_2(X_2)) \right)^2 \right]$, and, for $i = 1, 2$, $W_i(\theta_T, X_i) = -\int_I \mathbb{1}_{\{F_i(X_i) \leq u_i\}} \frac{\partial}{\partial \theta} \ell(\theta_T, u_1, u_2) \frac{\partial}{\partial u_i} \ell(\theta_T, u_1, u_2) c_{\theta_T}(u_1, u_2) du_1 du_2$.

Theorem 8.1. *Let the conditions (C.1-3) be fulfilled and $\beta > 0$. The following properties hold.*

(1) *Whenever θ_T is an interior point of Θ , the following properties are satisfied.*

- (a) *Let $B(\theta_T, n^{-1/3}) := \{\theta \in \Theta, |\theta - \theta_T| \leq n^{-1/3}\}$. Then, with probability 1, for all n sufficiently large, the function $\theta \mapsto \int_I \ell(\theta, u_1, u_2) dC_n(u_1, u_2)$ reaches its supremum for some $\tilde{\theta}_n \in B(\theta_T, n^{-1/3})$, so that $|\tilde{\theta}_n - \theta_T| = O(n^{-1/3})$ as $n \rightarrow \infty$. Moreover, $\int_I \frac{\partial}{\partial \theta} \ell(\tilde{\theta}_n, u_1, u_2) dC_n(u_1, u_2) = 0$.*
- (b) *We have the convergence in distribution, as $n \rightarrow \infty$,*

$$\sqrt{n}(\tilde{\theta}_n - \theta_T) \xrightarrow{d} N(0, \sigma^2/\beta^2). \quad (8.19)$$

(2) *Whenever $\theta_T = \theta_0$, is on the boundary of $\Theta = [\theta_0, \infty)$, the following properties are satisfied.*

- (a) *Let $B_+(\theta_0, n^{-1/3}) := [\theta_0, \theta_0 + n^{-1/3}]$. Then, with probability 1, for all n sufficiently large, the function $\theta \mapsto \int_I \ell(\theta, u_1, u_2) dC_n(u_1, u_2)$ reaches its supremum for some $\tilde{\theta}_n \in B_+(\theta_0, n^{-1/3})$, so that $|\tilde{\theta}_n - \theta_0| = O(n^{-1/3})$ as $n \rightarrow \infty$.*

(b) Let conditions (C.1-4) be fulfilled. Then, as $n \rightarrow \infty$, we have the convergence in distribution

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} X_+ := X\mathbf{1}_{(X>0)}, \quad (8.20)$$

where $X \stackrel{d}{=} N(0, \sigma^2/\beta^2)$ denotes a centered normal random variable with variance σ^2/β^2 .

Remark 8.1. A result in [Genest et al. \(1995\)](#) says that the MPL estimator, in general, is not as efficient as the ML one. However, under condition (C.5) below, it is efficient (its limit variance is the same as the ML one); see Proposition 2.2 in [Genest et al. \(1995\)](#), see also [Genest and Werker \(2002\)](#).

8.2.2 Tests of Independence

In this Section, we consider tests of marginal independence, under the previously considered parametric copula models. Throughout, we shall assume that marginal independence, namely the assumption, denoted hereafter by $\mathcal{H}_0$, that $C_{\theta_T}(u_1, u_2) = u_1 u_2$ for all $0 < u_1, u_2 < 1$, is equivalent to the condition that $\theta_T = \theta_0$, i.e., $C_{\theta_0}(u_1, u_2) = u_1 u_2$ for all $0 < u_1, u_2 < 1$. The alternative hypothesis, $\mathcal{H}_1 : \theta_T \neq \theta_0$, is, naturally, composite. The corresponding GPLR statistic is given by

$$\mathbf{S}_n = \mathbf{S}_n(\theta_0, \tilde{\theta}) := 2 \log \frac{\sup_{\theta \in \Theta} \prod_{k=1}^n c_{\theta}(\hat{F}_{1n}(X_{1k}), \hat{F}_{2n}(X_{2k}))}{\prod_{k=1}^n c_{\theta_0}(\hat{F}_{1n}(X_{1k}), \hat{F}_{2n}(X_{2k}))} = 2 \sup_{\theta \in \Theta} \sum_{k=1}^n \log c_{\theta}(\hat{F}_{1n}(X_{1k}), \hat{F}_{2n}(X_{2k})), \quad (8.21)$$

since $c_{\theta_0}(u_1, u_2) = 1$, for all $0 < u_1, u_2 < 1$. It turns out that, under the assumptions (C.1-2-3-5), and when θ_0 is an interior point of Θ (which is the case, in particular, for the FGM copula (9.8)), the limit law of $\mathbf{S}_n$ as $n \rightarrow \infty$, under the null hypothesis $\mathcal{H}_0$ of independence, is a χ_1^2 law with one degree of freedom. In the following, we will work under the additional condition (C.4), and provide the limit law, as $n \rightarrow \infty$, of the statistic $\mathbf{S}_n$, under the null hypothesis $\mathcal{H}_0$, and when θ_0 is a boundary point of Θ . We will also describe its limiting behavior, under the alternative hypothesis $\mathcal{H}_1$. In agreement with the assumptions and notations of the previous Section, we assume that the parameter space Θ is of the form $[\theta_0, \infty[$. We will make use of the following additional assumptions.

(C.5) We have $\lim_{\theta \rightarrow \theta_0+} \frac{\partial}{\partial u_i} \ell(\theta, u_1, u_2) = 0$, and there exist a neighborhood $N(\theta_0)$ of θ_0 , $M_1 > 0$ and $\delta_1 > 0$ such that, for all $\theta \in N(\theta_0) \cap \Theta$, one has, for $i = 1, 2$,
 $\left| \frac{\partial}{\partial \theta} \ell(\theta, u_1, u_2) \frac{\partial}{\partial u_i} \ell(\theta, u_1, u_2) c_{\theta}(u_1, u_2) \right| < M_1 r(u_i)^{-1.5+\delta_1} r(u_{3-i})^{0.5+\delta_1}$, where $r(u) := u(1-u)$ for $u \in (0, 1)$;

(C.6) There exist a neighborhood $N(\theta_T)$ of θ_T , and functions $r_i \in \mathcal{R}$ such that for all $\theta \in N(\theta_T) \cap \Theta$, we have $\left| \frac{\partial}{\partial \theta} \ell(\theta, u_1, u_2) \right| \leq r_1(u_1) r_2(u_2)$ with $\int_I \{r_1(u_1) r_2(u_2)\}^2 dC_{\theta_T}(u_1, u_2) < \infty$.

(C.7) There exist functions $r_i, \tilde{r}_i \in \mathcal{R}$, $q_i \in \mathcal{Q}$, $i = 1, 2$, such that $\left| \ell(\theta_T, u_1, u_2) \right| \leq r_1(u_1) r_2(u_2)$,
 $\left| \frac{\partial}{\partial u_i} \ell(\theta, u_1, u_2) \right| \leq \tilde{r}_i(u_i) r_j(u_j)$, $i, j = 1, 2$ and $i \neq j$,
 with $\int_I \{r_1(u_1) r_2(u_2)\}^2 dC_{\theta_T}(u_1, u_2) < \infty$ and $\int_I \{q_i(u_i) \tilde{r}_i(u_i) r_j(u_j)\} dC_{\theta_T}(u_1, u_2) < \infty$, $i, j = 1, 2$ and $i \neq j$.

Theorem 8.2. Assume that the conditions (C.1-2-3-4-5) hold. Then, under the null hypothesis $\mathcal{H}_0$, the statistic $\mathbf{S}_n$ converges in distribution, as $n \rightarrow \infty$, to the random variable $Z^2 \mathbf{1}_{\{Z>0\}}$, where $Z \stackrel{d}{=} N(0, 1)$ is a standard normal random variable.

We now describe the limiting behavior of the statistic $\mathbf{S}_n$ in (8.21), under the alternative hypothesis $\mathcal{H}_1 : \theta_T \neq \theta_0$. We will make use of the, so called, modified Kullback-Leibler divergence between probability measures. This quantity is defined by setting $KL_m(Q, P) := \int_{\mathcal{X}} \varphi \left(\frac{dQ}{dP} \right) dP$, where here, Q and P denote two probability measures, on measurable space $(\mathcal{X}, \mathcal{B})$, with $Q \ll P$, absolutely continuous with respect to P , and where we set $\varphi(x) := -\log x + x - 1$ for all $x > 0$. Note that $KL_m(Q, P) \geq 0$ is nonnegative,

and becomes equal to 0, if and only if $Q = P$. In the following, we will use the notation $KL_m(\theta_0, \theta_T)$, to denote the KL_m -divergence between the two distributions induced on I by C_{θ_0} and C_{θ_T} . Recalling the definition (8.21) of $\mathbf{S}_n$, we observe that $\frac{1}{2n}\mathbf{S}_n$ can be rewritten into

$$\frac{\mathbf{S}_n}{2n} = \int_I -\log \frac{c_{\theta_0}(u_1, u_2)}{c_{\hat{\theta}_n}(u_1, u_2)} dC_n(u_1, u_2) = \sup_{\theta \in \Theta} \int_I -\log \frac{c_{\theta_0}(u_1, u_2)}{c_{\theta}(u_1, u_2)} dC_n(u_1, u_2), \quad (8.22)$$

which is nothing else but an empirical version of the non-random quantity

$$\sup_{\theta \in \Theta} \int_I -\log \frac{c_{\theta_0}(u_1, u_2)}{c_{\theta}(u_1, u_2)} dC_{\theta_T}(u_1, u_2). \quad (8.23)$$

It is readily checked that the supremum in (8.23) is reached for the unique choice of $\theta = \theta_T$; see e.g. [Broniatowski and Keziou \(2006\)](#) Theorem 4.4 for a general proof in the context of φ -divergences. Therefore, we can write

$$\sup_{\theta \in \Theta} \int_I -\log \frac{c_{\theta_0}(u_1, u_2)}{c_{\theta}(u_1, u_2)} dC_{\theta_T}(u_1, u_2) = \int_I -\log \frac{c_{\theta_0}(u_1, u_2)}{c_{\theta_T}(u_1, u_2)} dC_{\theta_T}(u_1, u_2), \quad (8.24)$$

which is equal to $KL_m(\theta_0, \theta_T)$. By all this, we see that $\frac{\mathbf{S}_n}{2n}$ can be interpreted as an empirical version of $KL_m(\theta_0, \theta_T)$. This observation allows us to give a description of the limiting behavior of $\mathbf{S}_n$, under the alternative hypothesis $\mathcal{H}_1$. Our result concerning this problem is stated in the theorem below.

Theorem 8.3. *Let the conditions (C.1-2-3) and (C.6-7) be fulfilled. Then, under the alternative hypothesis $\mathcal{H}_1 : \theta_T \neq \theta_0$, we have, as $n \rightarrow \infty$, $\sqrt{n} \left(\frac{\mathbf{S}_n}{2n} - KL_m(\theta_0, \theta_T) \right) \xrightarrow{d} N(0, \gamma^2)$, converges to a centered normal random variable with variance equal to*

$$\gamma^2 := \text{Var}[h(\theta_T, F_1(X_1), F_2(X_2)) + Y_1(\theta_T, X_1) + Y_2(\theta_T, X_2)], \quad (8.25)$$

where $Y_i(\theta_T, X_i) = -\int_I \mathbf{1}_{\{F_i(X_i) \leq u_i\}} \frac{\partial}{\partial \theta} h\{\theta_T, u_1, u_2\} \frac{\partial}{\partial u_i} h\{\theta_T, u_1, u_2\} c_{\theta_T}(u_1, u_2) du_1 du_2$.

Remark 8.2. *The above regularity conditions are satisfied by a large number of single parameter families of bivariate copulas, including the standard bivariate normal, the Farlie-Gumbel-Morgenstern system, and copulas of the archimedean variety such as those of Ali-Mikhail-Haq, Clayton and Frank; see e.g. [Genest et al. \(1995\)](#) and [Tsukahara \(2005\)](#).*

Remark 8.3. *The asymptotic variances (8.18) and (10.19) may be consistently estimated respectively by the sample variances of $\frac{\partial}{\partial \theta} \ell \left(\tilde{\theta}_n, \hat{F}_{1n}(X_{1,k}), \hat{F}_{2n}(X_{2,k}) \right) + W_1(\tilde{\theta}_n, X_{1,k}) + W_2(\tilde{\theta}_n, X_{2,k})$, $k = 1, \dots, n$, and $h \left(\tilde{\theta}_n, \hat{F}_{1n}(X_{1,k}), \hat{F}_{2n}(X_{2,k}) \right) + Y_1(\tilde{\theta}_n, X_{1,k}) + Y_2(\tilde{\theta}_n, X_{2,k})$, $k = 1, \dots, n$, as was done in [Genest et al. \(1995\)](#). Similarly, the parameter β in (8.17) may be consistently estimated by the sample mean of $\left[\frac{\partial}{\partial \theta} \ell \left(\tilde{\theta}_n, \hat{F}_{1n}(X_{1,k}), \hat{F}_{2n}(X_{2,k}) \right) \right]^2$, $k = 1, \dots, n$.*

Remark 8.4. *An application of theorem 8.2, leads to reject the null hypothesis $\mathcal{H}_0 : \theta_T = \theta_0$, whenever the value of the statistic $\mathbf{S}_n$ exceeds $q_{1-\alpha}$, namely, the $(1-\alpha)$ -quantile of the law of the random variable $Z^2 \mathbf{1}_{\{Z > 0\}}$. The corresponding test is then, asymptotically of level α , when $n \rightarrow \infty$. The critical region is, accordingly, given by $CR := \{\mathbf{S}_n > q_{1-\alpha}\}$. The fact that this test is consistent follows from Theorem 2.3. Further, this theorem can be used to give an approximation to the power function $\theta_T \in \Theta \mapsto \beta(\theta_T) := P_{\theta_T}\{CR\}$ in a similar way to [Keziou and Leoni-Aubin \(2005\)](#) and [Keziou and Leoni-Aubin \(2008\)](#). We so obtain that*

$$\beta(\theta_T) \approx 1 - \Phi \left(\frac{\sqrt{n}}{\gamma(\theta_0, \theta_T)} \left(\frac{q_{1-\alpha}}{2n} - KL_m(\theta_0, \theta_T) \right) \right), \quad (8.26)$$

where Φ denotes, as usual, the cumulative distribution function of an $N(0, 1)$ standard normal random variable. A useful consequence of (10.20) is the possibility of computing an approximate value of the

sample size ensuring a specified power $\beta(\theta_T)$, with respect to some pre-assigned alternative $\theta_T \neq \theta_0$. Let n_0 be the positive root of the equation

$$\beta = 1 - \Phi \left(\frac{\sqrt{n}}{\gamma(\theta_0, \theta_T)} \left(\frac{q_{1-\alpha}}{2n} - KL_m(\theta_0, \theta_T) \right) \right), \quad (8.27)$$

which can be rewritten into

$$n_0 = \frac{(a + b) - \sqrt{a(a + 2b)}}{2KL_m(\theta_0, \theta_T)^2}, \quad (8.28)$$

where $a := \gamma^2(\theta_0, \theta_T) (\Phi^{-1}(1 - \beta))^2$ and $b := q_{1-\alpha} KL_m(\theta_0, \theta_T)$. The sought-after approximate value of the sample size is then given

$$n^* := \lfloor n_0 \rfloor + 1,$$

where $\lfloor u \rfloor$ denote the integer part of u .

8.3 Concluding remarks and possible developments

We have addressed the estimation and test problems for parametric copulas models with unknown marginals. We have studied the MPL estimator and the GPLR statistic (for testing the null hypothesis of independence) both when the parameter is an interior or a boundary point of the parameter space, under the null and the alternative hypotheses. The existence of the MPL estimator is proved using the law of iterated logarithm of the empirical copula process. Simulation results show that the law approximation of the GPLR statistic by its limiting law is good even for small sample sizes. It is well known that the MPL estimator under the hypothesis of independence, if the parameter is an interior point, is as efficient as the ML one (their limit variance is the same); it will be interesting to compare the efficiency of the MPL and ML estimators when the parameter is a boundary value of the parameter space. In regular case, in parametric estimation it is proved that the GLR test is optimal in Bahadur sense, and in Pitman sense (for local alternatives), and in general it is preferable than the score-type tests. Hence, it seems natural to see if these properties remain valid for the GPLR test in semiparametric copula models both when the parameter is an interior point or not.

Chapter 9

A test procedure of independence in copula models via χ^2 -divergence

In this Chapter, we present our main results of the paper [Bouzebda and Keziou \(2010b\)](#). We introduce a new test procedure of independence in the framework of parametric copulas with unknown marginals. The method is based essentially on the dual representation of χ^2 -divergence on signed finite measures. The asymptotic properties of the proposed estimate and the test statistic are studied under the null and alternative hypotheses, with simple and standard limit distributions both when the parameter is an interior point or not. The procedure has been generalized to other φ -divergence criteria and it has been showed that the the choice of chi-square divergence presents good properties in terms of efficiency-robustness; see [Bouzebda and Keziou \(2010a\)](#).

9.1 Introduction and motivations

Parametric models for copulas have been intensively investigated during the last decades. *Copulas* have become of great interest in applied statistics, because of the fact that they constitute a flexible and robust way to model dependence between the marginals of random vectors. The reader may refer to the following books for excellent expositions of the basics of copula theory : [Nelsen \(2006\)](#) and [Joe \(1997\)](#). In this framework, semiparametric inference methods, based on *pseudo-likelihood*, have been applied to copulas by a number of authors (see e.g. [Shih and Louis \(1995\)](#), [Wang and Ding \(2000\)](#), [Tsukahara \(2005\)](#) and the references therein). Throughout the available literature, investigations on the asymptotic properties of parametric estimators, as well as the relevant test statistics, have privileged the case where the parameter is an interior point of the admissible domain. However, for most parametric copula models of interest, the boundaries of the admissible parameter spaces include some important parameter values, typically among which, that corresponding to the independence of margins. This Chapter concentrates on this specific problem. We aim, namely, to investigate parametric inference procedures, in the case where the parameter belongs to the boundary of the admissible domain. In particular, the usual limit laws both for parametric copula estimators and test statistics become invalid under these limiting cases, and, in particular, under marginal independence. Motivated by this observation, we will introduce a new semiparametric inference procedure based on χ^2 -divergence and *duality* technique. We will show that the proposed estimator remains asymptotically normal, even under the marginal independence assumption. This will allow us to introduce test statistic of independence, his study will be made, both under the null and alternative hypotheses. It is well known since the work of [Sklar \(1959\)](#) that the joint behavior of a bivariate vector (X_1, X_2) with d.f. $\mathbf{F}(x_1, x_2) := \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2)$, and continuous marginal d.f.'s $F_i(x_i) := P(X_i \leq x_i)$, $i = 1, 2$, is characterized by the copula (or dependence function) $C(\cdot, \cdot)$ associated with $\mathbf{F}(\cdot, \cdot)$. The copula function is defined, for all $(u_1, u_2) \in (0, 1)^2$, through the identity $C(u_1, u_2) := \mathbf{P}\{F_1(X_1) \leq u_1, F_2(X_2) \leq u_2\}$. Many useful multivariate models for dependence between X_1 and X_2 turn out to be generated by *parametric* families of copulas of the form $\{C_\theta; \theta \in \Theta\}$,

typically indexed by a vector valued parameter $\theta \in \Theta \subseteq \mathbb{R}^p$ (see e.g. [Kimeldorf and Sampson \(1975b\)](#), [Kimeldorf and Sampson \(1975a\)](#), and [Joe \(1993\)](#)). The nonparametric approach to copula estimation has been initiated by [Deheuvels \(1979a\)](#), who introduced and investigated the *empirical copula process*. In addition, [Deheuvels \(1980, 1981\)](#) described the limiting behavior of this empirical process (see, also [Fermanian et al. \(2004\)](#) and the references therein). In this Chapter, we consider semiparametric copula models with unknown marginals. In order to estimate the unknown *true* value of the parameter $\theta \in \Theta$, which we denote, throughout the sequel, by $\theta_T \in \Theta$, some *semiparametric estimation* procedures, based on the maximization, on the parameter space Θ , of properly chosen *pseudo-likelihood* criterion, have been proposed by [Oakes \(1994\)](#), and studied by [Genest et al. \(1995\)](#), [Shih and Louis \(1995\)](#), [Wang and Ding \(2000\)](#) and [Tsukahara \(2005\)](#) among others. In each of these papers, some asymptotic normality properties are established for $\sqrt{n}(\tilde{\theta} - \theta_T)$, where $\tilde{\theta} = \tilde{\theta}_n$ denotes a properly chosen estimator of θ_T . This is achieved, provided that θ_T lies in the *interior*, denoted by $\overset{\circ}{\Theta}$, of the parameter space $\Theta \subseteq \mathbb{R}^p$. On the other hand, the case where $\theta_T \in \partial\Theta := \overline{\Theta} - \overset{\circ}{\Theta}$ is a *boundary value* of Θ , has not been studied in a systematical way until present. Moreover, it turns out that, for the above-mentioned estimators, the asymptotic normality of $\sqrt{n}(\tilde{\theta} - \theta_T)$, may fail to hold for $\theta_T \in \partial\Theta$; indeed, under some regularity conditions, when θ is univariate, we can prove that the limit law is the distribution of $Z\mathbf{1}_{(Z \geq 0)}$ where Z is a centred normal variable, and that the limit law of the generalized pseudo-likelihood ratio statistic is a mixture of chi-square laws with one degree of freedom and Dirac measure at zero; see [Bouzebda and Keziou \(2008\)](#). Furthermore, when the parameter is multivariate, the derivation of the limit distributions under the null hypothesis of independence, becomes much more complex; see [Self and Liang \(1987\)](#). Also, the limit distributions are not standard which yields formidable numerical difficulties to calculate the critical value of the test. We cite below some examples of parametric copulas, for which marginal independence is verified for some specific values of the parameter θ , on the boundary $\partial\Theta$ of the admissible parameter set Θ . We start with examples for which θ varies within subsets of $\mathbb{R}$. Such is the case for the extreme value copulas, namely, $C_A(u_1, u_2) := \exp \left\{ \log u_1 u_2 A \left(\frac{\log u_1}{\log u_1 u_2} \right) \right\}$, where $A(\cdot) : [0, 1] \mapsto [1/2, 1]$ is a convex function such that $\max(t, 1 - t) \leq A(t) \leq 1$, for all $0 \leq t \leq 1$.

For

$$A(t) := A_\theta(t) = (t^\theta + (1 - t)^\theta)^{1/\theta}; \quad \theta \in [1, \infty[, \quad (9.1)$$

we have [Gumbel \(1960\)](#) family of copulas, which is one of the most popular model used to model bivariate extreme values. For

$$A_\theta(t) = 1 - (t^{-\theta} + (1 - t)^{-\theta})^{-1/\theta}; \quad \theta \in [0, \infty[\quad (9.2)$$

we obtain [Galambos \(1975\)](#) family of copulas. Finally for

$$A_\theta(t) = t\Phi \left(\theta^{-1} + \frac{1}{2}\theta \log \left(\frac{t}{1 - t} \right) \right) + (1 - t)\Phi \left(\theta^{-1} - \frac{1}{2}\theta \log \left(\frac{t}{1 - t} \right) \right), \quad (9.3)$$

where $\theta \in [0, \infty[$ and $\Phi(\cdot)$ denoting the standard normal $N(0, 1)$ distribution function, we obtain the [Hüsler and Reiss \(1989\)](#) family of copulas. A useful family of copulas, due to [Joe \(1993\)](#), is given, for $0 < u_1, u_2 < 1$, by

$$C_\theta(u_1, u_2) := 1 - \left[(1 - u_1)^\theta + (1 - u_2)^\theta - (1 - u_1)^\theta (1 - u_2)^\theta \right]^{1/\theta}; \quad \theta \in [1, \infty[. \quad (9.4)$$

The Gumbel-Barnett copulas are given, for $0 < u_1, u_2 < 1$, by

$$C_\theta(u_1, u_2) := u_1 u_2 \exp \{ -(1 - \theta)(\log u_1)(\log u_2) \}; \quad \theta \in [0, 1]. \quad (9.5)$$

The Clayton copulas of positive dependence are such that, for $0 < u_1, u_2 < 1$,

$$C_\theta(u_1, u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1 \right)^{-1/\theta}; \quad \theta \in]0, \infty[. \quad (9.6)$$

Parametric families of copulas with parameter θ varying in $\mathbb{R}^p$, for some $p \geq 2$, include the following classical examples. Below, we set $\theta = (\theta_1, \theta_2)^\top \in \mathbb{R}^2$.

$$C_\theta(u_1, u_2) := \left\{ 1 + \left[(u_1^{-\theta_1} - 1)^{\theta_2} + (u_2^{-\theta_1} - 1)^{\theta_2} \right]^{1/\theta_2} \right\}^{-1/\theta_1}, \quad \theta \in]0, \infty[\times]1, \infty[; \quad (9.7)$$

$$C_\theta(u_1, u_2) := \exp \left\{ - \left[\theta_2^{-1} \log \left(\exp \left(-\theta_2 (\log u_1)^{\theta_1} \right) + \exp \left(-\theta_2 (\log u_2)^{\theta_1} \right) - 1 \right) \right]^{1/\theta_1} \right\},$$

$\theta \in [1, \infty[\times]0, \infty[$. For other examples of the kind, we refer to Joe (1997).

For each of the above examples, the independence case $C_{\theta_T}(u_1, u_2) = u_1 u_2$ (or $A(t) = 1$) occurs at the boundary of the parameter space Θ , i.e., when $\theta_T = 1$ for the models (9.1), (9.4) and (9.5), $\theta_T = 0$ for the models (9.2), (9.3) and (9.6), $\theta_T = (0, 1)^\top$ for the bivariate parameter model (9.7), and $\theta_T = (1, 0)^\top$ for the bivariate parameter model (9.8). In the sequel, we will denote by θ_0 the value of the parameter (when it exists), corresponding to the independence of the marginals, i.e., the value of the parameter for which we have $C_{\theta_0}(u_1, u_2) := u_1 u_2$, for all $(u_1, u_2) \in (0, 1)^2$. Hence, $\theta_0 = 1$ for the models (9.1), (9.4) and (9.5), $\theta_0 = 0$ for the models (9.2), (9.3) and (9.6), $\theta_0 = (0, 1)^\top$ for the model (9.7), and $\theta_0 = (1, 0)^\top$ for the model (9.8). Note that for the models (9.2), (9.3), (9.6), (9.7) and (9.8), $C_{\theta_0}(u_1, u_2) = u_1 u_2$ is naturally defined to be the limit of $C_\theta(\cdot, \cdot)$ when θ tends to θ_0 with values in Θ . We denote $c_\theta(\cdot, \cdot) := \frac{\partial^2}{\partial u_1 \partial u_2} C_\theta(\cdot, \cdot)$ the density of $C_\theta(\cdot, \cdot)$ and we define $c_{\theta_0}(\cdot, \cdot)$ to be the limit of $c_\theta(\cdot, \cdot)$ when θ tends to θ_0 with values in Θ . Hence, we can show that for all the above models $c_{\theta_0}(u_1, u_2) = 1$ for all $0 < u_1, u_2 < 1$. In contrast to the preceding examples, where $\theta_0 \in \partial\Theta$ is a boundary value of Θ , the case where θ_0 is an interior point of Θ may, at times, occur, but is more seldom. An example where $\theta_0 \in \mathring{\Theta}$ is given by the Farlie-Gumbel-Morgenstern (FGM) copula, defined by

$$C_\theta(u_1, u_2) := u_1 u_2 + \theta u_1 u_2 (1 - u_1)(1 - u_2), \quad \theta \in \Theta := [-1, 1], \quad (9.8)$$

and for which $\theta_0 = 0 \in \mathring{\Theta} =]-1, 1[$. In the present Chapter, we will treat parametric estimation of θ_T , and tests of the independence assumption $\theta_T = \theta_0$. We consider both the case where $\theta_0 \in \mathring{\Theta}$ is an interior point of Θ , and the case where $\theta_0 \in \partial\Theta$ is a boundary value of Θ . To treat this case, we propose a new inference procedure, based on an estimation of χ^2 -divergence by *duality* technique. This method may be applied independently of the dimension of the parameter space. Also the limit law of the estimate of the parameter is normal and the limiting distribution of the proposed test statistic is χ^2 under independence, either when θ_0 is an interior point, or when θ_0 is a boundary point of Θ . The idea is to include the parameter domain Θ into an enlarged space, say Θ_e , in order to render θ_0 an interior point of the new parameter space, Θ_e . The conclusion is then obtained through an application of χ^2 -divergence and duality technique. Our methods rely on the fact that, under appropriate assumptions, the definition of the density $c_\theta(\cdot, \cdot) := \frac{\partial^2}{\partial u_1 \partial u_2} C_\theta(\cdot, \cdot)$ of $C_\theta(\cdot, \cdot)$, pertaining to the models we consider, may be extended beyond the *standard* domain of variation Θ of θ . On the other hand, the definition of $c_\theta(\cdot, \cdot)$ which corresponds to these extensions, is then, in general, no longer a density, and may, at times, become negative. For example, such is the case for the parametric models (9.1), (9.4), (9.2) and (9.3), for which $c_\theta(\cdot, \cdot)$ is meaningful for some $\theta \notin \Theta$, but then, becomes negative over some non-negligible (with respect to Lebesgue's measure) subsets of $(0, 1)^2$. This implies that the log-likelihood of the data is not properly defined on the whole space Θ_e . For this reason, we will use the χ^2 -divergence between signed finite measures. We will discuss this problem in more details, below, in Section 9.2. The remainder of the present Chapter is organized as follows. In Section 9.2, we present our semiparametric inference procedure, based upon optimization of the χ^2 -divergence between the model $(C_\theta, \theta \in \Theta_e)$ and the empirical copula associated to the data, and by using the *dual representation* of χ^2 -divergence. We then derive the asymptotic limiting distribution of the proposed estimator. It will become clear later on from our results, that the asymptotic normality of the estimate holds, even under the independence assumption, when, either, θ_0 is an interior, or a boundary point of Θ . The proposed test statistic of independence is also studied, under the null hypothesis $\mathcal{H}_0$ of independence, as well as under the alternative hypothesis. The limiting asymptotic distribution of the test statistic under the alternative hypothesis is used to derive an approximation to the power function. An application of the forthcoming results will allow us to evaluate the sample size necessary to guarantee

a pre-assigned power level, with respect to a specified alternative. Simulation results are presented in Bouzebda and Keziou (2010b) Section 4, illustrating the performance of the proposed test statistic.

9.2 A semiparametric estimation procedure through χ^2 -divergence

As mentioned earlier, the problem of estimating θ , when $\theta \in \partial\Theta$, has not been systematically considered in the literature; and the classical asymptotic normality property of the estimators is no longer satisfied. To overcome this difficulty, in what follows, we enlarge the parameter space Θ into a wider space $\Theta_e \supset \Theta$. This is tailored to let θ_0 become an interior point of Θ_e . Naturally, we assume that the definition of the function $c_\theta(\cdot, \cdot)$ may be extended to Θ_e . The difficulty associated with this construction is that, subject to a proper definition, the densities $c_\theta(\cdot, \cdot)$ of $C_\theta(\cdot, \cdot)$ with respect to the Lebesgue's measure, may become negative on some non negligible subsets of $I := (0, 1)^2$ (in this case $c_\theta(\cdot, \cdot)$ becomes the density of a signed measure, see Remark 9.4). Note that just as Deheuvels's empirical copula is not a copula, $c_\theta(\cdot, \cdot)$ for $\theta \in \Theta_e$ is not necessarily a copula density and fail to integrate to 1. When such is the case, a semiparametric estimation of θ_T via log-likelihood cannot be used. To circumvent this difficulty, we introduce a new inference procedure, based on χ^2 -divergence method, and duality technique. Recall that the χ^2 -divergence between a finite signed measure $\mathbf{Q}$, and a probability $\mathbf{P}$, both defined on a same σ -field $(\mathcal{X}, \mathcal{B})$, when $\mathbf{Q}$ is absolutely continuous with respect to $\mathbf{P}$, is defined by

$$\chi^2(Q, P) := \int_{\mathcal{X}} \varphi \left(\frac{d\mathbf{Q}}{d\mathbf{P}}(x) \right) d\mathbf{P}(x), \text{ where } \varphi : x \in \mathbb{R} \mapsto \varphi(x) := \frac{1}{2} (x - 1)^2. \quad (9.9)$$

In the sequel, we denote by $\chi^2(\theta_0, \theta_T)$ the χ^2 -divergence between the copula distributions $C_{\theta_0}(\cdot, \cdot)$ and $C_{\theta_T}(\cdot, \cdot)$, i.e., $\chi^2(\theta_0, \theta_T) := \int_I \varphi \left(\frac{dC_{\theta_0}(u_1, u_2)}{dC_{\theta_T}(u_1, u_2)} \right) dC_{\theta_T}(u_1, u_2) = \int_I \varphi \left(\frac{1}{c_{\theta_T}(u_1, u_2)} \right) dC_{\theta_T}(u_1, u_2)$. Applying the dual representation of φ -divergence obtained by Broniatowski and Keziou (2006) Theorem 4.4, we readily obtain that $\chi^2(\theta_0, \theta_T)$ can be rewritten as

$$\chi^2(\theta_0, \theta_T) = \sup_{f \in \mathcal{F}} \left\{ \int_I f(u_1, u_2) dC_{\theta_0}(u_1, u_2) - \int_I \varphi^*(f(u_1, u_2)) dC_{\theta_T}(u_1, u_2) \right\}, \quad (9.10)$$

where $\varphi^*(\cdot)$ is used to denote the convex conjugate of $\varphi(\cdot)$, i.e., the function defined by $\varphi^* : t \in \mathbb{R} \mapsto \varphi^*(t) := \sup_{x \in \mathbb{R}} \{tx - \varphi(x)\} = \frac{1}{2}t^2 + t$, and $\mathcal{F}$ is an arbitrary class of measurable functions, fulfilling the following conditions : $\forall f \in \mathcal{F}$, $\int_I |f(u_1, u_2)| dC_{\theta_0}(u_1, u_2)$ is finite and $\varphi'(dC_{\theta_0}/dC_{\theta_T}) = \varphi'(1/c_{\theta_T}) \in \mathcal{F}$. Furthermore, the sup in (9.10) is unique and achieved at $f = \varphi'(1/c_{\theta_T})$. Define the new parameter space Θ_e of θ as follows. Set

$$\Theta_e := \left\{ \theta \in \mathbb{R}^d \text{ such that } \int |\varphi'(1/c_\theta(u_1, u_2))| du_1 du_2 < \infty \right\}. \quad (9.11)$$

Choosing the following class of functions $\mathcal{F} := \{(u_1, u_2) \in I \mapsto \varphi'(1/c_\theta(u_1, u_2)) - 1 ; \theta \in \Theta_e\}$, we infer from (9.10) the relation

$$\chi^2(\theta_0, \theta_T) = \sup_{\theta \in \Theta_e} \left\{ \int_I \left(\frac{1}{c_\theta(u_1, u_2)} - 1 \right) du_1 du_2 - \int_I \left(\frac{1}{2} \frac{1}{c_\theta(u_1, u_2)^2} - \frac{1}{2} \right) dC_{\theta_T}(u_1, u_2) \right\}. \quad (9.12)$$

It turns out that the supremum in (9.12) is reached iff $\theta = \theta_T$. Moreover, in general, θ_T is an interior point of the new parameter space Θ_e , especially under the null hypothesis of independence, namely, when $\theta_T = \theta_0$. Set

$$m(\theta, u_1, u_2) := \int_I \left(\frac{1}{c_\theta(u_1, u_2)} - 1 \right) du_1 du_2 - \left\{ \frac{1}{2} \frac{1}{c_\theta(u_1, u_2)^2} - \frac{1}{2} \right\}. \quad (9.13)$$

Consider a random sample $\{(X_{1,k}, X_{2,k}); k = 1, \dots, n\}$ from the distribution of (X_1, X_2) denoted by $\mathbf{F}_{\theta_T}(x_1, x_2) := C_{\theta_T}(F_1(x_1), F_2(x_2)) = \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2)$. In what follows, we propose to estimate the χ^2 -divergence $\chi^2(\theta_0, \theta_T)$, between $C_{\theta_0}(\cdot, \cdot)$ and $C_{\theta_T}(\cdot, \cdot)$, by

$$\widehat{\chi^2}(\theta_0, \theta_T) := \sup_{\theta \in \Theta_e} \int_I m(\theta, u_1, u_2) dC_n(u_1, u_2), \quad (9.14)$$

and to estimate the parameter θ_T by

$$\hat{\theta}_n := \arg \sup_{\theta \in \Theta_e} \left\{ \int_I m(\theta, u_1, u_2) dC_n(u_1, u_2) \right\}, \quad (9.15)$$

where $C_n(\cdot, \cdot)$ is the modified empirical copula, defined by

$$C_n(u_1, u_2) := \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{F_{1n}(X_{1k}) \leq u_1\}} \mathbb{1}_{\{F_{2n}(X_{2k}) \leq u_2\}}, \quad (u_1, u_2) \in I, \quad (9.16)$$

where

$$F_{jn}(t) := \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{]-\infty, t]}(X_{jk}), \quad j = 1, 2,$$

and $\mathbb{1}_A$ stands for the indicator function of the set A .

Remark 9.1. *The choice of the χ^2 -divergence among various divergences is motivated by the following statements:*

- Recall that the φ -divergences, between a finite signed measure $\mathbf{Q}$ and a probability $\mathbf{P}$ on $(\mathcal{X}, \mathcal{B})$, when $\mathbf{Q}$ is absolutely continuous with respect to $\mathbf{P}$, is defined by

$$D_\varphi(\mathbf{Q}, \mathbf{P}) := \int_{\mathcal{X}} \varphi \left(\frac{d\mathbf{Q}}{d\mathbf{P}}(x) \right) d\mathbf{P}(x),$$

where φ is a proper closed convex function from $]-\infty, \infty[$ to $[0, \infty[$ with $\varphi(1) = 0$ and such that the domain $\text{dom}_\varphi := \{x \in \mathbb{R} : \varphi(x) < \infty\}$ is an interval with end points $a_\varphi < 1 < b_\varphi$. The Kullback-Leibler, modified Kullback-Leibler, χ^2 , modified χ^2 , and Hellinger divergences are examples of φ -divergences; they are obtained, respectively, for $\varphi(x) = x \log x - x + 1$, $\varphi(x) = -\log x + x - 1$, $\varphi(x) = \frac{1}{2}(x - 1)^2$, $\varphi(x) = \frac{1}{2} \frac{(x-1)^2}{x}$, and $\varphi(x) = 2(\sqrt{x} - 1)^2$. We observe that Kullback-Leibler, modified Kullback-Leibler, modified χ^2 , and Hellinger divergences are infinite when $d\mathbf{Q}/d\mathbf{P}$ takes negative values on non negligible (with respect to $\mathbf{P}$) subset of $\mathcal{X}$, since the corresponding $\varphi(\cdot)$ is infinite on $(-\infty, 0)$. This problem does not hold in the case of χ^2 -divergence, indeed the corresponding $\varphi(\cdot)$ is finite on $\mathbb{R}$.

- We give an example of copulas for which the likelihood-based procedure fails. We consider the Gumbel copulas $C_\theta(\cdot, \cdot)$ given in (9.1), its corresponding density copula is defined by

$$c_\theta(u_1, u_2) := C_\theta(u_1, u_2)(u_1 u_2)^{-1} \frac{(\widetilde{u_1} \widetilde{u_2})^{(\theta-1)}}{(\widetilde{u_1}^\theta + \widetilde{u_2}^\theta)^{(2-1/\theta)}} \left[(\widetilde{u_1}^\theta + \widetilde{u_2}^\theta)^{(1/\theta)} + \theta - 1 \right], \quad (9.17)$$

where $\widetilde{x} = -\log x$. We can show that $c_\theta(\cdot, \cdot)$ may take negative values for some $\theta \in \Theta_e$. In fact $c_{0.7}(u_1, u_2)$ is negative for $(u_1, u_2) \in [0.9, 1]^2$, hence the likelihood function is not well defined.

Remark 9.2. *The set Θ_e defined in (9.11) is generally with non empty interior $\overset{\circ}{\Theta}_e$. In particular, we may check that θ_0 (the value corresponding to independence) belongs to $\overset{\circ}{\Theta}_e$, since the integral in (9.11) is finite; it is equal to zero when $\theta = \theta_0$, for any copula density $c_\theta(\cdot, \cdot)$. For example, in the case of FGM copulas (9.8), it is easy to show that $\Theta_e = \mathbb{R}$. However, it is hard to determine the whole set Θ_e for some copulas, but in order to test the independence, we need only to prove the existence of a neighborhood $N(\theta_0)$ of θ_0 for which the integral in (9.11) is finite since we calculate the estimate $\hat{\theta}_n$ in (9.15) by Newton-Raphson algorithm using θ_0 as initial point. The explicit calculation of the integral in (9.11) may be complicated for some copulas, in such cases we use the Monte Carlo method to compute this integral.*

Statistics of the form $\Psi_n := \int_I \psi(u_1, u_2) dC_n(u_1, u_2)$, belong to the general class of *multivariate rank statistics*. Their asymptotic properties have been investigated at length by a number of authors, among whom we may cite [Ruymgaart et al. \(1972\)](#), [Ruymgaart \(1974\)](#) and [Rüschendorf \(1976\)](#). In particular, the previous authors have provided regularity conditions, imposed on $\psi(\cdot, \cdot)$, which imply the asymptotic normality of Ψ_n . The corresponding arguments have been modified by [Genest et al. \(1995\)](#), [Tsukahara \(2005\)](#) and [Fermanian et al. \(2004\)](#), as to establish almost sure convergence of the estimators they consider. In the same spirit, the limiting behavior, as n tends to the infinity, of our estimator and test statistic $\mathbf{T}_n$, will make an instrumental use of the general theory of multivariate rank statistics. Using some similar arguments as in [Qin and Lawless \(1994\)](#), the existence and consistency of our estimator and test statistic will be established through an application of the law of the iterated logarithm for empirical copula processes, in combination with general arguments from multivariate rank statistics theory. In the sequel, without loss of generality, we will limit ourselves to the case where the parameter is univariate. The extensions of our result in a multivariate framework may be achieved, at the price of additional technicalities, and under similar assumptions. We will make use of the following definitions. Recall the definition of the sets of functions $\mathcal{Q}$ and $\mathcal{R}$, given in Definition 8.1. We will describe the asymptotic properties of the proposed estimate $\hat{\theta}_n$ under the following conditions.

- (C.1) There exist functions $r_{1,k}, r_{2,k} \in \mathcal{R}$ such that $\left| \frac{\partial^k}{\partial \theta^k} m(\theta_T, u_1, u_2) \right| \leq r_{1,k}(u_1)r_{2,k}(u_2)$, for $k = 1, 2$, and $\left| \frac{\partial^3}{\partial \theta^3} m(\theta, u_1, u_2) \right| \leq r_{1,3}(u_1)r_{2,3}(u_2)$ on a neighborhood $N(\theta_T)$ of θ_T , where $\int_I \{r_{1,k}(u_1)r_{2,k}(u_2)\}^2 dC_{\theta_T}(u_1, u_2) < \infty$ for $k = 1, 2, 3$;
- (C.2) The function $(u_1, u_2) \in I \mapsto \frac{\partial}{\partial \theta} m(\theta_T, u_1, u_2)$ is of bounded variation on I ;
- (C.3) For each θ , the function $\frac{\partial}{\partial \theta} m(\theta, u_1, u_2) : I \rightarrow \mathbb{R}$ is continuously differentiable, and there exist functions $r_i \in \mathcal{R}$, $\tilde{r}_i \in \mathcal{R}$ and $q_i \in \mathcal{Q}$ for $i = 1, 2$, such that $\left| \frac{\partial^2}{\partial \theta \partial u_i} m(\theta, u_1, u_2) \right| \leq \tilde{r}_i(u_i)r_j(u_j)$, $i, j = 1, 2$ and $i \neq j$, and $\int_I \{q_i(u_i)\tilde{r}_i(u_i)r_j(u_j)\} dC_{\theta_T}(u_1, u_2) < \infty$, $i, j = 1, 2$ and $i \neq j$.

Set

$$\Xi := -E \left[\frac{\partial^2}{\partial \theta^2} m(\theta_T, F_1(X_{1k}), F_2(X_{2k})) \right], \quad (9.18)$$

and

$$\Sigma^2 := \text{Var} \left[\frac{\partial}{\partial \theta} m(\theta_T, F_1(X_1), F_2(X_2)) + W_1(\theta_T, X_1) + W_2(\theta_T, X_2) \right], \quad (9.19)$$

where $W_i(\theta_T, X_i) := \int_I \mathbb{1}_{\{F_i(X_i) \leq u_i\}} \frac{\partial^2}{\partial \theta \partial u_i} m(\theta_T, u_1, u_2) c_{\theta_T}(u_1, u_2) du_1 du_2$, $i = 1, 2$. We can see that Ξ and $W_i(X_i)$ can be defined, respectively, by $\Xi = E \left[\left(\frac{\partial}{\partial \theta} m(\theta_T, F_1(X_{1k}), F_2(X_{2k})) \right)^2 \right]$ and, for $i = 1, 2$, $W_i(\theta_T, X_i) = - \int_I \mathbb{1}_{\{F_i(X_i) \leq u_i\}} \frac{\partial}{\partial \theta} m(\theta_T, u_1, u_2) \frac{\partial}{\partial u_i} m(\theta_T, u_1, u_2) c_{\theta_T}(u_1, u_2) du_1 du_2$. The following theorem describes the asymptotic behavior of the estimate $\hat{\theta}_n$ given in (9.15). From now on, “ $\xrightarrow{d}$ ” denotes the convergence in distribution.

Theorem 9.1. *Assume that the conditions C.1-C.3 hold.*

- (a) Let $B(\theta_T, n^{-1/3}) := \{\theta \in \Theta_e, |\theta - \theta_T| \leq n^{-1/3}\}$. Then, as $n \rightarrow \infty$, with probability one, the function $\theta \mapsto \int m(\theta, u_1, u_2) dC_n(u_1, u_2)$ reaches its maximum value at some point $\hat{\theta}_n$ in the interior of the interval $B(\theta_T, n^{-1/3})$. As a consequence, the estimate is consistent almost surely and satisfies $\int_I \frac{\partial}{\partial \theta} m(\hat{\theta}_n, u_1, u_2) dC_n(u_1, u_2) = 0$;
- (b) As $n \rightarrow \infty$, $\sqrt{n}(\hat{\theta}_n - \theta_T) \xrightarrow{d} N(0, \Sigma^2/\Xi^2)$.

9.3 A test based on “ χ^2 -divergence”

One of the motivations of the present work is to build a statistical test of independence, based on χ^2 -divergence. In the framework of the parametric copula model, the null hypothesis, namely, the

independence case $C_\theta(u_1, u_2) = u_1 u_2$ corresponds to the condition that $\mathcal{H}_0 : \theta_T = \theta_0$. We consider the alternative composite hypothesis $\mathcal{H}_1 : \theta_T \neq \theta_0$. The corresponding generalized pseudo-likelihood ratio statistic is then given by

$$\mathbf{S}_n(\theta_0, \tilde{\theta}) = 2 \log \frac{\sup_{\theta \in \Theta} \prod_{k=1}^n c_\theta(\hat{F}_{1n}(X_{1k}), \hat{F}_{2n}(X_{2k}))}{\prod_{k=1}^n c_{\theta_0}(\hat{F}_{1n}(X_{1k}), \hat{F}_{2n}(X_{2k}))} = 2 \sup_{\theta \in \Theta} \sum_{i=1}^n \log c_\theta(\hat{F}_{1n}(X_{1k}), \hat{F}_{2n}(X_{2k})),$$

where, for $j = 1, 2$, $\hat{F}_{jn}$ stands for $n/(n+1)$ times the marginal empirical distribution function of the j -th variable X_j . The rescaling by the factor $n/(n+1)$, avoids difficulties arising from potential unboundedness of $\log c_\theta(u_1, u_2)$, when either u_1 or u_2 tends to one. Since, θ_0 is a boundary value of the parameter space Θ , we can see, that the convergence in distribution of $\mathbf{S}_n$ to a χ^2 random variable is not likely to hold. In order to bring a solution to this problem, we introduce the statistic

$$\mathbf{T}_n(\theta_0, \hat{\theta}_n) := 2n \sup_{\theta \in \Theta_e} \int_I m(\theta, u_1, u_2) dC_n(u_1, u_2). \quad (9.20)$$

Below, we will show that, under the null hypothesis $\mathcal{H}_0$, the just-given statistic $\mathbf{T}_n(\theta_0, \hat{\theta}_n)$ converges in distribution to a χ^2 -random variable. This property allows us to build a test of $\mathcal{H}_0$ against $\mathcal{H}_1$, asymptotically of level α . The limit law of $\mathbf{T}_n(\theta_0, \hat{\theta}_n)$ will also be given under the alternative hypothesis $\mathcal{H}_1$. The following additional conditions will be needed for the statement of our results.

(C.4) We have $\lim_{\theta \rightarrow \theta_0} \frac{\partial}{\partial u_i} m(\theta, u_1, u_2) = 0$, and there exist $M_1 > 0$ and $\delta_1 > 0$ such that, for all θ in some neighborhood of θ_0 , one has, for $i = 1, 2$,

$$\left| \frac{\partial}{\partial \theta} m(\theta, u_1, u_2) \frac{\partial}{\partial u_i} m(\theta, u_1, u_2) c_\theta(u_1, u_2) \right| < M_1 r(u_i)^{-1.5+\delta_1} r(u_{3-i})^{0.5+\delta_1},$$

where $r(u) := u(1-u)$ for $u \in (0, 1)$;

(C.5) There exist a neighborhood $N(\theta_T)$ of θ_T , and functions $r_i \in \mathcal{R}$ such that for all $\theta \in N(\theta_T)$, we have $\left| \frac{\partial}{\partial \theta} m(\theta, u_1, u_2) \right| \leq r_1(u_1) r_2(u_2)$ with $\int_I \{r_1(u_1) r_2(u_2)\}^2 dC_{\theta_T}(u_1, u_2) < \infty$;

(C.6) There exist functions $r_i, \tilde{r}_i \in \mathcal{R}$, $q_i \in \mathcal{Q}$, $i = 1, 2$ such that $|m(\theta_T, u_1, u_2)| \leq r_1(u_1) r_2(u_2)$, $\left| \frac{\partial}{\partial u_i} m(\theta, u_1, u_2) \right| \leq \tilde{r}_i(u_i) r_j(u_j)$, $i, j = 1, 2$ and $i \neq j$, with $\int_I \{r_1(u_1) r_2(u_2)\}^2 dC_{\theta_T}(u_1, u_2) < \infty$, and $\int_I \{q_i(u_i) \tilde{r}_i(u_i) r_j(u_j)\} dC_{\theta_T}(u_1, u_2) < \infty$, $i, j = 1, 2$ and $i \neq j$.

Theorem 9.2. Assume that conditions C.1-C.4 hold. Then, under the null hypothesis $\mathcal{H}_0$, the statistic $\mathbf{T}_n$ converges in distribution to a χ_1^2 -random variable (with 1 degree of freedom).

Theorem 9.3. Assume that conditions C.1-C.3 and C.5-C.6 hold. Then, under the alternative hypothesis $\mathcal{H}_1$, we have $\sqrt{n} \left(\frac{\mathbf{T}_n}{2n} - \chi^2(\theta_0, \theta_T) \right)$ converges to a centered normal random variable with variance

$$\sigma_{\chi^2}^2 := \text{Var} [m(\theta_T, F_1(X_{1k}), F_2(X_{2k})) + Y_1(\theta_T, X_1) + Y_2(\theta_T, X_2)], \quad (9.21)$$

where $Y_i(\theta_T, X_i) := \int_I \mathbf{1}_{\{F_i(X_i) \leq u_i\}} \frac{\partial}{\partial u_i} m(\theta_T, u_1, u_2) c_{\theta_T}(u_1, u_2) du_1 du_2$.

Remark 9.3. An application of Theorem 9.2, leads to reject the null hypothesis $\mathcal{H}_0 : \theta_T = \theta_0$, whenever the value of the statistic $\mathbf{T}_n$ exceeds $q_{1-\alpha}$, namely, the $(1-\alpha)$ -quantile of the χ_1^2 -distribution. The test corresponding to this rejection rule is then, asymptotically of level α , when $n \rightarrow \infty$. Accordingly, the critical region is given by $CR := \{\mathbf{T}_n > q_{1-\alpha}\}$. The fact that this test is consistent follows from Theorem 9.3. Further, this theorem can be used to give an approximation to the power function $\theta_T \mapsto \beta(\theta_T) := P_{\theta_T}\{CR\}$ in a similar way to Keziou and Leoni-Aubin (2005) and Keziou and Leoni-Aubin (2008). We so obtain that

$$\beta(\theta_T) \approx 1 - \Phi \left(\frac{\sqrt{n}}{\sigma_{\chi^2}} \left(\frac{q_{1-\alpha}}{2n} - \chi^2(\theta_0, \theta_T) \right) \right), \quad (9.22)$$

where Φ denotes, as usual, the distribution function of $N(0, 1)$ standard normal random variable. A useful consequence of (10.20) is the possibility of computing an approximate value of the sample size ensuring a specified power $\beta(\theta_T)$, with respect to some pre-assigned alternative $\theta_T \neq \theta_0$. Let n_0 be the positive root of the equation

$$\beta = 1 - \Phi \left(\frac{\sqrt{n}}{\sigma_{\chi^2}} \left(\frac{q_{1-\alpha}}{2n} - \chi^2(\theta_0, \theta_T) \right) \right), \quad (9.23)$$

which can be rewritten as

$$n_0 = \frac{(a + b) - \sqrt{a(a + 2b)}}{2\chi^2(\theta_0, \theta_T)}, \quad (9.24)$$

where $a := \sigma_{\chi^2}^2 (\Phi^{-1}(1 - \beta))^2$ and $b := q_{1-\alpha} \chi^2(\theta_0, \theta_T)$. The sought-after approximate value of the sample size is then given by

$$n^* := \lfloor n_0 \rfloor + 1,$$

where $\lfloor u \rfloor$ denote the integer part of u .

Remark 9.4. If the parameter θ does not belong to Θ , i.e., $\theta \in \Theta_e - \Theta$, the densities $c_\theta(\cdot, \cdot)$, with respect to the Lebesgue's measure, associated to $C_\theta(\cdot, \cdot)$, may become negative on some non negligible subsets of I . So, in this case

$$C_\theta(u_1, u_2) := \int_{(0, u_1) \times (0, u_2)} c_\theta(u_1, u_2) d\lambda(u_1, u_2); \quad \forall u \in I, \quad (9.25)$$

may be a signed measure (and not a copula) for some $\theta \in \Theta_e - \Theta$. Hence, the formula (9.15) may lead to a value of the estimator $\hat{\theta}_n$ belonging to $\Theta_e - \Theta$, and the corresponding $C_{\hat{\theta}_n}$ is not necessarily a copula. So, for point estimation, the pseudo-maximum likelihood estimator (restricted to vary in the admissible domain) should be used instead of (9.15). The latter estimator may not be meaningful, and is likely to have a larger mean square error. The main advantage of our formula (9.14), where the supremum is taken on the extended space Θ_e , instead of the admissible domain Θ , is that it permits easily to build a test of independence even when the dimension of the parameter space Θ is larger than one. We can show that the proposed test statistic, based on the formula (9.14), has a χ_d^2 limit law with $d = \dim(\Theta)$ degrees of freedom.

Remark 9.5. The asymptotic variances (9.19) and (9.21) may be consistently estimated respectively by the sample variances of $\frac{\partial}{\partial \theta} m(\hat{\theta}_n, F_{1n}(X_{1,k}), F_{2n}(X_{2,k})) + W_1(\hat{\theta}_n, X_{1,k}) + W_2(\hat{\theta}_n, X_{2,k})$, $k = 1, \dots, n$, and $m(\hat{\theta}_n, F_{1n}(X_{1,k}), F_{2n}(X_{2,k})) + Y_1(\hat{\theta}_n, X_{1,k}) + Y_2(\hat{\theta}_n, X_{2,k})$, $k = 1, \dots, n$, as was done in Genest et al. (1995). Similarly, the parameter Ξ in (9.18) may be consistently estimated by the sample mean of $\left[\frac{\partial}{\partial \theta} m(\hat{\theta}_n, F_{1n}(X_{1,k}), F_{2n}(X_{2,k})) \right]^2$, $k = 1, \dots, n$.

9.4 Conclusion and possible developments

We have proposed a new test procedure based on χ^2 -divergence and duality technique in the framework of parametric copulas. It seems that the procedure introduced here is particularly well adapted to the boundary problem. For point estimation, the estimator based on χ^2 -divergence when we extend the parameter space, may not have a meaningful interpretation and most probably has a larger mean square error. However, Theorem 9.1 may be useful to construct the confidence region for the parameter of interest in connection with the intersection method (see Feng and McCulloch (1992)) which is easy to implement, while the asymptotic distribution of the maximum likelihood estimator is difficult to use, since the limiting distributions are complex when incorporating boundary constraints.

For clarity, our arguments are presented in the bivariate context. Note, however, that a d -variate generalization is possible with obvious changes of notation and modification assumptions. We mention that the popular multivariate Archimedean copulas, $C_\theta(u_1, \dots, u_d) = \phi^{-1} \{ \phi(u_1) + \dots + \phi(u_d) \}$, can only admit positive dependence, so that as long as the independence copula belongs to a given Archimedean family, θ_0 is on the boundary. Note, also that the proposed test of independence is not *omnibus* but depends on the hypothesis that the dependence structure belongs to a certain family.

Chapter 10

A semiparametric likelihood ratio test to the change point in copula models

In the present Chapter, a semiparametric maximum-likelihood-type test is proposed and proved to have the same limiting null distribution as the classical parametric likelihood one. Under some mild conditions, the limiting law of the proposed test statistic, suitably normalized and centralized, is shown to be double exponential, under the null hypothesis of no change in the parameter of copula models. The asymptotic distribution of the proposed statistic under specified alternatives is shown to be normal, and an approximation to the power function is given. For more detail and proofs, we refer to [Bouzebda and Keziou \(2013\)](#), where simulation results are provided to illustrate the finite sample performance of the proposed test statistics.

10.1 Introduction and motivations

Change point models are used in many fields, for instance, archaeology, econometrics, epidemiology, medicine, reliability and financial applications. For a broader presentation of the field of change-point analysis along with statistical applications, we refer the reader to the monographs by [Brodsky and Darkhovsky \(1993\)](#), [Csörgő and Horváth \(1997\)](#), [Basseville and Nikiforov \(1993\)](#) and [Chen and Gupta \(2000\)](#). As mentioned in [Csörgő and Horváth \(1988\)](#), change point problems have originally arisen in the context of quality control; see [Page \(1954, 1955\)](#). Interesting applications of change point models in the analysis of DNA sequences may be found in [Fu and Curnow \(1990\)](#) and [Braun and Muller \(1998\)](#) for the segmentation of such sequences, in this setting, in general there is multiple changes in the sequences. We may refer to [He and Severini \(2010\)](#), [Döring \(2010\)](#), [Aue and Horváth \(2013\)](#), [Chan *et al.* \(2013\)](#) and [Horváth and Hušková \(2012\)](#), for more recent references. We start by giving some notation and definitions that are needed for the forthcoming sections. Consider a random vector $\mathbf{X} := (X_1, \dots, X_d)^\top \in \mathbb{R}^d$, $d \geq 1$, with joint cumulative distribution function [d.f.] $\mathbb{F}(\mathbf{x}) := \mathbb{F}(x_1, \dots, x_d) := \mathbb{P}(\mathbf{X} \leq \mathbf{x}) := \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d)$ with continuous marginal d.f.'s $F_j(x_j) := \mathbb{P}(X_j \leq x_j)$, for $j = 1, \dots, d$. Throughout the Chapter, we write bold letters for vectors, e.g., $\mathbf{x} = (x_1, \dots, x_d)^\top$ is a d -dimensional vector and $\mathbf{x}^\top$ denotes the transpose of $\mathbf{x}$. Inequalities $\mathbf{x} \leq \mathbf{y}$ are understood componentwise, i.e., $x_j \leq y_j$ for all $j = 1, \dots, d$. The characterization theorem of [Sklar \(1959\)](#) implies that there exists a unique d -variate function $\mathbb{C}(\cdot)$, such that,

$$\mathbb{F}(\mathbf{x}) = \mathbb{C}(F_1(x_1), \dots, F_d(x_d)), \quad \text{for all } \mathbf{x} \in \mathbb{R}^d. \quad (10.1)$$

The function $\mathbb{C}(\cdot)$ is called a copula and it is in itself a joint cumulative distribution function on $[0, 1]^d$ with uniform marginals. Conversely, for any marginal distributions $F_1(\cdot), \dots, F_d(\cdot)$, and any copula function $\mathbb{C}(\cdot)$, the function $\mathbb{C}(F_1(x_1), \dots, F_d(x_d))$ is a multivariate distribution function with given marginal distributions $F_1(\cdot), \dots, F_d(\cdot)$. In the monographs by [Nelsen \(2006\)](#) and [Joe \(1997\)](#) the reader may find detailed ingredients of the modelling theory as well as surveys of the commonly used copulas. For in

depth and overview historical notes, we refer to [Schweizer \(1991\)](#). We can refer also to [Sklar \(1973\)](#), where the author has sketched the proof of (10.1), developed some of its consequences, and surveyed some of the work on copulas. Copulas have proved to be a flexible and versatile tool in the analysis of dependency structures. More specifically, copula $\mathbb{C}(\cdot)$ “couples” the joint distribution function $\mathbb{F}(\cdot)$ to its univariate marginals, capturing as such the dependence structure between the components of $\mathbf{X} = (X_1, \dots, X_d)^\top$. Indeed, most conventional measures of dependence can be explicitly expressed in terms of the copula. This feature has led to successful applications in actuarial science and survival analysis (see e.g. [Frees and Valdez \(1998\)](#), [Cui and Sun \(2004\)](#)). In the literature on risk management and, more generally, in mathematical economics and mathematical finance modelling, a number of illustrations are provided (refer to books of [Cherubini et al. \(2004\)](#) and [McNeil et al. \(2005\)](#)), in particular, in the context of asset pricing and credit risk management. Many useful multivariate models for dependence between $X_1, \dots, X_d$ turn out to be generated by *parametric* families of copulas of the form $\{\mathbb{C}(\cdot, \boldsymbol{\theta}); \boldsymbol{\theta} := (\theta_1, \dots, \theta_p)^\top \in \boldsymbol{\Theta}\}$, typically indexed by a vector valued parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, for $p \geq 1$, (see e.g. [Kimeldorf and Sampson \(1975a,b\)](#) and [Joe \(1993\)](#) among others). In this framework, semiparametric inference methods, based on *pseudo-likelihood*, proposed by [Oakes \(1994\)](#), have been investigated by a number of authors (see e.g. [Genest et al. \(1995\)](#), [Shih and Louis \(1995\)](#), [Wang and Ding \(2000\)](#), [Chen and Fan \(2005\)](#), [Tsukahara \(2005\)](#), [Bouzebda and Keziou \(2008, 2010c\)](#) and the references therein), we may refer also to [Bouzebda and Keziou \(2009, 2010c,a\)](#) for semiparametric estimation based on φ -divergences, including as a particular case the semiparametric likelihood estimators. Although the idea of our testing approach follows that in [Gombay and Horváth \(1994\)](#), we allow for infinite-dimensional nuisance parameters (marginal distributions) in our proposed test. Hence, our tests are semiparametric, and the limiting distribution is distribution free. [Gombay and Horváth \(1994\)](#) have established asymptotic distribution of the likelihood ratio (LR) statistic for parametric models of copulas with known uniform marginals, their results are not directly applicable here since the two-step estimation of the semiparametric copula model depends on the estimates of unknown marginal distributions. These results are not only useful in their own right but essential to the construction of our semiparametric likelihood ratio (SLR) tests. Conventionally, the copula parameter is assumed to remain constant over time. However, empirical evidence suggests that the dependence structure is likely to vary due to some financial adjustments and critical social events. To cope with the change point, in the copula setting, for known marginals, [Dias and Embrechts \(2004\)](#) suggested a change point test for the copula parameter based on maximum likelihood estimation. [Guégan and Zhang \(2010\)](#) analyzed two types of copula parameter changes and proposed a test based on generalized likelihood ratio assuming that the copula family is known a priori. However, neither of these authors have examined in details the theoretical properties of the test. In the other hand, in [Dias and Embrechts \(2004\)](#), the problem has been considered in a parametric framework since the marginals have been assumed uniform. In the present work, in the derivation of our asymptotic results, we state explicitly the regularity conditions in the most general form and we give a rigorous proof of asymptotic properties of the test both under the null and the alternative hypotheses. Here, we deal with testing changes in dependence structures, provided there is no change-point in the marginal distributions. In the sequel, we assume that $\mathbb{C}(\cdot, \boldsymbol{\theta})$ admits a density $\mathbf{c}(\cdot, \boldsymbol{\theta})$ with respect to ν , where ν is a σ -finite measure on the borel σ -field $([0, 1]^d, \mathcal{B}([0, 1]^d))$. In what follows, we precise the problem that we will treat. Let $\mathbf{X}_1, \dots, \mathbf{X}_n$, be independent random vectors with values in $\mathbb{R}^d$ with copula functions $\mathbb{C}(\mathbf{u}, \boldsymbol{\theta}_1), \dots, \mathbb{C}(\mathbf{u}, \boldsymbol{\theta}_n)$, where $\mathbf{u} \in [0, 1]^d$ and $\boldsymbol{\theta}_i := (\theta_{1i}, \dots, \theta_{pi})^\top \in \boldsymbol{\Theta}$, $1 \leq i \leq n$. We want to test the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta}_1 = \dots = \boldsymbol{\theta}_n$, against the change-point alternative that there is k^* , $1 \leq k^* \leq n - 1$, such that $\mathcal{H}_1 : \boldsymbol{\theta}_1 = \dots = \boldsymbol{\theta}_{k^*} \neq \boldsymbol{\theta}_{k^*+1} = \dots = \boldsymbol{\theta}_n$. The copula functions $\mathbb{C}(\cdot, \boldsymbol{\theta}_1), \dots, \mathbb{C}(\cdot, \boldsymbol{\theta}_n)$ have probability densities $\mathbf{c}(\cdot, \boldsymbol{\theta}_1), \dots, \mathbf{c}(\cdot, \boldsymbol{\theta}_n)$ with respect to ν , where ν is a σ -finite measure on $([0, 1]^d, \mathcal{B}([0, 1]^d))$. If we know when the changes occur in the sample, then this is simply a two-sample problem. Let, for $i = 1, \dots, p$, $\ell_i(\mathbf{F}_n(\mathbf{x}), \boldsymbol{\theta}) := \frac{\partial}{\partial \theta_i} \log\{\mathbf{c}(\mathbf{F}_n(\mathbf{x}), \boldsymbol{\theta})\}$, where $\mathbf{F}_n(\mathbf{x}) := (F_{1n}(x_1), \dots, F_{dn}(x_d))^\top$, $\mathbf{F}(\mathbf{x}) := (F_1(x_1), \dots, F_d(x_d))^\top$, and for $j = 1, \dots, d$, $F_{jn}(\cdot)$ stands for $n/(n+1)$ times the marginal empirical distribution function of the j -th variable X_j . We have, namely, for all $x \in \mathbb{R}$,

$$F_{jn}(x) := \frac{1}{n+1} \sum_{k=1}^n \mathbb{1}_{]-\infty, x]}(X_{jk}), \quad j = 1, \dots, d, \quad (10.2)$$

where $\mathbb{1}_A$ stands for the indicator function of the set A . As usual, the rescaling by the factor $n/(n+1)$, avoids difficulties arising from potential unboundedness of $\log\{\mathbf{c}(u_1, \dots, u_d, \boldsymbol{\theta}_k)\}$, for $k = 1, \dots, n$, when either of u_i 's, for $i = 1, \dots, d$, tends to 1. We need to the following condition to be fulfilled.

(C.1) For each k , $1 \leq k \leq n$, we can find unique estimators $\hat{\boldsymbol{\theta}}_k$ and $\hat{\boldsymbol{\theta}}_{n-k}$, such that,

$$\sum_{j=1}^k \ell_i(\mathbf{F}_n(\mathbf{X}_j), \hat{\boldsymbol{\theta}}_k) = 0, \quad \text{for all } 1 \leq i \leq p, \quad (10.3)$$

$$\sum_{j=k+1}^n \ell_i(\mathbf{F}_n(\mathbf{X}_j), \hat{\boldsymbol{\theta}}_{n-k}) = 0, \quad \text{for all } 1 \leq i \leq p. \quad (10.4)$$

If the change occurs at $k^* = k$, then we should reject $\mathcal{H}_0$ for small value of $\boldsymbol{\Lambda}_k$, where

$$\boldsymbol{\Lambda}_k := \frac{\prod_{j=1}^n \mathbf{c}(\mathbf{F}_n(\mathbf{X}_j), \hat{\boldsymbol{\theta}}_n)}{\prod_{j=1}^k \mathbf{c}(\mathbf{F}_n(\mathbf{X}_j), \hat{\boldsymbol{\theta}}_k) \prod_{j=k+1}^n \mathbf{c}(\mathbf{F}_n(\mathbf{X}_j), \hat{\boldsymbol{\theta}}_{n-k})}. \quad (10.5)$$

Since k^* is unknown, it is natural to use the maximally selected pseudo-likelihood ratio and reject $\mathcal{H}_0$, if

$$\mathbf{Z}_n := \max_{1 \leq k \leq n-1} (-2 \log(\boldsymbol{\Lambda}_k)), \quad (10.6)$$

is large. Unless otherwise specified, we will denote the true value of the parameter $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$.

The rest of this Chapter is organized as follows. Section 10.2 establishes the asymptotic properties of the proposed statistical test $\mathbf{Z}_n$, both under the null and the alternative hypothesis.

10.2 Main results

The following notation is needed for the statement of our forthcoming results. Let $\ell(\mathbf{F}_n(\mathbf{x}), \boldsymbol{\theta}) := \log\{\mathbf{c}(\mathbf{F}_n(\mathbf{x}), \boldsymbol{\theta})\}$, and $\ell_{i_1 \dots i_r}(\mathbf{F}_n(\mathbf{x}), \boldsymbol{\theta}) := \frac{\partial^r}{\partial \theta_{i_1} \dots \partial \theta_{i_r}} \log\{\mathbf{c}(\mathbf{F}_n(\mathbf{x}), \boldsymbol{\theta})\}$. Let $\boldsymbol{\Sigma}$ be the variance-covariance matrix of p -dimensional vector whose i th component is given by $\ell_i(F_1(X_1), \dots, F_d(X_d), \boldsymbol{\theta}_0) + \sum_{j=1}^d W_{ij}(X_j)$, where $W_{ij}(X_j) := \int_{[0,1]^d} \mathbb{1}\{F_j(X_j) \leq u_j\} \frac{\partial}{\partial u_j} \ell_i(u_1, \dots, u_d, \boldsymbol{\theta}_0) d\mathbb{C}(u_1, \dots, u_d, \boldsymbol{\theta}_0)$. In the sequel $\|\cdot\|$ denotes as usual Euclidian norm. We assume the following regularity conditions.

(C.2) The copula $\mathbb{C}(\mathbf{u}, \boldsymbol{\theta}_0)$ has partial continuous derivatives with respect to $\mathbf{u}$;

(C.3) (i) For any $\mathbf{u} \in [0, 1]^d$, $\ell(\mathbf{u}, \boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$;

(ii) $\mathbb{E}_{\boldsymbol{\theta}_0} \{\sup_{\boldsymbol{\theta} \in \Theta} |\ell(\mathbf{F}(\mathbf{X}), \boldsymbol{\theta})|\} < \infty$;

(C.4) (i) $\mathbb{E}_{\boldsymbol{\theta}_0} \{\ell(\mathbf{F}(\mathbf{X}), \boldsymbol{\theta})\}$ has a unique maximum $\boldsymbol{\theta}_0$ in the interior of Θ , $\bar{\Theta}$ say, where Θ is a compact subset of $\mathbb{R}^p$;

(ii) The $p \times p$ -matrix $\mathbf{I}(\boldsymbol{\theta}_0)$, with entries $I_{ij} := -\mathbb{E}_{\boldsymbol{\theta}_0} \{\ell_{ij}(\mathbf{F}(\mathbf{X}), \boldsymbol{\theta}_0)\}$, is positive definite matrix and $\boldsymbol{\Sigma}$ is positive definite;

(C.5) For all $i = 1, \dots, p$, for all $j = 1, \dots, d$, $\frac{\partial}{\partial u_j} \ell_i(\mathbf{u}, \boldsymbol{\theta}_0)$ is well defined and continuous in $\mathbf{u} = (u_1, \dots, u_d)^\top \in (0, 1)^d$;

(C.6) (i) We have $\|\ell_j(\mathbf{u}, \boldsymbol{\theta}_0)\| \leq \text{constant} \times \prod_{i=1}^d \{u_i(1-u_i)\}^{-a_i}$ for some $a_i \geq 0$, such that

$$\mathbb{E}_{\boldsymbol{\theta}_0} \left\{ \prod_{i=1}^d \{F_i(X_i)(1-F_i(X_i))\}^{-2a_i} \right\} < \infty;$$

(ii) We have $\left\| \frac{\partial}{\partial u_j} \ell_i(\mathbf{u}, \boldsymbol{\theta}_0) \right\| \leq \text{constant} \times \{u_j(1-u_j)^{b_j}\} \prod_{i=1, i \neq j}^d \{u_i(1-u_i)\}^{-a_i}$ for some $b_i > a_i \geq 0$, such that, $\mathbb{E}_{\boldsymbol{\theta}_0} \left\{ \{F_j(X_j)(1-F_j(X_j))\}^{\xi_j-b_j} \prod_{i=1, i \neq j}^d \{F_i(X_i)(1-F_i(X_i))\}^{-2a_i} \right\} < \infty$, for some $\xi_j \in (0, 1/2)$.

(C.7) There is a function $M(\cdot)$, such that, for all $\mathbf{u} \in [0, 1]^d$ and $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$, $|\ell_{ijk}(\mathbf{u}, \boldsymbol{\theta})| \leq M(\mathbf{u})$ and $\mathbb{E}_{\boldsymbol{\theta}_0} \{M(\mathbf{F}(X))\} < \infty$.

Remark 10.1. *Csörgő and Horváth (1997) have mentioned that there are cases where we cannot find unique maximum likelihood estimators for each $1 \leq k \leq n$ if we have only few observations. Instead of (C.1), we can assume, according to Csörgő and Horváth (1997), the following condition*

(C'.1) *There is a constant k_0 such that we can find unique estimators $\hat{\boldsymbol{\theta}}_k$ and $\hat{\boldsymbol{\theta}}_{n-k}$ for all $k_0 \leq k \leq n - k_0$ such that (10.3) and (10.4) hold. Also, we can find unique $\hat{\boldsymbol{\theta}}_n$, such that, $\sum_{j=1}^n \ell_i(\mathbf{F}_n(\mathbf{X}_j), \hat{\boldsymbol{\theta}}_n) = 0$, for all $1 \leq i \leq p$.*

Recall the definitions of the sets of functions $\mathcal{Q}$ and $\mathcal{R}$, given in Definition 8.1.

Remark 10.2. *Condition (C.6) above is usual in the copula setting and is the same that used in (Ruyngaert, 1974, Theorem 2.1), see also (Ruyngaert et al., 1972, Theorem 2.1), (Genest et al., 1995, Proposition A.1(ii)), (Tsukahara, 2005, Proposition 3). Condition (C.6) allows the score function and its partial derivatives with respect to the first d arguments to blow up at the boundaries, which occur for many popular copula functions such as the Gaussian copula, t -copula and Clayton copula. Using the preceding definition, condition (C.6) may be written in the most general form as follows:*

(C'.6) *There exist $r_i(\cdot) \in \mathcal{R}$, $\tilde{r}_i(\cdot) \in \mathcal{R}$ and $q(\cdot) \in \mathcal{Q}$ with $\|\ell_j(\mathbf{u}, \boldsymbol{\theta}_0)\| \leq \prod_{i=1}^d r_i(u_i)$, $\left\| \frac{\partial}{\partial u_j} \ell_i(\mathbf{u}, \boldsymbol{\theta}_0) \right\| \leq \tilde{r}_i(u_i) \prod_{i \neq j}^d r_j(u_j)$; $\int_{[0,1]^d} \left\{ \prod_{i=1}^d r_i(u_i) \right\}^2 d\mathbb{C}(\mathbf{u}, \boldsymbol{\theta}_0) < \infty$, $\int \left\{ q_i(u_i) \tilde{r}_i(u_i) \prod_{i \neq j}^d r_j(u_j) \right\} d\mathbb{C}(\mathbf{u}, \boldsymbol{\theta}_0) < \infty$, for $i = 1, \dots, d$.*

Let

$$a(t) := (2 \log t)^{1/2} \quad \text{and} \quad b_p(t) := 2 \log t + \frac{1}{2} p \log \log t - \log \Gamma\left(\frac{p}{2}\right), \quad (10.7)$$

where $\Gamma(t) := \int_0^\infty y^{t-1} \exp(-y) dy$ is the Gamma function.

The main result may now be stated precisely as follows.

Theorem 10.1. *Assume that the conditions (C.1 – 7) hold. Then, under the null hypothesis $\mathcal{H}_0$, for all t , we have*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ a(\log n) \mathbf{Z}_n^{1/2} \leq t + b_p(\log n) \right\} = \exp(-2 \exp(-t)). \quad (10.8)$$

Remark 10.3. *One can replace condition (C.7) by the following one.*

(C'.7) (i) *For every $(u_1, \dots, u_d)$, $\ell_{ij}(u_1, \dots, u_d, \boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ in neighborhood of $\boldsymbol{\theta}_0$;*
(ii) *For some $\delta > 0$, we have $\mathbb{E}_{\boldsymbol{\theta}_0} \left\{ \left| \ell_i(F_1(X_1), \dots, F_d(X_d), \boldsymbol{\theta}_0) + \sum_{j=1}^d W_{ij}(X_j) \right|^{2+\delta} \right\} < \infty$, $i = 1, \dots, p$.*

Remark 10.4. *Note that the proof techniques of our Theorem 10.1 are largely inspired from those of Gombay and Horváth (1994) and changes have been made in order to adapt them to our semiparametric purpose.*

Remark 10.5. (Test 1). *Based on the above theorem, we propose the following test of level $\alpha \in [0, 1]$. Denote by $[a_\alpha, b_\alpha]$ the optimal dispersion interval of order $(1 - \alpha)$ of the double exponential distribution in (10.8), i.e., the shortest one that satisfies $\mathbb{P}(Z \in [a_\alpha, b_\alpha]) = 1 - \alpha$, where Z is a double exponential random variable. We reject the null hypothesis $\mathcal{H}_0$, if the statistic $S_n := a(\log n) \mathbf{Z}_n^{1/2} - b_p(\log n) \notin [a_\alpha, b_\alpha]$. The critical region is, accordingly, given by $RC_\alpha := \{S_n \notin [a_\alpha, b_\alpha]\}$. This leads, obviously, to a test of the null hypothesis $\mathcal{H}_0$, asymptotically of level α , by the above theorem.*

Remark 10.6. *As it has been observed by (Csörgő and Horváth, 1997, p. 23) in the context of parametric models of probability densities, simulation results, at section in our paper Bouzebda and Keziou (2013), show that the rate of convergence in (10.8) is very slow. In fact the convergence of the nominal level (computed by Monte-Carlo simulations) to the asymptotic level α is very slow. This leads to a conservative test. Hence, we need very large samples if we want to use the proposed test 1, in Remark 10.5. In the subsection below, we give a gaussian approximation to the likelihood ratio statistic $\mathbf{Z}_n$, which turns out to be better than the above double exponential one.*

10.2.1 Gaussian approximation

Following [Vostrikova \(1983\)](#), [Gombay \(1996\)](#) and [Gombay and Horváth \(1996b\)](#), we define the likelihood ratio process

$$\mathbf{V}_n(t) = -2 \log \mathbf{\Lambda}_{\lfloor (n+1)t \rfloor}, \quad \text{for } 1/(n+1) \leq t \leq n/(n+1), \quad (10.9)$$

where $\lfloor u \rfloor \leq u < \lfloor u \rfloor + 1$ denotes the integer part of u . Here it is assumed that $\mathbf{V}_n(t) = 0$, if $0 \leq t < 1/(n+1)$ and $\mathbf{V}_n(t) = 0$, if $n/(n+1) \leq t < 1$. Let $\{B_i(t), 0 \leq t \leq 1\}$, $1 \leq i \leq p$, be independent Brownian bridges and define $B^{(p)} := \sum_{1 \leq i \leq p} B_i^2(t)$.

Theorem 10.2. *Assume that the conditions (C.1)-(C'.7) hold. Then, under the null hypothesis $\mathcal{H}_0$, we can find a sequence of stochastic processes $\{B_n^{(p)}(t), 0 \leq t \leq 1\}$ such that the following distributional identity holds $\{B_n^{(p)}(t), 0 \leq t \leq 1\} \stackrel{d}{=} \{B^{(p)}(t), 0 \leq t \leq 1\}$, for each n ,*

$$n^\alpha \sup_{\lambda/n \leq t \leq 1-\lambda/n} (t(1-t))^\alpha \left| \mathbf{V}_n(t) - B_n^{(p)}(t)/(t(1-t)) \right| = O_{\mathbb{P}}(1) \quad (10.10)$$

for all $\lambda > 0$ and $0 \leq \alpha < 1/2 - 1/\delta$, and $\sup_{0 \leq t \leq 1} |t(1-t)\mathbf{V}_n - B_n^{(p)}(t)| = o_{\mathbb{P}}(n^{1/\delta-1/2})$.

Let $\psi(\cdot)$ be a positive function on $(0, 1)$ increasing in a neighborhood of zero and decreasing in a neighborhood of one. Define the function

$$I(\psi, \epsilon) := \int_0^1 (s(1-s))^{-1} \exp\left(-\frac{\epsilon \psi^2(s)}{s(1-s)}\right) ds, \quad (10.11)$$

where $\epsilon > 0$. The following results are the analogue of those in Theorems 1.3 and 1.4 of [Gombay and Horváth \(1996b\)](#).

Theorem 10.3. *Assume that the conditions (C.1)-(C'.7) and $\mathcal{H}_0$ hold.*

(i) *We can define a sequence of stochastic processes $\{B_n^{(p)}(t), 0 \leq t \leq 1\}$, such that $\sup_{0 \leq t \leq 1} |t(1-t)\mathbf{V}_n - B_n^{(p)}(t)|/\psi^2(t) = o_{\mathbb{P}}(1)$, if and only if $I(\psi, \epsilon) < \infty$ for all $\epsilon > 0$.*

(i) *$I(\psi, \epsilon) < \infty$ for some $\epsilon > 0$ if and only if the following convergence in distribution holds*

$$\sup_{0 \leq t \leq 1} |t(1-t)\mathbf{V}_n|/\psi^2(t) \xrightarrow{d} \sup_{0 \leq t \leq 1} |B_n^{(p)}(t)|/\psi^2(t). \quad (10.12)$$

It is clear that (10.10) yields $\left| \mathbf{Z}_n^{1/2} - \sup_{1/n \leq t \leq 1-1/n} (B_n^{(p)}(t)/(t(1-t)))^{1/2} \right| = O_{\mathbb{P}}(1)$. It will be of interest to give the rate of convergence in the last statement in the similar way as in Theorem 2.2 of [Gombay and Horváth \(1996b\)](#) and Theorem 1.3.2. in [Csörgő and Horváth \(1997\)](#). The following theorem gives the rate of convergence.

Theorem 10.4. *Assume that the conditions (C.1)-(C'.7) and $\mathcal{H}_0$ hold. Then*

$$\left| \mathbf{Z}_n^{1/2} - \sup_{1/n \leq t \leq 1-1/n} (B_n^{(p)}(t)/(t(1-t)))^{1/2} \right| = o_{\mathbb{P}}(\exp(-\log n)^{1-\epsilon}), \quad (10.13)$$

for all $0 < \epsilon < 1$, where $\{B_n^{(p)}(t), 0 \leq t \leq 1\}$ is defined in Theorem 10.2. Let $h(n) \geq 1/n$, $l(n) \geq 1/n$ satisfying $\limsup_{n \rightarrow \infty} n(h(n) + l(n)) \exp(-(\log n)^{1-\epsilon^*}) < \infty$, for some $0 < \epsilon^* \leq 1$. Then we have

$$\left| \mathbf{Z}_n^{1/2} - \sup_{h(n) \leq t \leq 1-l(n)} (B_n^{(p)}(t)/(t(1-t)))^{1/2} \right| = o_{\mathbb{P}}(\exp(-\log n)^{1-\epsilon}), \quad \text{for all } 0 < \epsilon < \epsilon^*. \quad (10.14)$$

The proof of this theorem follows the same lines of that of ([Gombay and Horváth, 1996b](#), Theorem 2.2 and Theorem 3.1, see pp. 149-151) or ([Csörgő and Horváth, 1997](#), Theorem 1.3.2, see the proof page 24).

Remark 10.7. (Test 2). *There is no known explicit formula for the distribution function of the random variable $Z(h, l) := \sup_{h(n) \leq t \leq 1-l(n)} (B_n^{(p)}(t)/(t(1-t)))^{1/2}$. In [(Csörgő and Horváth, 1997, p. 25) and (Vostrikova, 1981, eq. (6))], we can find the following approximation*

$$\mathbb{P}(Z(h, l) \geq x) \approx \frac{x^p \exp(-x^2/2)}{2^{p/2} \Gamma(p/2)} \left(\log \frac{(1-h)(1-l)}{hl} - \frac{p}{x^2} \log \frac{(1-h)(1-l)}{hl} + \frac{4}{x^2} + O\left(\frac{1}{x^4}\right) \right) \quad (10.15)$$

as $x \rightarrow \infty$, for any $0 < h < l < 1$. We use this approximation to compute the quantiles $q_\alpha(h, l)$ of order α of the random variable $Z(h, l)$, namely, the real numbers satisfying $P(Z(h, l) \geq q_\alpha(h, l)) = \alpha$. In view of the above theorem, we propose to reject the null hypothesis $\mathcal{H}_0$ when $\mathbf{Z}_n^{1/2} > q_\alpha(h, l)$, which leads to a test of $\mathcal{H}_0$ against $\mathcal{H}_1$ with asymptotic level equals to α . The critical region is, accordingly, given by

$$RC_\alpha := \left\{ \mathbf{Z}_n^{1/2} > q_\alpha(h, l) \right\}. \quad (10.16)$$

Simulation results show that the choice $h(n) = l(n) = (\log n)^{3/2}/n$ gives good rate of convergence of the nominal level (computed by Monte-Carlo simulation) to the asymptotic level α , as it has been observed, in a simulation study, by Dias and Embrechts (2004) for the Gumbel copula with known uniform marginals.

10.3 The asymptotic distribution of the SLR statistic under the alternative hypothesis

We now describe the limiting behavior of the statistic $\mathbf{Z}_n$ in (10.6), under the alternative hypothesis, that there exists k^* , $1 < k^* < n$, such that $\mathcal{H}_1 : \boldsymbol{\theta}_1 = \dots = \boldsymbol{\theta}_{k^*} \neq \boldsymbol{\theta}_{k^*+1} = \dots = \boldsymbol{\theta}_n$. Let $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}_1 = \dots = \boldsymbol{\theta}_{k^*}$ and $\boldsymbol{\theta}^{(2)} = \boldsymbol{\theta}_{k^*+1} = \dots = \boldsymbol{\theta}_n$ denote, respectively, the values of the parameter before and after the change. We will make use of the, so called, modified Kullback-Leibler divergence between probability measures. This quantity is defined by setting $KL_m(Q, P) := \int_{\mathcal{X}} \varphi \left(\frac{dQ}{dP}(x) \right) dP(x)$, where here, Q and P denote two probability measures on a same σ -field $(\mathcal{X}, \mathcal{B})$, with $Q \ll P$, absolutely continuous with respect to P , and where we set $\varphi(x) := -\log x + x - 1$ for all $x > 0$. Note that $KL_m(Q, P) \geq 0$ is nonnegative, and becomes equal to 0, if and only if $Q = P$. In the following, we will use the notation $KL_m(\mathbb{C}_{\alpha_1}, \mathbb{C}_{\alpha_2})$, to denote the KL_m -divergence between the two measures of probability, on $([0, 1]^d, \mathcal{B}([0, 1]^d))$, induced by the copulas $\mathbb{C}(\cdot, \alpha_1)$ and $\mathbb{C}(\cdot, \alpha_2)$. Let $\mathbb{C}_n(\cdot)$ denote the empirical copula associated to the data, i.e.,

$$\mathbb{C}_n(\mathbf{u}) := \frac{1}{n} \sum_{i=1}^n \prod_{j=1}^d \mathbb{1}_{\{F_{jn}(X_{ji}) \leq u_j\}}, \quad \text{for } \mathbf{u} \in [0, 1]^d. \quad (10.17)$$

Let $\mathbb{C}_k(\cdot)$ be the empirical copula based on the first k observations and let $\mathbb{C}_{n-k}(\cdot)$ denote the empirical copula based on the last $n - k$ observations, namely,

$$\mathbb{C}_k(\mathbf{u}) := \frac{1}{k} \sum_{i=1}^k \prod_{j=1}^d \mathbb{1}_{\{F_{jn}(X_{ji}) \leq u_j\}}, \quad \mathbb{C}_{n-k}(\mathbf{u}) := \frac{1}{n-k} \sum_{i=1}^{n-k} \prod_{j=1}^d \mathbb{1}_{\{F_{jn}(X_{ji}) \leq u_j\}}, \quad \text{for all } \mathbf{u} \in [0, 1]^d.$$

Note that the marginals are estimated using the whole data. In the sequel, we will use the following additional assumptions.

(C.8) $k^* = k^*(n)$, $0 < \lim_{n \rightarrow \infty} k^*(n)/n = \lambda < 1$ and $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$ are fixed elements in the interior of $\boldsymbol{\Theta}$;

(C.9) $\boldsymbol{\theta}^* := \arg \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} KL_m(\mathbf{C}_{\boldsymbol{\theta}}, \lambda \mathbf{C}_{\boldsymbol{\theta}^{(1)}} + (1 - \lambda) \mathbf{C}_{\boldsymbol{\theta}^{(2)}})$ exists, is unique and interior point of $\boldsymbol{\Theta}$.

Recalling the definition $\mathbf{A}_k$ in (10.5), we observe that $\frac{-1}{n} \log \mathbf{A}_{k^*}$ can be rewritten into

$$\begin{aligned} \frac{-1}{n} \log \mathbf{A}_{k^*} &= -\frac{k^*}{n} \int_{[0,1]^d} \log \frac{\mathbf{c}(\mathbf{u}, \hat{\boldsymbol{\theta}}_n)}{\mathbf{c}(\mathbf{u}, \hat{\boldsymbol{\theta}}_{k^*})} d\mathbb{C}_{k^*}(\mathbf{u}) - \frac{(n-k^*)}{n} \int_{[0,1]^d} \log \frac{\mathbf{c}(\mathbf{u}, \hat{\boldsymbol{\theta}}_n)}{\mathbf{c}(\mathbf{u}, \hat{\boldsymbol{\theta}}_{n-k^*})} d\mathbb{C}_{n-k^*}(\mathbf{u}) \\ &= -\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{k^*}{n} \int_{[0,1]^d} \log \frac{\mathbf{c}(\mathbf{u}, \hat{\boldsymbol{\theta}}_n)}{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta})} d\mathbb{C}_{k^*}(\mathbf{u}) - \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{(n-k^*)}{n} \int_{[0,1]^d} \log \frac{\mathbf{c}(\mathbf{u}, \hat{\boldsymbol{\theta}}_n)}{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta})} d\mathbb{C}_{n-k^*}(\mathbf{u}), \end{aligned}$$

which, under some suitably regularity conditions, converges in probability to the non-random quantity

$$\begin{aligned} \Xi(\lambda, \boldsymbol{\theta}^*, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) &:= -\inf_{\boldsymbol{\theta} \in \Theta} \lambda \int_{[0,1]^d} \log \frac{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta}^*)}{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta})} d\mathbb{C}(\mathbf{u}, \boldsymbol{\theta}^{(1)}) - \inf_{\boldsymbol{\theta} \in \Theta} (1-\lambda) \int_{[0,1]^d} \log \frac{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta}^*)}{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta})} d\mathbb{C}(\mathbf{u}, \boldsymbol{\theta}^{(2)}) \\ &= -\lambda \int_{[0,1]^d} \log \frac{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta}^*)}{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta}^{(1)})} d\mathbb{C}(\mathbf{u}, \boldsymbol{\theta}^{(1)}) - (1-\lambda) \int_{[0,1]^d} \log \frac{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta}^*)}{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta}^{(2)})} d\mathbb{C}(\mathbf{u}, \boldsymbol{\theta}^{(2)}) \\ &=: \lambda KL_m(\mathbb{C}_{\boldsymbol{\theta}^*}, \mathbb{C}_{\boldsymbol{\theta}^{(1)}}) + (1-\lambda) KL_m(\mathbb{C}_{\boldsymbol{\theta}^*}, \mathbb{C}_{\boldsymbol{\theta}^{(2)}}). \end{aligned} \quad (10.18)$$

At this point we may refer to (Broniatowski and Keziou, 2006, Theorem 4.4) and (Bouzebda and Keziou, 2010a, Theorem 4.2) for a general proof of the above convergence in the case of divergences in different frameworks. By (10.18), we can see that $\frac{-1}{n} \log \mathbf{\Lambda}_k$ can be interpreted as an empirical version of the mixture of KL_m divergences,

$$\Xi(\lambda, \boldsymbol{\theta}^*, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) := \lambda KL_m(\mathbb{C}_{\boldsymbol{\theta}^*}, \mathbb{C}_{\boldsymbol{\theta}^{(1)}}) + (1-\lambda) KL_m(\mathbb{C}_{\boldsymbol{\theta}^*}, \mathbb{C}_{\boldsymbol{\theta}^{(2)}}).$$

This observation and the above heuristic arguments allow us to give a description of the limiting behavior of $\mathbf{Z}_n$, under the alternative hypothesis $\mathcal{H}_1$. We will use the following additional conditions.

(C.10) We have $\int_{[0,1]^d} |h_\ell(\boldsymbol{\theta}_0, \mathbf{u})|^2 d\mathbb{C}(\mathbf{u}, \boldsymbol{\theta}_0) < \infty$, for $\ell = 1, 2$, where h_ℓ is defined by

$$h_\ell(\boldsymbol{\theta}, \mathbf{u}) := -\log \left\{ \frac{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta})}{\mathbf{c}(\mathbf{u}, \boldsymbol{\theta}^{(\ell)})} \right\}.$$

(C.11) (i) We have, for $\ell = 1, 2$, $\left\| \frac{\partial}{\partial \theta_j} h_\ell(\boldsymbol{\theta}, \mathbf{u}) \right\| \leq \text{constant} \times \prod_{i=1}^d \{u_i(1-u_i)\}^{-a_i}$ for some $a_i \geq 0$, such that $\mathbb{E}_{\boldsymbol{\theta}_0} \left\{ \prod_{i=1}^d \{F_i(X_i)(1-F_i(X_i))\}^{-2a_i} \right\} < \infty$;

(ii) We have, for $\ell = 1, 2$, $\left\| \frac{\partial^2}{\partial \theta_j \partial u_j} h_\ell(\boldsymbol{\theta}, \mathbf{u}) \right\| \leq \text{constant} \times \{u_j(1-u_j)^{b_j}\} \prod_{i=1, i \neq j}^d \{u_i(1-u_i)\}^{-a_i}$ for some $b_i > a_i \geq 0$, s.t. $\mathbb{E}_{\boldsymbol{\theta}_0} \left\{ \{F_j(X_j)(1-F_j(X_j))\}^{\xi_j - b_j} \prod_{i=1, i \neq j}^d \{F_i(X_i)(1-F_i(X_i))\}^{-2a_i} \right\} < \infty$, for some $\xi_j \in (0, 1/2)$.

Our result concerning this problem is stated in the theorem below.

Theorem 10.5. *Let the conditions (C.1-11) be fulfilled. Then, under the alternative hypothesis $\mathcal{H}_1$, we have, as $n \rightarrow \infty$, $\sqrt{n} \left(\frac{-1}{n} \log \mathbf{\Lambda}_k - \Xi(\lambda, \boldsymbol{\theta}^*, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) \right) \xrightarrow{d} N(0, \sigma^2)$, a centered normal distribution with variance equal to $\sigma^2 := \sigma^2(\lambda, \boldsymbol{\theta}^*, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) =$*

$$= \text{Var} \left\{ \sum_{i=1}^2 \lambda_i \left(h_i(\boldsymbol{\theta}^*, \mathbf{F}(\mathbf{X})) + \sum_{j=1}^d \int_{[0,1]^d} \frac{\partial}{\partial u_j} h_i(\boldsymbol{\theta}^*, \mathbf{u}) \mathbb{1}_{\{F_j(X_j) \leq u_j\}} d\mathbb{C}(\mathbf{u}, \boldsymbol{\theta}^{(i)}) \right) \right\} \quad (10.19)$$

where, $\lambda_1 := \lambda$ and $\lambda_2 := 1 - \lambda$.

The proof of this theorem follows the same lines of that of (Bouzebda and Keziou, 2010a, Theorem 4.2). The following corollary is a consequence of Theorem 10.5 and the use of the fact that $\widehat{k}_n - k^*(n) = O_P(1)$ under suitably conditions.

Corollary 10.1. *Let the conditions (C.1-11) be fulfilled. Then, under the alternative hypothesis $\mathcal{H}_1$, we have, as $n \rightarrow \infty$, $\frac{\sqrt{n}}{\sigma(\lambda, \boldsymbol{\theta}^*, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})} \left(\frac{\mathbf{Z}_n}{2n} - \Xi(\lambda, \boldsymbol{\theta}^*, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) \right) \xrightarrow{d} N(0, 1)$.*

Notice that Gombay (1997) was established the asymptotic distribution of the likelihood ratio under noncontiguous alternatives for the exponential family of distributions. In Theorem 1.5.1. of Csörgő and Horváth (1997), the asymptotic distribution of the likelihood ratio was obtained under condition similar to condition (C.8) and also for the exponential family of distributions, we may refer also to Theorem 1.3 of Gombay and Horváth (1996a). Our results cover the general case since there is no assumption on the form of $\mathbf{c}(\cdot, \boldsymbol{\theta}_0)$. We mention also our methodology for deriving the limiting distribution under the alternatives is quite different from those used in the preceding cited references.

10.3.1 Approximation of the power function of the test 2

Corollary 10.1 can be used to give an approximation to the power function $(\theta^{(1)}, \theta^{(2)}) \mapsto \beta(\theta^{(1)}, \theta^{(2)}) := \mathbb{P}_{\mathcal{H}_1}(\mathbf{Z}_n^{1/2} > q_\alpha(h, l))$ of the test 2 introduced in remark 10.7. We so obtain, for any fixed $\theta^{(1)} \neq \theta^{(2)}$ in the interior of Θ , the following approximation of the power

$$\beta(\theta^{(1)}, \theta^{(2)}) \approx 1 - \Phi \left(\frac{\sqrt{n}}{\hat{\sigma}(\hat{\lambda}, \hat{\theta}^*, \theta^{(1)}, \theta^{(2)})} \left(\frac{(q_\alpha(h, l))^2}{2n} - \hat{\Xi}(\hat{\lambda}, \hat{\theta}^*, \theta^{(1)}, \theta^{(2)}) \right) \right), \quad (10.20)$$

where Φ denotes, as usual, the distribution function of an $N(0, 1)$ standard normal random variable, and $\hat{\sigma}(\hat{\lambda}, \hat{\theta}^*, \theta^{(1)}, \theta^{(2)})$ and $\hat{\Xi}(\hat{\lambda}, \hat{\theta}^*, \theta^{(1)}, \theta^{(2)})$ are, respectively, the empirical versions of the parameters $\sigma(\lambda, \theta^*, \theta^{(1)}, \theta^{(2)})$ and $\Xi(\lambda, \theta^*, \theta^{(1)}, \theta^{(2)})$, where $\hat{\lambda} := \frac{\hat{k}_n}{n}$ and $\hat{\theta}^* := \arg \max_{\theta \in \Theta} \int_{[0,1]^d} \log(c(\mathbf{u}, \hat{\theta}_n)) d(\hat{\lambda} \mathbb{C}_{\hat{k}_n}(\mathbf{u}) + (1 - \hat{\lambda}_n) \mathbb{C}_{n - \hat{k}_n}(\mathbf{u}))$. Hence, $\hat{\Xi}(\hat{\lambda}, \hat{\theta}^*, \theta^{(1)}, \theta^{(2)})$ is the sample mean of

$$-\hat{\lambda}_n \log \frac{c(\mathbf{F}_n(\mathbf{X}_j), \hat{\theta}^*)}{c(\mathbf{F}_n(\mathbf{X}_j), \theta^{(1)})} \mathbb{1}_{\{1, \dots, \hat{k}_n\}}(j) - (1 - \hat{\lambda}_n) \log \frac{c(\mathbf{F}_n(\mathbf{X}_j), \hat{\theta}^*)}{c(\mathbf{F}_n(\mathbf{X}_j), \theta^{(2)})} \mathbb{1}_{\{\hat{k}_n + 1, \dots, n\}}(j), \quad j = 1, \dots, n,$$

and $\hat{\sigma}^2(\hat{\lambda}, \hat{\theta}^*, \theta^{(1)}, \theta^{(2)})$ is the sample variance of

$$\begin{aligned} & \hat{k}_n \left(h_1(\hat{\theta}^*, \mathbf{F}_n(\mathbf{X}_j)) + \sum_{m=1}^d W_m^{(1)}(\hat{\theta}^*, \mathbf{F}_n(\mathbf{X}_j)) \right) \mathbb{1}_{\{1, \dots, \hat{k}_n\}}(j) + \\ & (1 - \hat{k}_n) \left(h_2(\hat{\theta}^*, \mathbf{F}_n(\mathbf{X}_j)) + \sum_{m=1}^d W_m^{(2)}(\hat{\theta}^*, \mathbf{F}_n(\mathbf{X}_j)) \right) \mathbb{1}_{\{\hat{k}_n + 1, \dots, n\}}(j), \quad j = 1, \dots, n, \end{aligned}$$

where $W_m^{(i)}(\hat{\theta}^*, \mathbf{F}_n(\mathbf{X}_j)) := \int_{[0,1]^d} \frac{\partial}{\partial u_m} h_i(\hat{\theta}^*, \mathbf{u}) \mathbb{1}_{\{F_m(X_{mj}) \leq u_m\}} d\mathbb{C}(\mathbf{u}, \theta^{(i)}), \quad i = 1, 2, m = 1, \dots, d.$

A useful consequence of (10.20) is the obtaining of an approximate value of the minimal sample size ensuring a specified power $\beta := \beta(\theta^{(1)}, \theta^{(2)})$, with respect to some pre-assigned alternative $\mathcal{H}_1$. Let n_0 be the positive root of the equation

$$\beta = 1 - \Phi \left(\frac{\sqrt{n}}{\hat{\sigma}(\hat{\lambda}, \hat{\theta}^*, \theta^{(1)}, \theta^{(2)})} \left(\frac{(q_\alpha(h, l))^2}{2n} - \hat{\Xi}(\hat{\lambda}, \hat{\theta}^*, \theta^{(1)}, \theta^{(2)}) \right) \right). \quad (10.21)$$

The sought-after approximate value of the sample size is then given by $n^* := \lfloor n_0 \rfloor + 1$, where $\lfloor n_0 \rfloor$ is the integer part of n_0 .

10.4 Concluding remarks and directions for future research

We have considered the statistical test for the change point problem in the semiparametric copula models. The test is based on pseudo-likelihood ratio and the limiting distributions are characterized under both the null and alternative hypotheses. An other alternative to the present proposed test, is to define, in a semiparametric context with finite or infinite dimension, test statistics based on the estimates of φ -divergences in tow-samples semiparametric density ratio models (for the ratio joint distribution/product distribution of marginals), see Keziou and Leoni-Aubin (2008) and Keziou and Regnault (2015), or more generally for the change point problem in distributions. The problem of multiple change point can be investigated using our proposed estimates of multivariate ψ -divergences in multi-sample density ratio models; see Keziou (2015). These developments will be investigated in future works.

Chapter 11

Other contribution : ICA and BSS using divergences and copulas

We present in this Chapter our contribution in ICA, and Blind sources separation of (free-noise or noised) mixtures of independent/dependent sources, using φ -divergences and copulas.

In [Keziou *et al.* \(2008\)](#); [Ould Mohamed *et al.* \(2010\)](#), we have proposed BSS criterion for free-noise mixtures of second order cyclo-stationnary independent sources (with known and different cyclic frequencies). This work is a part of the PhD thesis of S. M. Ould Mohamed, see [Ould Mohamed \(2010\)](#), that I have supervised at 50%. The method results from the spectral analysis; it is based on the use of the second order cyclostationary characteristics of the source signals. The theoretical results are presented and proved in [Keziou *et al.* \(2008\)](#); [Ould Mohamed *et al.* \(2010\)](#), and the simulation results show a good performance of the proposed method in comparison with Sobi, CycloSobi and Jade algorithms.

In our paper [El Rhabi *et al.* \(2013\)](#), we have considered the BSS problem for noised mixtures of independent sources. We have proposed a new methodology in order to enhance separation performances in terms of efficiency and robustness. Our approach consists in denoising the observed signals through the minimization of their total variation, and then minimizing divergence separation criteria (between probability densities) combined with the total variation of the estimated source signals. We have showed by the way that the method leads to some projection problems that are solved by means of projected gradient algorithms. The efficiency and robustness of the proposed algorithm using Hellinger and Kullback-Leibler divergences, are illustrated and compared with the classical mutual information approach, through numerical simulations; see [El Rhabi *et al.* \(2013\)](#).

In [Keziou *et al.* \(2014\)](#), we have considered BSS problem for free-noise mixtures of independent/dependent sources. We have introduced a new blind source separation approach, based on minimization of Kullback-Leibler divergence between copula densities, for both independent or dependent source component signals. This work is a part of the PhD thesis of A. Ghazdali, see [Ghazdali \(2015\)](#), that I have supervised at 50%. In the classical case of independent source components, the proposed method generalizes the mutual information (between probability densities) procedure. Moreover, it has the great advantage to be naturally extensible to separate mixtures of dependent source components. Simulation results shows the convergence and the efficiency of the proposed algorithms, see [Keziou *et al.* \(2014\)](#).

In our paper [Ghazdali *et al.* \(2017\)](#), we have considered the BSS problem in the case of noised mixtures of independent/dependent sources. We have introduced a new method

for Blind Source Separation (BSS) in noisy instantaneous mixtures of both independent or dependent source component signals. The proposed approach is based on the minimization of a regularized criterion. Precisely, it consists in combining the total variation method for denoising with the Kullback-Leibler divergence between copula densities. This latter takes advantage of the copula to model the structure of the dependence between signal components. The obtained algorithm achieves separation in a noisy context where standard BSS methods fail. The efficiency and robustness of the proposed approach are illustrated by numerical simulations; for more detail we refer to [Ghazdali *et al.* \(2017\)](#). We limit our self to present hereafter the main results of only the paper [Ghazdali *et al.* \(2017\)](#), since it treats the more general case and uses our previous works [El Rhabi *et al.* \(2013\)](#), [Keziou *et al.* \(2014\)](#).

11.1 Introduction

The blind source separation (BSS) problem is a fundamental issue in applications of many different fields such as signal and image processing, biological and medical data analysis, communications, etc. The BSS aims to recover unknown source signals from a set of observations which are unknown mixtures of source signals. In order to separate the mixtures, different assumptions on the mixture and on the sources have to be made. In the literature, the most common assumptions are the linearity of the mixtures and the statistical independence of the source components and the condition that at most one of the components is gaussian. Under these assumptions, the BSS problem is linked to the well known problem of Independent Component Analysis (ICA), see e.g. [Comon \(1994\)](#). Using the above conditions, particularly, for noise-free mixtures, many methods of BSS have been proposed in the literature, see e.g. [Mansour and Jutten \(1996\)](#); [Taleb and Jutten \(1997\)](#); [Hyvärinen and Oja \(1997\)](#); [Pham \(2000\)](#). Some of these procedures use second or higher order statistics [Belouchrani *et al.* \(1997\)](#); [Pesquet and Moreau \(2001\)](#), maximize likelihood [Cardoso \(1998\)](#), maximize nongaussianity [Novey and Adali \(2005\)](#), minimize the mutual information [Pham \(2002\)](#) or φ -divergences [Keziou *et al.* \(2009\)](#). An interesting overview on the subject can be found in [Comon and Jutten \(2010\)](#). Recently, under the assumption that the source components are independent, based on estimation of mutual information and total variation (TV) regularization, [El Rhabi *et al.* \(2013\)](#) provided a new BSS algorithm to separate noisy mixtures of instantaneous or covolutive mixtures of independent source components. Further, it has been shown in [Keziou *et al.* \(2014\)](#) that, based on copula models, without the assumption of independence of the source components, we can still identify both mixing matrix and sources uniquely (up to scale and permutation indeterminacies) of (noise-free) mixtures of both independent and dependent source components. Motivated by various applications, we investigate, in the present Chapter, models of noisy linear instantaneous mixtures of independent/dependent sources, for which we propose, based on our previous works [El Rhabi *et al.* \(2013\)](#) and [Keziou *et al.* \(2014\)](#), a new BSS procedure. We successfully combined the two methods of [El Rhabi *et al.* \(2013\)](#) and [Keziou *et al.* \(2014\)](#) to obtain a new BSS algorithm able to separate noisy mixtures of independent/dependent source components. Our methodology consists in denoising the observed signals through the minimization of their total variation, and then minimizing Kullback-Leibler divergence, between copula densities, penalized by the total variation of the estimated source signals. The outline of this Chapter is organized as follows. In Section [11.2](#), we state the BSS problem in a noisy mixture context. In Section [11.3](#), we give a brief introduction on copulas and some of their main properties. In Section [11.4](#), we describe our approach. Section [11.5](#) presents how to implement the proposed approach using both numerical and statistical techniques. Section [6](#) in [Ghazdali *et al.* \(2017\)](#) gives some numerical results, illustrating the efficiency and robustness of our proposed method. Conclusions and possible developments are drawn in Section [11.6](#).

11.2 The BSS model

11.2.1 BSS model in the noise-free case

BSS problem can be modeled as follows. Denoting $\mathbf{A}[\cdot]$ the (unknown) mixing operator, the relationship between the observed and source signals can be written as

$$\mathbf{x}(t) := \mathbf{A}[\mathbf{s}(t)] \in \mathbb{R}^m, \quad (11.1)$$

where $\mathbf{s}(t) \in \mathbb{R}^m$ is the unknown vector of source signals to be estimated, and $\mathbf{x}(t)$ represents the observed signal vector at time $t \in [0, T]$. The goal of BSS, is therefore to estimate the unknown sources $\mathbf{s}(t)$ from the observed mixtures $\mathbf{x}(t)$. The estimation is performed with no prior information about either the sources or the mixing operator $\mathbf{A}[\cdot]$. Specific restrictions are made on both the mixing model and the source signals in order to limit the generality. We will restrict our self to the case where the number of source components and the number of observed mixture ones are equal, and we assume, in the present Chapter, that the mixtures are linear and instantaneous, so that the mixing operator $\mathbf{A}$ can be considered as a $m \times m$ matrix. In this case, supposing in addition that $\mathbf{A}$ is invertible, the candidate estimates of the sources will be obviously of the form

$$\mathbf{y}(t) := \mathbf{B} \mathbf{x}(t) \in \mathbb{R}^m, \quad t \in [0, T], \quad (11.2)$$

where $\mathbf{B} \in \mathbb{R}^{m \times m}$ represents an appropriate demixing matrix. In other words, the problem is to obtain an estimate, denote it $\hat{\mathbf{B}}$, “closing” as much as possible to the (unknown) ideal solution $\mathbf{B} = \mathbf{A}^{-1}$, by the use of only the observed vector signal $\mathbf{x}(t)$, leading to accurate estimate $\hat{\mathbf{s}}(t)$ of the source vector signal $\mathbf{s}(t)$,

$$\hat{\mathbf{s}}(t) := \hat{\mathbf{B}} \mathbf{x}(t) \in \mathbb{R}^m, \quad t \in [0, T]. \quad (11.3)$$

11.2.2 BSS model for noisy mixtures

We will focus on the BSS problem for noisy linear instantaneous mixtures

$$\bar{\mathbf{x}}(t) := \mathbf{A} \mathbf{s}(t) + \mathbf{n}(t) \in \mathbb{R}^m, \quad t \in [0, T], \quad (11.4)$$

where $\bar{\mathbf{x}}(t) \in \mathbb{R}^m$ is the vector of noisy observations, $\mathbf{s}(t) \in \mathbb{R}^m$ is the unknown vector of sources to be estimated, $\mathbf{n}(t) \in \mathbb{R}^m$ is the (unknown) noise vector, and $\mathbf{A} \in \mathbb{R}^{m \times m}$ is the unknown mixing $m \times m$ -matrix. We assume that the source signal $\mathbf{s}(t)$ and the noise $\mathbf{n}(t)$ are statistically independent. The mixing model (11.4) can also be written as

$$\bar{\mathbf{x}}(t) = \mathbf{x}(t) + \mathbf{n}(t), \quad (11.5)$$

where

$$\mathbf{x}(t) := \mathbf{A} \mathbf{s}(t) \quad (11.6)$$

is the inaccessible noise-free mixed vector signals. The aim here, is to estimate the sources $\mathbf{s}(t)$ using only the observed noisy signal $\bar{\mathbf{x}}(t)$. The presence of noise $\mathbf{n}(t)$ within the mixing model (11.4), as well as the possible dependency structure of the source components, complicate significantly the BSS problem. Unlike the above case of free-noise mixtures, see Subsection 11.2.1, the separating system of the form

$$\bar{\mathbf{y}}(t) := \mathbf{B} \bar{\mathbf{x}}(t) \quad (11.7)$$

does not provide a valid estimation for the source signals $\mathbf{s}(\cdot)$. In fact, the obtained signals according to the above system, can be written as

$$\begin{aligned} \bar{\mathbf{y}}(t) := \mathbf{B} \bar{\mathbf{x}}(t) &= \mathbf{B} \mathbf{x}(t) + \mathbf{B} \mathbf{n}(t) \\ &=: \mathbf{y}(t) + \bar{\mathbf{n}}(t), \end{aligned} \quad (11.8)$$

where $\bar{\mathbf{y}}(t) \in \mathbb{R}^m$ is the “noisy estimate” of $\mathbf{s}(t)$. That is the noisy estimated source $\bar{\mathbf{y}}(t)$ is the sum of

$$\mathbf{y}(t) := \mathbf{B} \mathbf{x}(t) \quad (11.9)$$

the “noisy-free” estimate of the sources, which is not feasible, and the unknown “noise”

$$\bar{\mathbf{n}}(t) := \mathbf{B} \mathbf{n}(t). \quad (11.10)$$

Ideally, we would like to retrieve $\mathbf{y}(t)$ by denoising $\bar{\mathbf{y}}(t)$, but it is rather difficult since the noise $\bar{\mathbf{n}}(t)$ is in turn unknown. During last years, several algorithms have been proposed to tackle the noisy BSS problem; In [Sahmoudi et al. \(2006\)](#), the authors proposed a two-step approach by combining the fractional lower order statistics for the mixing matrix estimation and minimum entropy criterion for noise-free source components estimation with gradient-based BSS algorithms. In [Belouchrani and Cichocki \(2000\)](#), a whitening procedure is proposed to reduce the noise effect. At our knowledge, the noisy BSS problem for possibly dependent sources has not been considered in the literature. In [Keziou et al. \(2014\)](#), the authors proposed a new criterion to successfully separate noise-free mixtures of both independent and dependent sources, based on estimation of Kullback-Leibler divergence between copula densities. [El Rhabi et al. \(2013\)](#) provides a BSS procedure for noisy mixtures using TV-regularization of φ -mutual information, which applies exclusively under the assumption that the source components are independent. Based on these two last works, the present Chapter introduces a new procedure for estimating sources in the context of noisy mixtures of possibly dependent source components. It proceeds on two stages : (i) denoising the observed signal $\bar{\mathbf{x}}(t)$ before demixing; (ii) a simultaneous BSS-denoisy procedure via minimizing a TV-regularized measure between copula densities.

11.3 Brief recall on copula

When modeling multivariate distributions, one has to take into account the effects of the marginal distributions as well as the dependence between them. This can be achieved by using the copula approach, which allows to deal with the margins and the dependency structure separately. The concept of copulas was introduced by [Sklar \(1959\)](#) as a function which couples a joint distribution function with its univariate margins. Copula is considered as a multivariate distribution function on the unit cube $[0, 1]^p$, with uniformly distributed marginals. Precisely, it can be presented as follows. Consider any random vector $\mathbf{Y} := (Y_1, \dots, Y_m)^\top \in \mathbb{R}^m$, $m \geq 2$, with joint distribution function (d.f.) $\mathbb{F}_{\mathbf{Y}}(\cdot) : \mathbf{y} \in \mathbb{R}^m \mapsto \mathbb{F}_{\mathbf{Y}}(\mathbf{y}) := \mathbb{F}_{\mathbf{Y}}(y_1, \dots, y_m) := \mathbb{P}(Y_1 \leq y_1, \dots, Y_m \leq y_m)$, and continuous marginal d.f.'s $F_{Y_j}(\cdot) : y_j \in \mathbb{R} \mapsto F_{Y_j}(y_j) := \mathbb{P}(Y_j \leq y_j)$, $\forall j = 1, \dots, m$. The characterization theorem of [Sklar \(1959\)](#) shows that there exists a unique m -variate function $\mathbb{C}_{\mathbf{Y}}(\cdot) : [0, 1]^m \mapsto [0, 1]$, such that, $\mathbb{F}_{\mathbf{Y}}(\mathbf{y}) = \mathbb{C}_{\mathbf{Y}}(F_{Y_1}(y_1), \dots, F_{Y_m}(y_m))$, $\forall \mathbf{y} := (y_1, \dots, y_m)^\top \in \mathbb{R}^m$. The function $\mathbb{C}_{\mathbf{Y}}(\cdot)$ is called a copula, and it is in itself a joint d.f. on $[0, 1]^m$, with uniform margins. We have $\forall \mathbf{u} := (u_1, \dots, u_m)^\top \in [0, 1]^m$, $\mathbb{C}_{\mathbf{Y}}(\mathbf{u}) = \mathbb{P}(F_{Y_1}(Y_1) \leq u_1, \dots, F_{Y_m}(Y_m) \leq u_m)$. Conversely, for any marginal d.f.'s $F_1(\cdot), \dots, F_m(\cdot)$, and copula $\mathbb{C}(\cdot)$, the function $\mathbb{C}(F_1(\cdot), \dots, F_m(\cdot))$ is a multivariate d.f. on $\mathbb{R}^m$. On the other hand, since the marginal d.f.'s $F_{Y_j}(\cdot)$, $j = 1, \dots, m$, are assumed to be continuous, the random variables $U_1 := F_{Y_1}(Y_1), \dots, U_m := F_{Y_m}(Y_m)$ are uniformly distributed on the interval $[0, 1]$, and the distribution function of the random vector $\mathbf{U} := (U_1, \dots, U_m)^\top \in [0, 1]^m$ is a copula. Moreover, we have that the components $Y_1, \dots, Y_m$ are statistically independent if and only if (iff) the copula $\mathbb{C}_{\mathbf{Y}}(\cdot)$ of the random vector $\mathbf{Y} := (Y_1, \dots, Y_m)^\top \in \mathbb{R}^m$ writes $\mathbb{C}_{\mathbf{Y}}(\mathbf{u}) = \prod_{j=1}^m u_j =: \mathbb{C}_{\Pi}(\mathbf{u})$, $\forall \mathbf{u} \in [0, 1]^m$, which will be denoted $\mathbb{C}_{\Pi}(\cdot)$; It is called the copula of independence. Define, when it exists, the copula density $c_{\mathbf{Y}}(\mathbf{u}) := \frac{\partial^m \mathbb{C}_{\mathbf{Y}}(\mathbf{u})}{\partial u_1 \dots \partial u_m}$, $\forall \mathbf{u} \in [0, 1]^m$. Hence, the copula density of independence, denote it $c_{\Pi}(\cdot)$, is the function taking the value 1 on $[0, 1]^m$ and zero otherwise, namely,

$$c_{\Pi}(\mathbf{u}) := \mathbf{1}_{[0, 1]^m}(\mathbf{u}), \quad \forall \mathbf{u} \in [0, 1]^m. \quad (11.11)$$

Let $f_{\mathbf{Y}}(\cdot)$, if it exists, be the probability density of the random vector $\mathbf{Y} := (Y_1, \dots, Y_m)^\top$, and, respectively, $f_{Y_1}(\cdot), \dots, f_{Y_m}(\cdot) \in \mathbb{R}^m$, the marginal probability densities of the random variables $Y_1, \dots, Y_m$.

Then, a straightforward computation shows that, for all vector $\mathbf{y} := (y_1, \dots, y_m)^\top \in \mathbb{R}^m$, we have the relation

$$f_{\mathbf{Y}}(\mathbf{y}) = c_{\mathbf{Y}}(F_{Y_1}(y_1), \dots, F_{Y_m}(y_m)) \prod_{j=1}^m f_{Y_j}(y_j). \quad (11.12)$$

In the literature, many models have been proposed for copulas. An important class of those models is that of semiparametric copula models, which has been widely used in survival analysis, where modeling and estimating the dependency structure, between survival variables, are fundamental problems. Models of this class are based on parametric copulas $\mathbb{C}(\cdot, \boldsymbol{\theta})$, indexed by a parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, with nonparametric margins. For more details on modeling theory as well as surveys of the commonly semiparametric used copulas, we can refer to [Nelsen \(2006\)](#); [Joe \(1997\)](#).

11.4 A combined TV-Copula approach for BSS

Our TV-Copula approach for BSS proceeds in two steps. (i) uses the TV regularization technique for denoising the observed signals. (ii) separates the mixtures by minimizing Kullback-Leibler divergence between copulas, penalized by a TV-regularization term. In the following, we describe these two steps.

11.4.1 Setp 1 : Denoising the observed signals

As in [El Rhabi et al. \(2013\)](#), we would obtain the denoised mixture signal

$$\mathbf{x}(t) := (x_1(t), \dots, x_m(t))^\top, \quad t \in [0, T], \quad (11.13)$$

from the (noisy) observed one $\bar{\mathbf{x}}(t)$, see (11.4), by means of the following variational minimization problem

$$x_i(\cdot) = \arg \min_{w_i \in \mathcal{X}_c} \left\{ \frac{1}{T} \int_0^T \frac{1}{2} (w_i(t) - \bar{x}_i(t))^2 dt + \lambda \frac{1}{T} \int_0^T \phi(w'_i(t)) dt \right\}, \quad i = 1, \dots, m, \quad (11.14)$$

where $\lambda > 0$ is the parameter of penalization, $\phi(\cdot) : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a well chosen function, $w'_i(t)$ is the first derivative of the signal $w_i(t)$ with respect to time t , and $\mathcal{X}_c$ is some appropriate space of real functions on $[0, T]$. In practice, in order to reduce the noise, the function $\phi(\cdot)$ is chosen to encourage smoothing in regions where the variations of the signal are weak, that is $w'_i \approx 0$, and to preserve discontinuities where w'_i is strong. The total variation case corresponds to the choice $\phi(x) = |x|$, which will be adopted here. One can make other choices like $\phi(x) = \sqrt{1 + x^2}$. A natural space when dealing with the above continuous variational problem (11.14), with $\phi(x) = |x|$, is $\mathcal{X}_c = BV([0, T])$, the space of all real valued functions on the interval $[0, T]$ with bounded variation; see e.g. [Evans and Gariepy \(1992\)](#) and [Cohen et al. \(2003\)](#). Here T is the observation time of the signals. The mathematical treatment of the above problem, in its continuous form, is beyond the scope of this Chapter. We will consider only the corresponding discrete version which will be described in Subsection 11.5.1 below.

11.4.2 Step 2 : BSS via minimizing TV-regularized KL-divergence between copulas

The aim of the following step is to construct an estimated source signal $\hat{\mathbf{s}}(t) := \hat{\mathbf{B}} \mathbf{x}(t)$ from the denoised signal $\mathbf{x}(t) := (x_1(t), \dots, x_m(t))^\top$ obtained as solution of (11.14). Our approach consists in minimizing, with respect to $\mathbf{B}$ on the demixing matrices space, an estimate (to be defined below) with some criterion of the form

$$\mathbf{B} \mapsto \mathcal{J}(\mathbf{B}) := \mathcal{J}_{\text{sep}}(\mathbf{B}) + \mathcal{J}_{\text{reg}}(\mathbf{y}), \quad (11.15)$$

where $\mathbf{y}(t) := \mathbf{B} \mathbf{x}(t) \in \mathbb{R}^m$, and $\mathbf{x}(t) = (x_1(t), \dots, x_m(t))^\top \in \mathbb{R}^m$ is obtained from (11.14); $\mathcal{J}_{\text{sep}}(\cdot)$ is some separating criterion, while $\mathcal{J}_{\text{reg}}(\mathbf{y})$ is a regularization term devoted to denoising. For the denoising phase, for both cases, of independent or dependent source components, we propose to use as regularization term the following one

$$\mathcal{J}_{\text{reg}}(\mathbf{y}) := \gamma \frac{1}{2T} \int_0^T \|\mathbf{y}(t) - \bar{\mathbf{y}}(t)\|^2 dt + \mu \frac{1}{T} \int_0^T |\nabla \mathbf{y}(t)| dt, \quad \gamma > 0, \mu > 0, \quad (11.16)$$

where $\bar{\mathbf{y}} = \mathbf{B} \bar{\mathbf{x}}$; the real numbers γ and μ are regularization parameters to be suitably chosen by the user. In the above display, $\|\cdot\|$ is used to denote the Euclidian norm on $\mathbb{R}^m$, and $|\nabla \mathbf{y}(t)| := \sum_{i=1}^m |y'_i(t)|$, for all t . The first term $\mathcal{J}_{\text{sep}}(\cdot)$, in (11.15), is a measure of dependence, or similarity between copula densities. Its form depends on the fact that the source components are independent or dependent. Hence, to describe the criterion $\mathcal{J}_{\text{sep}}(\cdot)$, we will consider separately, the case where the source components are independent and the case where they are dependent. First, we will consider the following stochastic modeling of the continuous time signals $\mathbf{s}(t)$, $\mathbf{x}(t)$ and $\mathbf{y}(t) := \mathbf{B} \mathbf{x}(t)$, $t \in [0, T]$. We assume that the above last random processes are stationary, so that the corresponding sampled versions, with certain time period, say T_e ,

$$\mathbf{s}(n) := \mathbf{s}(nT_e), \mathbf{x}(n) := \mathbf{x}(nT_e) \text{ and } \mathbf{y}(n) := \mathbf{y}(nT_e) = \mathbf{B} \mathbf{x}(n), n = 1, \dots, N, \quad (11.17)$$

can be considered as realizations of random vectors in $\mathbb{R}^m$, which will be denoted

$$\mathbf{S}, \quad \mathbf{X} \quad \text{and} \quad \mathbf{Y} := \mathbf{B} \mathbf{X}, \quad (11.18)$$

respectively. It has been shown, in Keziou *et al.* (2014), for the noisy-free mixtures case, that if we dispose of some prior information about the copula density of the random source vector $\mathbf{S}$, one can detect, under some assumptions, both the mixing matrix and the sources uniquely (up to scale and permutation indeterminacies) for both cases of independent or dependent source components. Denote by

$$\mathbb{F}_{\mathbf{Y}}(\cdot) : \mathbf{y} \in \mathbb{R}^m \mapsto \mathbb{F}_{\mathbf{Y}}(\mathbf{y}) := \mathbb{F}_{\mathbf{Y}}(y_1, \dots, y_m) := \mathbb{P}(Y_1 \leq y_1, \dots, Y_m \leq y_m), \quad (11.19)$$

the joint distribution function of the random vector $\mathbf{Y} := (Y_1, \dots, Y_m)^\top \in \mathbb{R}^m$, with continuous marginal distribution functions $F_{Y_i}(\cdot) : y_i \in \mathbb{R} \mapsto F_{Y_i}(y_i) := \mathbb{P}(Y_i \leq y_i)$, $\forall i = 1, \dots, m$. Let, if it exists, $f_{\mathbf{Y}}(\cdot)$ be the probability density of the random vector $\mathbf{Y}$, and, respectively, $f_{Y_1}, \dots, f_{Y_m}$, the marginal probability densities of the components $Y_1, \dots, Y_m$. The mutual information (MI) of $\mathbf{Y}$ is defined by

$$\mathbb{I}(\mathbf{Y}) := \int_{\mathbb{R}^m} \log \frac{f_{\mathbf{Y}}(\mathbf{y})}{\prod_{i=1}^m f_{Y_i}(y_i)} f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} = \mathbb{E} \left(\log \frac{f_{\mathbf{Y}}(\mathbf{Y})}{\prod_{i=1}^m f_{Y_i}(Y_i)} \right) \quad (11.20)$$

which equals the Kullback-Leibler divergence, denote it $\mathbb{K}(\cdot, \cdot)$, between the joint density $f_{\mathbf{Y}}(\cdot)$ and the product density $\prod_{i=1}^m f_{Y_i}(\cdot)$ of the margins $f_{Y_i}(\cdot)$, $i = 1, \dots, m$, that is

$$\mathbb{I}(\mathbf{Y}) = \mathbb{K} \left(f_{\mathbf{Y}}, \prod_{i=1}^m f_{Y_i} \right). \quad (11.21)$$

In (11.20), $\mathbb{E}(\cdot)$ is the mathematical expectation. Note also that the above criterion $\mathbb{I}(\mathbf{Y})$ is nonnegative and achieves its minimum value zero iff $f_{\mathbf{Y}}(\cdot) = \prod_{i=1}^m f_{Y_i}(\cdot)$, i.e., iff the components of the vector $\mathbf{Y}$ are statistically independent. We give now, the form of the separating term $\mathcal{J}_{\text{sep}}(\cdot)$ of the criterion (11.15).

The case of independent source components

Assume that the source components are independent. Using the relation (11.12), and applying the change variable formula for multiple integrals, one can show that $\mathbb{I}(\mathbf{Y})$ can be written, in terms of copula densities, as follows

$$\begin{aligned} \mathbb{I}(\mathbf{Y}) &= \int_{[0,1]^m} \log(c_{\mathbf{Y}}(\mathbf{u})) c_{\mathbf{Y}}(\mathbf{u}) d\mathbf{u} = \mathbb{E}(\log c_{\mathbf{Y}}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m))) =: -\mathbb{H}(c_{\mathbf{Y}}) \\ &= \int_{[0,1]^m} \log \left(\frac{c_{\mathbf{Y}}(\mathbf{u})}{1} \right) c_{\mathbf{Y}}(\mathbf{u}) d\mathbf{u} = \int_{[0,1]^m} \log \left(\frac{c_{\mathbf{Y}}(\mathbf{u})}{c_{\Pi}(\mathbf{u})} \right) c_{\mathbf{Y}}(\mathbf{u}) d\mathbf{u} \\ &= \mathbb{E} \left(\log \frac{c_{\mathbf{Y}}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m))}{c_{\Pi}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m))} \right) = \mathbb{K}(c_{\mathbf{Y}}, c_{\Pi}), \end{aligned} \quad (11.22)$$

where $c_{\mathbf{Y}}(\cdot)$ is the copula density of $\mathbf{Y}$, $c_{\Pi}(\cdot) := \mathbb{1}_{[0,1]^m}(\cdot)$ is the copula density of independence, and $\mathbb{E}(\cdot)$ the mathematical expectation operator. The above equations mean that the MI of a random vector

$\mathbf{Y}$ can be seen as the opposite of the Shannon entropy $\mathbb{H}(\cdot)$ of the copula density $c_{\mathbf{Y}}(\cdot)$ of $\mathbf{Y}$, or as the Kullback-Leibler divergence between the copula density $c_{\mathbf{Y}}(\cdot)$ of $\mathbf{Y}$ and the copula density $c_{\Pi}(\cdot)$ of independence. Moreover, $\mathbb{I}(\mathbf{Y}) = \mathbb{K}(c_{\mathbf{Y}}, c_{\Pi})$ is nonnegative and achieves its minimum value zero iff $c_{\mathbf{Y}}(\mathbf{u}) = c_{\Pi}(\mathbf{u})$, $\forall \mathbf{u} \in [0, 1]^m$, namely, iff the components of the vector $\mathbf{Y}$ are independent. The separating term will be chosen then to be

$$\mathcal{J}_{\text{sep}}(\mathbf{B}) := \mathcal{J}_{\text{sep}}^{\text{ind}}(\mathbf{B}) := \mathbb{K}(c_{\mathbf{Y}}, c_{\Pi}) = \mathbb{E}(\log c_{\mathbf{Y}}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m))). \quad (11.23)$$

We have that the function $\mathbf{B} \mapsto \mathcal{J}_{\text{sep}}^{\text{ind}}(\mathbf{B})$ is nonnegative and attains its minimum value zero iff $\mathbf{B} = \mathbf{A}^{-1}$ (up to scale and permutation indeterminacies of rows).

The case of dependent source components

In the case where the source components are dependent, we assume that we have at hand some prior information about the copula density of the random source vector $\mathbf{S}$. Note that this is possible for many practical problems, it can be done, see Section 11.3, from training sample $\mathbf{S}$, by a model selection procedure, see e.g. [Chen and Fan \(2006\)](#), among semiparametric copula density models $\{c_{\theta}(\cdot); \theta \in \Theta \subset \mathbb{R}^d\}$, typically indexed by a multivariate parameter θ ; see e.g. [Joe \(1997\)](#) and [Nelsen \(2006\)](#) for many examples of such models. The parameter θ can be estimated using maximum semiparametric likelihood, see e.g. [Genest et al. \(1995\)](#) and [Tsukahara \(2005\)](#). Denote then by $\hat{\theta}$, the obtained value of θ and $c_{\hat{\theta}}(\cdot)$ the copula density modeling the dependency structure of the source components. Obviously, since the source components are assumed to be dependent, $c_{\hat{\theta}}(\cdot)$ should be different from the density copula of independence $c_{\Pi}(\cdot)$. Hence, we naturally replace in (11.22), $c_{\Pi}(\cdot)$ by $c_{\hat{\theta}}(\cdot)$, and propose the following separating criterion $\mathcal{J}_{\text{sep}}(\cdot)$

$$\mathcal{J}_{\text{sep}}(\mathbf{B}) := \mathcal{J}_{\text{sep}}^{\text{dep}}(\mathbf{B}) := \int_{[0,1]^m} \log \left(\frac{c_{\mathbf{Y}}(\mathbf{u})}{c_{\hat{\theta}}(\mathbf{u})} \right) c_{\mathbf{Y}}(\mathbf{u}) d\mathbf{u} := \mathbb{E} \left(\log \frac{c_{\mathbf{Y}}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m))}{c_{\hat{\theta}}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m))} \right), \quad (11.24)$$

which equals $\mathbb{K}(c_{\mathbf{Y}}, c_{\hat{\theta}})$, the Kullback-Leibler divergence between the copula densities $c_{\mathbf{Y}}(\cdot)$ and $c_{\hat{\theta}}(\cdot)$. Moreover, one can show that the function $\mathbf{B} \mapsto \mathcal{J}_{\text{sep}}^{\text{dep}}(\mathbf{B})$ is nonnegative, and attains its minimum value iff $\mathbf{B} = \mathbf{A}^{-1}$ (up to scale and permutation indeterminacies of rows), provided that the copula density $c_{\hat{\theta}}(\cdot)$ of the random vector source $\mathbf{S}$ satisfies the following assumption : for any regular matrix $\mathbf{M}$, if the copula density of the random vector $\mathbf{M}\mathbf{S}$ equals $c_{\hat{\theta}}(\cdot)$, then $\mathbf{M} = \mathbf{D}\mathbf{P}$, for some diagonal matrix $\mathbf{D}$ and permutation matrix $\mathbf{P}$.

Remark 11.1. *The above criterion (11.24) supposes the knowledge of the copula density model $c_{\theta}(\cdot)$ (with known parameter θ), or the availability of training sample of the random vector source $\mathbf{S}$ from which we obtain the copula density $c_{\hat{\theta}}(\cdot)$ modeling the dependency structure of the source components, as described in Subsection 11.4.2. When only the model is known (with unknown parameter), we can adapt the above criterion as follows. Denote by $\{c_{\theta}(\cdot); \theta \in \Theta\}$ the density copula model of the dependency structure of the source components. The criterion (11.24) can be in this case replaced by*

$$\mathcal{J}_{\text{sep}}(\mathbf{B}) := \inf_{\theta \in \Theta} \int_{[0,1]^m} \log \left(\frac{c_{\mathbf{Y}}(\mathbf{u})}{c_{\theta}(\mathbf{u})} \right) c_{\mathbf{Y}}(\mathbf{u}) d\mathbf{u} := \inf_{\theta \in \Theta} \mathbb{E} \left(\log \frac{c_{\mathbf{Y}}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m))}{c_{\theta}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m))} \right). \quad (11.25)$$

In the more general case, when both the model and the parameter are unknown, the criterion may be adapted in a similar way as in our previous work [Keziou et al. \(2014\)](#) Section 5.3.

11.5 Discretization and Statistical estimation

In this Section, we describe how to make operational the above methodology. Precisely, using both numerical and statistical techniques, we give hereafter the discrete versions of the regularization terms (11.14) and (11.16), as well as statistical estimates of the separation terms (11.23) and (11.24). In all the sequel, the continuous time signals are sampled with a certain period T_e . We need to define additional

notation. To each scalar signal, say $u(t)$, $t \in [0, T]$, one associate a vector $\mathbb{U} := (u(1), \dots, u(N))^\top \in \mathcal{X}_d := \mathbb{R}^N$ such that $u(k) := u(kT_e)$ for all $k = 1, \dots, N$. The vector space $\mathcal{X}_d := \mathbb{R}^N$ will be equipped with the Euclidian inner product $\langle \mathbb{U}, \mathbb{V} \rangle := \sum_{k=1}^N u(k)v(k)$, for all $\mathbb{U}, \mathbb{V} \in \mathcal{X}_d$. For any $\mathbb{U} \in \mathcal{X}_d$, denote by $\mathbb{U}' := (u'(1), \dots, u'(N))^\top \in \mathcal{X}_d$, the vector obtained by numerical differentiation $u'(k) := \frac{u(k+1) - u(k)}{T_e}$, for all $k = 1, \dots, N-1$, and $u'(N) = 0$. Due to the discretization of the first derivatives, and in order to avoid the *edge effect*, we define the *backward derivative* of any $\mathbb{U} \in \mathcal{X}_d$, denoted $\mathbb{U}^* := (u^*(1), \dots, u^*(N))^\top$, by $u^*(1) := \frac{u(1)}{T_e}$, $u^*(N) := -\frac{u(N-1)}{T_e}$ and $u^*(k) := \frac{u(k) - u(k-1)}{T_e}$ for all $k = 2, \dots, N-1$. Here, the backward derivatives $(\cdot)^*$ is the discrete adjoint operator of $-(\cdot)'$. That is, for all $\mathbb{U}, \mathbb{V} \in \mathcal{X}_d$, we have $\langle \mathbb{U}^*, \mathbb{V} \rangle = -\langle \mathbb{U}, \mathbb{V}' \rangle$. All these definitions can be extended in a natural way to elements of the cartesian product space $\mathcal{X}_d^m := \mathbb{R}^m \times \mathbb{R}^N$. In the sequel, elements of the vector space $\mathcal{X}_d = \mathbb{R}^N$ will be denoted by double capital symbols, while the elements of the vector space $\mathcal{X}_d^m$ are denoted by bold double capital symbols, which can be considered as $m \times N$ -matrices.

11.5.1 Denoising the discrete observed signal

In this Section, we show how to compute the denoised observations

$$\mathbb{X} := \left(\mathbf{x}(n) = (x_1(n), \dots, x_m(n))^\top, n = 1, \dots, N \right)$$

from the observed noisy ones $\overline{\mathbb{X}} := \left(\overline{\mathbf{x}}(n) = (\overline{x}_1(n), \dots, \overline{x}_m(n))^\top, n = 1, \dots, N \right)$. This will be obtained by solving the discrete version

$$x_i(\cdot) = \arg \min_{\mathbb{W}_i \in \mathcal{X}_d} \left\{ \frac{1}{N} \sum_{n=1}^N \frac{1}{2} (w_i(n) - \overline{x}_i(n))^2 + \lambda \frac{1}{N} \sum_{n=1}^N |w'_i(n)| \right\}, i = 1, \dots, m, \quad (11.26)$$

of the optimization problem (11.14). We start with the following proposition which gives a simple characterization of the solution; see [Chambolle \(2004\)](#) for a proof.

Proposition 11.1. ([Chambolle \(2004\)](#)). *The discrete version (11.26) of the optimization problem (11.14) has a unique solution given by*

$$\mathbb{X} = \overline{\mathbb{X}} - \prod_{\lambda G} \overline{\mathbb{X}}, \quad (11.27)$$

where $\prod_{\lambda G}(\cdot)$ is the Euclidian orthogonal projection operator on the convex set λG with

$$G := \{\mathbb{V}^* \text{ such that } \mathbb{V} \in \mathcal{X}_d^m \text{ with } |v_i(n)| \leq 1, \forall (i, n) \in \{1, \dots, m\} \times \{1, \dots, N\}\}. \quad (11.28)$$

Thus, to compute $\mathbb{X}$, we are led to compute the projection operator $\prod_{\lambda G}(\cdot)$ on the convex set λG . For the denoising step, we propose then the Algorithm 1; the reader may refer to [El Rhabi et al. \(2013\)](#) for more details, particularly, for a proof on convergence of this algorithm. In all the sequel, the filtered observed signal $\mathbb{X}$ will be considered as the denoised version of the observed signal $\overline{\mathbb{X}}$.

Data: $\overline{\mathbb{X}}$ the observed noisy signal

Result: $\mathbb{X}$ the denoised signal

Initialization: $q = 0$, $\mathbb{X}^{(0)} = \overline{\mathbb{X}}$. Given $\epsilon > 0$, $\lambda > 0$ and $\rho > 0$ suitably chosen

Do: $q = q + 1$

for $j = 1, \dots, m$

$\mathbb{X}_j^{(q)} = \prod_{\lambda G}(\mathbb{X}_j^{(q-1)} + 2\lambda\rho(\lambda(\mathbb{X}_j^{(q-1)})^* - \overline{\mathbb{X}}_j)')$

end

Until $\|\mathbb{X}^{(q)} - \mathbb{X}^{(q-1)}\| < \epsilon$

$\mathbb{X} = \mathbb{X}^{(q)}$

Algorithm 1: The denoising step.

11.5.2 BSS by minimizing TV-regularized KL-divergence between copulas

The case of independent source components

Recall that, when the source components are independent, the criterion function (11.15) is defined by

$$\mathcal{J}^{\text{ind}}(\cdot) : \mathbf{B} \mapsto \mathcal{J}^{\text{ind}}(\mathbf{B}) := \mathcal{J}_{\text{sep}}^{\text{ind}}(\mathbf{B}) + \mathcal{J}_{\text{reg}}(\mathbf{y}), \quad (11.29)$$

where $\mathcal{J}_{\text{sep}}^{\text{ind}}(\mathbf{B})$ is given by, see (11.23), $\mathcal{J}_{\text{sep}}^{\text{ind}}(\mathbf{B}) := \mathbb{K}(c_{\mathbf{Y}}, c_{\Pi}) = \mathbb{E}(\log c_{\mathbf{Y}}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m)))$ and $\mathcal{J}_{\text{reg}}(\mathbf{y}) := \gamma \frac{1}{2T} \int_0^T \|\mathbf{y}(t) - \bar{\mathbf{y}}(t)\|^2 dt + \mu \frac{1}{T} \int_0^T |\nabla \mathbf{y}(t)| dt$, $\gamma > 0$, $\mu > 0$, with $\mathbf{y}(t) = \mathbf{B} \mathbf{x}(t)$ and $\bar{\mathbf{y}} = \mathbf{B} \bar{\mathbf{x}}(t)$. Using the stochastic modeling (11.17) and the relation (11.23), we propose to approximate the criterion (11.29) by

$$\mathbf{B} \mapsto \widehat{\mathcal{J}}^{\text{ind}}(\mathbf{B}) := \widehat{\mathcal{J}}_{\text{sep}}^{\text{ind}}(\mathbf{B}) + \mathcal{J}_{\text{reg,d}}(\mathbf{y}), \quad (11.30)$$

where

$$\mathcal{J}_{\text{reg,d}}(\mathbf{y}) := \frac{\gamma}{2N} \sum_{i=1}^N \|\mathbf{y}(i) - \bar{\mathbf{y}}(i)\|^2 + \frac{\mu}{N} \sum_{i=1}^N |\nabla \mathbf{y}(i)| \quad (11.31)$$

is the discrete version of $\mathcal{J}_{\text{reg}}(\mathbf{y})$, and $\mathbf{B} \mapsto \widehat{\mathcal{J}}_{\text{sep}}^{\text{ind}}(\mathbf{B})$ is the statistical estimate, of the separating criterion $\mathbf{B} \mapsto \mathcal{J}_{\text{sep}}^{\text{ind}}(\mathbf{B})$, which we define as follows

$$\mathbf{B} \mapsto \widehat{\mathcal{J}}_{\text{sep}}^{\text{ind}}(\mathbf{B}) := \frac{1}{N} \sum_{i=1}^N \log \left(\widehat{c}_{\mathbf{Y}}(\widehat{F}_{Y_1}(y_1(i)), \dots, \widehat{F}_{Y_m}(y_m(i))) \right), \quad (11.32)$$

where $\widehat{c}_{\mathbf{Y}}(\widehat{F}_{Y_1}(y_1(i)), \dots, \widehat{F}_{Y_m}(y_m(i))) := \frac{1}{NH_1 \dots H_m} \sum_{\ell=1}^m \prod_{j=1}^m k \left(\frac{\widehat{F}_{Y_j}(y_j(i)) - \widehat{F}_{Y_j}(y_j(\ell))}{H_j} \right)$, is the kernel estimate of the copula density $c_{\mathbf{Y}}(\cdot)$, and $\widehat{F}_{Y_j}(\cdot)$, $\forall j = 1, \dots, m$, is the smooth estimate of the marginal distribution function $F_{Y_j}(\cdot)$ of the random variable Y_j , defined for any real number $r \in \mathbb{R}$, by $\widehat{F}_{Y_j}(r) := \frac{1}{N} \sum_{\ell=1}^N K \left(\frac{r - y_j(\ell)}{h_j} \right)$, where $K(\cdot)$ is the primitive of a kernel $k(\cdot)$, a symmetric centered probability density.

An appropriate choice of the kernel $k(\cdot)$, for estimating the copula density, can be done according to Omelka *et al.* (2009), which copes with the boundary effect. The bandwidth parameters $H_1, \dots, H_m$ and $h_1, \dots, h_m$ will be chosen according to Silverman's rule of thumb, see e.g. Silverman (1986), i.e., for all $j = 1, \dots, m$, we take $H_j = \left(\frac{4}{m+2} \right)^{\frac{1}{m+4}} N^{\frac{-1}{m+4}} \widehat{\Sigma}_j$, and $h_j = \left(\frac{4}{3} \right)^{\frac{1}{5}} N^{\frac{-1}{5}} \widehat{\sigma}_j$, where $\widehat{\Sigma}_j$ and $\widehat{\sigma}_j$ are, respectively, the empirical standard deviation of the data $\widehat{F}_{Y_j}(y_j(1)), \dots, \widehat{F}_{Y_j}(y_j(N))$ and $y_j(1), \dots, y_j(N)$. The source signal vector $\mathbf{s}(i)$, $i = 1, \dots, N$, will be then estimated by $\widehat{\mathbf{s}}(i) = \widehat{\mathbf{B}} \mathbf{x}(i)$, $i = 1, \dots, N$, where $\widehat{\mathbf{B}} := \arg \inf_{\mathbf{B}} \widehat{\mathcal{J}}^{\text{ind}}(\mathbf{B})$, which can be computed using a gradient descent type algorithm. We can then derive the following Algorithm.

Data: $\bar{\mathbf{x}}$ the observed signal

Result: $\widehat{\mathbf{s}}$ the estimated source signal

Initialization: Compute $\mathbf{x} = \bar{\mathbf{x}} - \prod_{\lambda \in \mathcal{G}} \bar{\mathbf{x}}$ from Algorithm 1, $\mathbf{B}^{(0)} = \mathbf{I}_m$, $\mathbf{y}^{(0)} = \mathbf{B}^{(0)} \mathbf{x}$.

Given $\varepsilon > 0$, $\nu > 0$.

Do: Update $\mathbf{B}$ and $\mathbf{y}$:

$$\mathbf{B}^{(q+1)} = \mathbf{B}^{(q)} - \nu \frac{d\widehat{\mathcal{J}}^{\text{ind}}(\mathbf{B})}{d\mathbf{B}}$$

$$\mathbf{y}^{(q+1)} = \mathbf{B}^{(q+1)} \mathbf{x}.$$

Until $\|\mathbf{B}^{(q+1)} - \mathbf{B}^{(q)}\| < \epsilon$

$$\widehat{\mathbf{s}} = \mathbf{y}^{(q+1)}$$

Algorithm 2: BSS algorithm for independent source components.

The case of dependent source components

Recall that, in the case where the source components are dependent, the criterion function (11.15) is defined by

$$\mathcal{J}^{\text{dep}}(\cdot) : \mathbf{B} \mapsto \mathcal{J}^{\text{dep}}(\mathbf{B}) := \mathcal{J}_{\text{sep}}^{\text{dep}}(\mathbf{B}) + \mathcal{J}_{\text{reg}}(\mathbf{y}), \quad (11.33)$$

where $\mathcal{J}^{\text{dep}}(\mathbf{B})$ is defined by, see (11.24),

$$\mathcal{J}_{\text{sep}}^{\text{dep}}(\mathbf{B}) := \int_{[0,1]^m} \log \left(\frac{c_{\mathbf{Y}}(\mathbf{u})}{c_{\hat{\theta}}(\mathbf{u})} \right) c_{\mathbf{Y}}(\mathbf{u}) d\mathbf{u} = \mathbb{E} \left(\log \frac{c_{\mathbf{Y}}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m))}{c_{\hat{\theta}}(F_{Y_1}(Y_1), \dots, F_{Y_m}(Y_m))} \right).$$

and $\mathcal{J}_{\text{reg}}(\mathbf{y}) := \gamma \frac{1}{2T} \int_0^T \|\mathbf{y}(t) - \bar{\mathbf{y}}(t)\|^2 dt + \mu \frac{1}{T} \int_0^T |\nabla \mathbf{y}(t)| dt$, $\gamma > 0$, $\mu > 0$. Using the stochastic modeling (11.17) and the relation (11.24), we propose to approximate the criterion (11.33) by

$$\mathbf{B} \mapsto \widehat{\mathcal{J}^{\text{dep}}}(\mathbf{B}) := \widehat{\mathcal{J}_{\text{sep}}^{\text{dep}}}(\mathbf{B}) + \mathcal{J}_{\text{reg,d}}(\mathbf{y}), \quad (11.34)$$

where

$$\mathcal{J}_{\text{reg,d}}(\mathbf{y}) := \frac{\gamma}{2N} \sum_{i=1}^N \|\mathbf{y}(i) - \bar{\mathbf{y}}(i)\|^2 + \frac{\mu}{N} \sum_{i=1}^N |\nabla \mathbf{y}(i)| \quad (11.35)$$

is the discrete version of $\mathcal{J}_{\text{reg}}(\mathbf{y})$, and $\widehat{\mathcal{J}_{\text{sep}}^{\text{dep}}}(\cdot) : \mathbf{B} \mapsto \widehat{\mathcal{J}_{\text{sep}}^{\text{dep}}}(\mathbf{B})$ is the statistical estimate, of the criterion $\mathcal{J}_{\text{sep}}^{\text{dep}}(\cdot) : \mathbf{B} \mapsto \mathcal{J}_{\text{sep}}^{\text{dep}}(\mathbf{B})$, which we define as follows

$$\widehat{\mathcal{J}_{\text{sep}}^{\text{dep}}}(\cdot) : \mathbf{B} \mapsto \widehat{\mathcal{J}_{\text{sep}}^{\text{dep}}}(\mathbf{B}) := \frac{1}{N} \sum_{i=1}^N \log \left(\frac{\widehat{c}_{\mathbf{Y}}(\widehat{F}_{Y_1}(y(i)), \dots, \widehat{F}_{Y_m}(y_m(i)))}{\widehat{c}_{\hat{\theta}}(\widehat{F}_{Y_1}(y(i)), \dots, \widehat{F}_{Y_m}(y_m(i)))} \right). \quad (11.36)$$

The source signal vector $\mathbf{s}(i)$, $i = 1, \dots, N$, will be then estimated by $\widehat{\mathbf{s}}(i) = \widehat{\mathbf{B}} \mathbf{x}(i)$, $i = 1, \dots, N$, where $\widehat{\mathbf{B}} := \arg \inf_{\mathbf{B}} \widehat{\mathcal{J}^{\text{dep}}}(\mathbf{B})$, which can be computed using a gradient descent type algorithm. We obtain then the following Algorithm 3.

Data: $\bar{\mathbf{x}}$ the observed signal

Result: $\widehat{\mathbf{s}}$ the estimated source signal

Initialization: Compute $\mathbf{x} = \bar{\mathbf{x}} - \prod_{\lambda \in \Lambda} \bar{\mathbf{x}}$ from Algorithm 1, $\mathbf{B}^{(0)} = \mathbf{I}_m$, $\mathbf{y}^{(0)} = \mathbf{B}^{(0)} \mathbf{x}$.

Given $\varepsilon > 0$, $\nu > 0$.

Do: Update $\mathbf{B}$ and $\mathbf{y}$:

$$\mathbf{B}^{(q+1)} = \mathbf{B}^{(q)} - \nu \frac{d\widehat{\mathcal{J}^{\text{dep}}}(\mathbf{B})}{d\mathbf{B}}$$

$$\mathbf{y}^{(q+1)} = \mathbf{B}^{(q+1)} \mathbf{x}$$

Until $\|\mathbf{B}^{(q+1)} - \mathbf{B}^{(q)}\| < \epsilon$

$$\widehat{\mathbf{s}} = \mathbf{y}^{(q+1)}$$

Algorithm 3: BSS algorithm for dependent source components.

11.6 Conclusion and directions for future research

We have presented a new BSS approach in presence of noise, by minimizing a new regularized criterion. The approach is able to separate instantaneous linear mixtures of both independent and dependent

source components. It proceeds in two steps : a preprocessing which consists in reducing the noise, and a second one minimizing a separation criterion based on Kullback-Leibler divergence between copula densities regularized by total variation term. The accuracy and the consistency of the obtained algorithms have been illustrated by simulation. It should be mentioned, however, that our proposed algorithms based on copula densities, rather than the classical ones based on probability densities, are more time consuming, since we estimate both copulas density of the vector and the marginal distribution function of each component. Moreover, the number of mixtures/sources should be small, since the accuracy estimation of the sources decreases when the number of mixtures/sources increases, caused by the curse of dimensionality in nonparametric multivariate density estimation. The present approach can be extended to deal with convolutive mixtures of independent/dependent sources. In the case of independent sources, in order to cope with the problem of the curse of dimensionality (in nonparametric multivariate density estimation used for estimating the separation criterion, namely, the mutual information), we propose to develop the following idea for ICA. We propose to use, instead of the above estimate of the mutual information based on kernel density estimation of densities or copula densities, a new estimate of the separation criterion (a φ -mutual information) using conjugate representation of the φ -divergence between the joint distribution of $\mathbf{Y}$ and the product distribution of the components of $\mathbf{Y}$, in semiparametric ratio models. We have then to solve many problems, identifiability of the problem, model selection for the ratio, consistency of the estimates, the complexity of the corresponding algorithms and the extension of the proposed method to the case of dependent sources.

Bibliography

- Agresti, A. (1990). *Categorical data analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.
- Ali, S. and Silvey, S. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc., B*, **28**, 131–142.
- Anderson, J. A. (1979). Multivariate logistic compounds. *Biometrika*, **66**(1), 17–26.
- Aue, A. and Horváth, L. (2013). Structural breaks in time series. *J. Time Series Anal.*, **34**(1), 1–16.
- Baggerly, K. A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika*, **85**(3), 535–547.
- Basseville, M. and Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application*. Prentice Hall Information and System Sciences Series. Prentice Hall, Inc., Englewood Cliffs, NJ.
- Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Ann. Inst. Statist. Math.*, **46**(4), 683–705.
- Beirlant, J., Dudewicz, E. J., Györfi, L., and van der Meulen, E. C. (1997). Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.*, **6**(1), 17–39.
- Belouchrani, A. and Cichocki, A. (2000). Robust whitening procedure in blind source separation context. *Electronics letters*, **36**(24), 2050–2051.
- Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., and Moulines, E. (1997). A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, **45**(2), 434–444.
- Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.*, **5**(3), 445–463.
- Berlinet, A. (1999). How to get central limit theorems for global errors of estimates. *Appl. Math.*, **44**(2), 81–96.
- Berlinet, A., Vajda, I., and van der Meulen, E. C. (1998). About the asymptotic accuracy of Barron density estimates. *IEEE Trans. Inform. Theory*, **44**(3), 999–1009.
- Bertail, P. (2004). Empirical likelihood in nonparametric and semiparametric models. In *Parametric and semiparametric models with applications to reliability, survival analysis, and quality of life*, Stat. Ind. Technol., pages 291–306. Birkhäuser Boston, Boston, MA.
- Bertail, P. (2006). Empirical likelihood in some semiparametric models. *Bernoulli*, **12**(2), 299–331.
- Biau, G. and Devroye, L. (2005). Density estimation by the penalized combinatorial method. *J. Multivariate Anal.*, **94**(1), 196–208.
- Borwein, J. M. and Lewis, A. S. (1991). Duality relationships for entropy-like minimization problems. *SIAM J. Control Optim.*, **29**(2), 325–338.

- Borwein, J. M. and Lewis, A. S. (1992a). Partially finite convex programming. I. Quasi relative interiors and duality theory. *Math. Programming*, **57**(1, Ser. B), 15–48.
- Borwein, J. M. and Lewis, A. S. (1992b). Partially finite convex programming. II. Explicit lattice models. *Math. Programming*, **57**(1, Ser. B), 49–83.
- Borwein, J. M. and Lewis, A. S. (1993). Partially-finite programming in L_1 and the existence of maximum entropy estimates. *SIAM J. Optim.*, **3**(2), 248–267.
- Bouzebda, S. and Keziou, A. (2008). A test of independence in some copula models. *Math. Methods Statist.*, **17**(2), 123–137.
- Bouzebda, S. and Keziou, A. (2009). Estimation and tests of independence in copula models via divergences. *C. R. Math. Acad. Sci. Paris*, **347**(11-12), 667–672.
- Bouzebda, S. and Keziou, A. (2010a). New estimates and tests of independence in semiparametric copula models. *Kybernetika (Prague)*, **46**(1), 178–201.
- Bouzebda, S. and Keziou, A. (2010b). A new test procedure of independence in copula models via χ^2 -divergence. *Comm. Statist. Theory Methods*, **39**(1-2), 1–20.
- Bouzebda, S. and Keziou, A. (2010c). A semiparametric test of independence in copula models for censored data. *C. R. Math. Acad. Sci. Paris*, **348**(7-8), 449–453.
- Bouzebda, S. and Keziou, A. (2013). A semiparametric maximum likelihood ratio test for the change point in copula models. *Stat. Methodol.*, **14**, 39–61.
- Braun, J. V. and Muller, H.-G. (1998). Statistical methods for dna sequence segmentation. *Statistical Science*, pages 142–162.
- Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change-point problems*, volume 243 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht.
- Brøndsted, A. (1964). Conjugate convex functions in topological vector spaces. *Mat.-Fys. Medd. Danske Vid. Selsk.*, **34**(2), 27 pp. (1964).
- Broniatowski, M. (2003). Estimation of the Kullback-Leibler divergence. *Math. Methods Statist.*, **12**(4), 391–409 (2004).
- Broniatowski, M. and Keziou, A. (2004). Estimation and tests for models satisfying linear constraints with unknown parameter. *Preprint 2004-2, L.S.T.A - Université Paris 6*.
- Broniatowski, M. and Keziou, A. (2006). Minimization of ϕ -divergences on sets of signed measures. *Studia Sci. Math. Hungar.*; *arXiv:1003.5457*, **43**(4), 403–442.
- Broniatowski, M. and Keziou, A. (2009). Parametric estimation and tests through divergences and the duality technique. *J. Multivariate Anal.*, **100**(1), 16–36.
- Broniatowski, M. and Keziou, A. (2012). Divergences and duality for estimation and test under moment condition models. *J. Statist. Plann. Inference*, **142**(9), 2554–2573.
- Cai, S., Chen, J., and Zidek, J. V. (2013). Hypothesis testing in the presence of multiple samples under density ratio models. *arXiv preprint arXiv:1309.4740*.
- Cardoso, J.-F. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE*, **86**(10), 2009–2025.
- Castella, M., Rhioui, S., Moreau, E., and Pesquet, J.-C. (2007). Quadratic higher order criteria for iterative blind separation of a MIMO convolutive mixture of sources. *IEEE Trans. Signal Process.*, **55**(1), 218–232.

- Cellucci, C. J., Albano, A. M., and Rapp, P. E. (2005). Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Phys. Rev. E*, **71**, 066208.
- Chambolle, A. (2004). An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, **20**(1-2), 89–97.
- Chan, J., Horváth, L., and Hušková, M. (2013). Darling-Erdős limit results for change-point detection in panel data. *J. Statist. Plann. Inference*, **143**(5), 955–970.
- Chan, S.-O., Diakonikolas, I., Valiant, G., and Valiant, P. (2014). Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203. ACM, New York.
- Chao, A. and Shen, T.-J. (2003). Nonparametric estimation of Shannon’s index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.*, **10**(4), 429–443.
- Chen, J. and Gupta, A. K. (2000). *Parametric statistical change point analysis*. Birkhäuser Boston, Inc., Boston, MA.
- Chen, J. and Liu, Y. (2013). Quantile and quantile-function estimations under density ratio model. *Ann. Statist.*, **41**(3), 1669–1692.
- Chen, X. and Fan, Y. (2005). Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. *Canad. J. Statist.*, **33**(3), 389–414.
- Chen, X. and Fan, Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *J. Econometrics*, **135**(1-2), 125–154.
- Chen, X., Hong, H., and Shum, M. (2007). Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models. *J. Econometrics*, **141**(1), 109–140.
- Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. Wiley Finance Series. John Wiley & Sons, Ltd., Chichester.
- Choi, S. and Cichocki, A. (1997). Adaptive blind separation of speech signals: Cocktail party problem. In *International Conference on Speech Processing*, pages 617–622.
- Cichocki, A. and Amari, S.-i. (2010). Families of alpha- beta- and gamma-divergences: flexible and robust measures of similarities. *Entropy*, **12**(6), 1532–1568.
- Cohen, A., Wolfgang, D., , Daubechies, I., and DeVore, R. (2003). Harmonic analysis of the space bv. *Rev. Mat. Iberoamericana*, **19**(1), 235–263.
- Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing, IEEE*, **36**(3), 287–314.
- Comon, P. and Jutten, C. (2010). *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.
- Congedo, M., Gouy-Pailler, C., and Jutten, C. (2008). On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics. *Clinical Neurophysiology*, **119**(12), 2677 – 2686.
- Corcoran, S. (1998). Bertlett adjustment of empirical discrepancy statistics. *Biometrika*, **85**, 967–972.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.

- Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B*, **46**(3), 440–464.
- Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **8**, 85–108.
- Csiszár, I. (1967a). Information-type indices of the divergence of distributions . I, II. *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.* *17* (1967), 123–149; *ibid.*, **17**, 267–291.
- Csiszár, I. (1967b). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, **2**, 299–318.
- Csiszár, I. (1967c). On topology properties of f -divergences. *Studia Sci. Math. Hungar.*, **2**, 329–339.
- Csiszár, I. (1975). I -divergence geometry of probability distributions and minimization problems. *Ann. Probability*, **3**, 146–158.
- Csiszár, I. (1984). Sanov property, generalized I -projection and a conditional limit theorem. *Ann. Probab.*, **12**(3), 768–793.
- Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Math. Hungar.*, **68**(1-2), 161–186.
- Csiszár, I., Gamboa, F., and Gassiat, E. (1999). MEM pixel correlated solutions for generalized moment and interpolation problems. *IEEE Trans. Inform. Theory*, **45**(7), 2253–2270.
- Csörgő, M. and Horvath, L. (1988). Nonparametric methods for changepoint problems. *Handbook of statistics*, **7**, 403–425.
- Csörgő, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. With a foreword by David Kendall.
- Cui, S. and Sun, Y. (2004). Checking for the gamma frailty distribution under the marginal proportional hazards frailty model. *Statist. Sinica*, **14**(1), 249–267.
- Darbellay, G. A. and Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Inform. Theory*, **45**(4), 1315–1321.
- Darmois, G. (1953). Analyse générale des liaisons stochastiques. Etude particulière de l’analyse factorielle linéaire. *Rev. Inst. Internat. Statistique*, **21**, 2–8.
- De Young, M. and Evans, B. (2007). Blind source separation with a time-varying mixing matrix. In *in Proc. of the Forty-First Asilomar Conference on Signal, System & Computer, Pacific Grove California*, pages 626–630.
- Deheuvels, P. (1979a). La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d’indépendance. *Acad. Roy. Belg. Bull. Cl. Sci. (5)*, **65**(6), 274–292.
- Deheuvels, P. (1979b). Propriétés d’existence et propriétés topologiques des fonctions de dépendance avec applications à la convergence des types pour des lois multivariées. *C. R. Acad. Sci. Paris Sér. A-B*, **288**(2), A145–A148.
- Deheuvels, P. (1980). Nonparametric test of independence. In *Nonparametric asymptotic statistics (Proc. Conf., Rouen, 1979) (French)*, volume 821 of *Lecture Notes in Math.*, pages 95–107. Springer, Berlin.
- Deheuvels, P. (1981). A Kolmogorov-Smirnov type test for independence and multivariate samples. *Rev. Roumaine Math. Pures Appl.*, **26**(2), 213–226.

- Dembo, A. and Zeitouni, O. (1998). *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics*. Springer-Verlag, New York, second edition.
- Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York.
- Devroye, L., Györfi, L., and Lugosi, G. (2002). A note on robust hypothesis testing. *IEEE Trans. Inform. Theory*, **48**(7), 2111–2114.
- Dias, A. and Embrechts, P. (2004). Change-point analysis for dependence structures in finance and insurance. *Risk Measures of the 21st Century*, page 321335.
- Döring, M. (2010). Multiple change-point estimation with U -statistics. *J. Statist. Plann. Inference*, **140**(7), 2003–2017.
- Dudewicz, E. J. and van der Meulen, E. C. (1981). Entropy-based tests of uniformity. *J. Amer. Statist. Assoc.*, **76**(376), 967–974.
- Dugué, D. (1951). Analyticité et convexité des fonctions caractéristiques. *Ann. Inst. H. Poincaré*, **12**, 45–56.
- Dunford, N. and Schwartz, J. (1962). *Linear Operators*. Interscience Publishers.
- Eichelsbacher, P. and Schmock, U. (2002). Large deviations of U -empirical measures in strong topologies and applications. *Ann. Inst. H. Poincaré Probab. Statist.*, **38**(5), 779–797.
- Ekeland, I. and Témam, R. (1999). *Convex analysis and variational problems*, volume 28 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english edition. Translated from the French.
- El Rhabi, M., Fenniri, H., Keziou, A., and Moreau, E. (2013). A robust algorithm for convolutive blind source separation in presence of noise. *Signal Processing*, **93**(4), 818–827.
- Evans, L. and Gariepy, R. (1992). Measure theory and fine properties of function. *Studies in Advanced Mathematics*, CRC Press, Boca Raton, FL.
- Fenchel, W. (1949). On conjugate convex functions. *Canadian J. Math.*, **1**, 73–77.
- Feng, Z. and McCulloch, C. E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statist. Probab. Lett.*, **13**(4), 325–332.
- Ferguson, T. S. (1982). An inconsistent maximum likelihood estimate. *J. Amer. Statist. Assoc.*, **77**(380), 831–834.
- Fermanian, J.-D., Radulović, D., and Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli*, **10**(5), 847–860.
- Ferreira, D., Sá Ant6nio, M., Cerqueira, A., and Seixas, J. (2009). Ica-based method for quantifying eeg event-related desynchronization. *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*, **ICA’09**, 403–410.
- Fokianos, K. (2003). *Box-Cox Transformation for semiparametric Comparison of Two Samples*. In *Foundations of Statistical Inference* (eds H.R.L.Y. Haitovsky and Y. Ritov), pp.131–140. Haidelberg : Physica.
- Fokianos, K., Kedem, B., Qin, J., and Short, D. A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, **43**(1), 56–65.

- Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *N. Am. Actuar. J.*, **2**(1), 1–25.
- Fu, Y.-X. and Curnow, R. N. (1990). Maximum likelihood estimation of multiple change points. *Biometrika*, **77**(3), 563–573.
- Galambos, J. (1975). Order statistics of samples from multivariate distributions. *J. Amer. Statist. Assoc.*, **70**(351, part 1), 674–680.
- Gänssler, P. (1971). Compactness and sequential compactness in spaces of measures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **17**, 124–146.
- Genest, C. and MacKay, R. J. (1986). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canad. J. Statist.*, **14**(2), 145–159.
- Genest, C. and Verret, F. (2005). Locally most powerful rank tests of independence for copula models. *J. Nonparametr. Stat.*, **17**(5), 521–539.
- Genest, C. and Werker, B. J. M. (2002). Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. In *Distributions with given marginals and statistical modelling*, pages 103–112. Kluwer Acad. Publ., Dordrecht.
- Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, **82**(3), 543–552.
- Ghazdali, A. (2015). *Séparation aveugle de sources indépendantes/dépendantes par minimisation de critères entre copules*. Ph.D. thesis, Université Cadi-Ayyad, Marrakech, Maroc.
- Ghazdali, A., El Rhabi, M., Fenniri, H., Hakim, A., and Keziou, A. (2017). Blind noisy mixture separation for independent/dependent sources through a regularized criterion on copulas. *Signal Processing*, **131**, 502–513.
- Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, **31**, 1208–1211.
- Gombay, E. (1996). The weighted sequential likelihood ratio. *Canad. J. Statist.*, **24**(2), 229–239.
- Gombay, E. (1997). The likelihood ratio under noncontiguous alternatives. *Canad. J. Statist.*, **25**(3), 417–423.
- Gombay, E. and Horváth, L. (1994). An application of the maximum likelihood test to the change-point problem. *Stochastic Process. Appl.*, **50**(1), 161–171.
- Gombay, E. and Horváth, L. (1996a). Approximations for the time of change and the power function in change-point models. *J. Statist. Plann. Inference*, **52**(1), 43–66.
- Gombay, E. and Horváth, L. (1996b). On the rate of approximations for maximum likelihood tests in change-point models. *J. Multivariate Anal.*, **56**(1), 120–152.
- Groeneboom, P., Oosterhoff, J., and Ruymgaart, F. H. (1979). Large deviation theorems for empirical probability measures. *Ann. Probab.*, **7**(4), 553–586.
- Guégan, D. and Zhang, J. (2010). Change analysis of a dynamic copula for measuring dependence in multivariate financial data. *Quant. Finance*, **10**(4), 421–430.
- Gumbel, E. J. (1960). Bivariate exponential distributions. *J. Amer. Statist. Assoc.*, **55**, 698–707.
- Györfi, L. and Nemetz, T. (1978). f -dissimilarity: a generalization of the affinity of several distributions. *Ann. Inst. Statist. Math.*, **30**(1), 105–113.

- Györfi, L. and Vajda, I. (2002). Asymptotic distributions for goodness-of-fit statistics in a sequence of multinomial models. *Statist. Probab. Lett.*, **56**(1), 57–67.
- Györfi, L., Liese, F., Vajda, I., and van der Meulen, E. C. (1998). Distribution estimates consistent in χ^2 -divergence. *Statistics*, **32**(1), 31–57.
- Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Ann. Statist.*, **12**(3), 971–988.
- Hansen, L., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, **14**, 462–2800.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**(4), 1029–1054.
- He, H. and Severini, T. A. (2010). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, **16**(3), 759–779.
- Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.*, **37**(3), 1079–1111.
- Hollander, M. and Wolfe, D. A. (1999). *Nonparametric statistical methods*. John Wiley & Sons Inc., New York, second edition.
- Horváth, L. and Hušková, M. (2012). Change-point detection in panel data. *J. Time Series Anal.*, **33**(4), 631–648.
- Hosmer, Jr., D. W. and Lemeshow, S. (1999). *Applied survival analysis*. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons Inc., New York. Regression modeling of time to event data, A Wiley-Interscience Publication.
- Hosmer, Jr., D. W. and Lemeshow, S. (2000). *Applied logistic regression*. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons Inc., New York.
- Hüsler, J. and Reiss, R.-D. (1989). Maxima of normal random vectors: between independence and complete dependence. *Statist. Probab. Lett.*, **7**(4), 283–286.
- Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, **9**, 1483–1492.
- Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Rev. Econom. Stud.*, **64**(3), 359–383.
- Jiao, J., Venkat, K., Han, Y., and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inform. Theory*, **61**(5), 2835–2885.
- Jiménez, R. and Shao, Y. (2001). On robustness and efficiency of minimum divergence estimators. *Test*, **10**(2), 241–248.
- Joe, H. (1993). Parametric families of multivariate distributions with given margins. *J. Multivariate Anal.*, **46**(2), 262–282.
- Joe, H. (1997). *Multivariate models and dependence concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.
- Kanamori, T., Suzuki, T., and Sugiyama, M. (2012). f -divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Trans. Inform. Theory*, **58**(2), 708–720.

- Kay, R. and Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, **74**(3), 495–501.
- Keziou, A. (2003a). Dual representation of ϕ -divergences and applications. *C. R. Math. Acad. Sci. Paris*, **336**(10), 857–862.
- Keziou, A. (2003b). *Utilisation des divergences entre mesures en statistique inferentielle*. Ph.D. thesis, Université Pierre et Marie Curie-Paris VI, <https://tel.archives-ouvertes.fr/tel-00004069>.
- Keziou, A. (2015). Multivariate divergences with application in multisample density ratio models. In *Geometric Science of Information*, pages 444–453. Springer.
- Keziou, A. (2016). Multivariate divergences : Applications in inference for multiple samples under density ratio models. *Working paper*.
- Keziou, A. and Leoni-Aubin, S. (2005). Test of homogeneity in semiparametric two-sample density ratio models. *C. R. Math. Acad. Sci. Paris*, **340**(12), 905–910.
- Keziou, A. and Leoni-Aubin, S. (2008). On empirical likelihood for semiparametric two-sample density ratio models. *J. Statist. Plann. Inference*, **138**(4), 915–928.
- Keziou, A. and Regnault, P. (2015). Generalized mutual-information based independence tests. In *Geometric Science of Information*, pages 454–463. Springer.
- Keziou, A. and Regnault, P. (2016). Semiparametric estimation of mutual information and related criteria : Optimal test of independence. *To appear in IEEE Tans. Inform. Theory*, *arXiv:1508.04671*.
- Keziou, A., Ould Mohamed, M. S., Fenniri, H., and Delaunay, G. (2008). A new criterion for second order cyclostationary blind source separation. In *2nd International Conference on Signals, Circuits and Systems, 2008. SCS 2008.*, pages 1–6. IEEE.
- Keziou, A., Fenniri, H., Ould Mohamed, M. S., and Delaunay, G. (2009). Séparations aveugle de sources par minimisation des α -divergences. In *XXIIIe colloque GRETSI, Dijon (FRA), 8-11 septembre 2009*. GRETSI, Groupe d’Etudes du Traitement du Signal et des Images.
- Keziou, A., Fenniri, H., Ghazdali, A., and Moreau, E. (2014). New blind source separation method of independent/dependent sources. *Signal Processing*, **104**, 319–324.
- Khan, S., Bandyopadhyay, S., Ganguly, A. R., Saigal, S., Erickson, D. J., Protopopescu, V., and Ostrouchov, G. (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys. Rev. E*, **76**, 026209.
- Kimeldorf, G. and Sampson, A. (1975a). One-parameter families of bivariate distributions with fixed marginals. *Comm. Statist.*, **4**, 293–301.
- Kimeldorf, G. and Sampson, A. (1975b). Uniform representations of bivariate distributions. *Comm. Statist.*, **4**(7), 617–627.
- Kitamura, Y. (2001). Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica*, **69**(6), 1661–1672.
- Kitamura, Y. (2007). *Empirical likelihood methods in econometric theory and practice*. Cambridge University Press.
- Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Series in Statistics. Springer, New York.
- Kontoyiannis, I. and Skoularidou, M. (2015). Estimating the directed information and testing for causality. *CoRR*, **abs/1507.01234**.

- Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E*, **69**, 066138.
- Lehmann, E. L. (1986). *Testing statistical hypotheses*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, second edition.
- Léonard, C. (2001a). Convex conjugates of integral functionals. *Acta Math. Hungar.*, **93**(4), 253–280.
- Léonard, C. (2001b). Minimization of energy functionals applied to some inverse problems. *Appl. Math. Optim.*, **44**(3), 273–297.
- Léonard, C. (2001c). Minimizers of energy functionals. *Acta Math. Hungar.*, **93**(4), 281–325.
- Liese, F. (1977). On the existence of f -projections. In *Topics in information theory (Second Colloq., Keszthely, 1975)*, pages 431–446. Colloq. Math. Soc. János Bolyai, Vol. 16. North-Holland, Amsterdam.
- Liese, F. and Vajda, I. (1987). *Convex statistical distances*, volume 95 of *Teubner-Texte zur Mathematik [Teubner Texts in Mathematics]*. BSB B. G. Teubner Verlagsgesellschaft, Leipzig. With German, French and Russian summaries.
- Liese, F. and Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Trans. Inform. Theory*, **52**(10), 4394–4412.
- Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Statist.*, **22**(2), 1081–1114.
- Mansour, A. and Jutten, C. (1996). A direct solution for blind separation of sources. *IEEE Transactions on Signal Processing*, **44**(3), 746–748.
- McCullagh, P. and Nelder, J. A. (1983). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.
- McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative risk management*. Princeton Series in Finance. Princeton University Press, Princeton, NJ. Concepts, techniques and tools.
- Menéndez, M. L., Morales, D., Pardo, L., and Vajda, I. (1998). Asymptotic distributions of ϕ -divergences of hypothetical and observed frequencies on refined partitions. *Statist. Neerlandica*, **52**(1), 71–89.
- Menéndez, M. L., Pardo, J. A., Pardo, L., and Zografos, K. (2003). On tests of homogeneity based on minimum ϕ -divergence estimator with constraints. *Comput. Statist. Data Anal.*, **43**(2), 215–234.
- Moon, Y.-I., Rajagopalan, B., and Lall, U. (1995). Estimation of mutual information using kernel density estimators. *Phys. Rev. E*, **52**, 2318–2321.
- Morales, D. and Pardo, L. (2001). Some approximations to power functions of ϕ -divergences tests in parametric models. *Test*, **10**(2), 249–269.
- Morales, D., Pardo, L., and Vajda, I. (1995). Asymptotic divergence of estimates of discrete distributions. *J. Statist. Plann. Inference*, **48**(3), 347–369.
- Morales, D., Pardo, L., and Zografos, K. (1998). Informational distances and related statistics in mixed continuous and categorical variables. *J. Statist. Plann. Inference*, **75**(1), 47–63.
- Moreau, J. (1962). Fonctions convexes en dualité. *Séminaire de Mathématiques de la Faculté des Sciences de Montpellier*, **no. 1**.
- Nelsen, R. B. (2006). *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition.

- Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, **72**(1), 219–255.
- Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, **56**(11), 5847–5861.
- Nikitin, Y. (1995). *Asymptotic efficiency of nonparametric tests*. Cambridge University Press, Cambridge.
- Novey, M. and Adali, T. (2005). Ica by maximization of nongaussianity using complex functions. *Proc. MLSP*.
- Oakes, D. (1994). Multivariate survival distributions. *J. Nonparametr. Statist.*, **3**(3-4), 343–354.
- Omelka, M., Gijbels, I., and Veraverbeke, N. (2009). Improved kernel estimation of copulas: weak convergence and goodness-of-fit testing. *Ann. Statist.*, **37**(5B), 3023–3058.
- Ould Mohamed, M. S. (2010). *Contribution à la séparation aveugle de sources par utilisation des divergences entre densités de probabilité: application à l'analyse vibratoire*. Ph.D. thesis, Université de Reims Champagne-Ardenne.
- Ould Mohamed, M. S., Keziou, A., Fenniri, H., and Delaunay, G. (2010). Nouveau critère de séparation aveugle de sources cyclostationnaires au second ordre. *ARIMA Rev. Afr. Rech. Inform. Math. Appl.*, **12**, 1–14.
- Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**(1), 90–120.
- Owen, A. (1991). Empirical likelihood for linear models. *Ann. Statist.*, **19**(4), 1725–1747.
- Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**(2), 237–249.
- Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall, New York.
- Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41**, 100–115.
- Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527.
- Pardo, J. A., Pardo, L., and Zografos, K. (2002). Minimum ϕ -divergence estimators with constraints in multinomial populations. *J. Statist. Plann. Inference*, **104**(1), 221–237.
- Pardo, L. (2006). *Statistical inference based on divergence measures*, volume 185 of *Statistics: Textbooks and Monographs*. Chapman & Hall/CRC, Boca Raton, FL.
- Pesquet, J.-C. and Moreau, E. (2001). Cumulant-based independence measures for linear mixtures. *IEEE Trans. Inform. Theory*, **47**(5), 1947–1956.
- Pfanzagl, J. and Wefelmeyer, W. (1978). A third-order optimum property of the maximum likelihood estimator. *J. Multivariate Anal.*, **8**(1), 1–29.
- Pham, D. (2000). Blind separation of instantaneous mixture of sources based on order statistics. *IEEE Transactions on Signal Processing*, **48**(2), 363–375.
- Pham, D. T. (2002). Mutual information approach to blind separation of stationary sources. *IEEE Trans. Inform. Theory*, **48**(7), 1935–1946.
- Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, **85**(3), 619–630.

- Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**(1), 300–325.
- Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, **84**(3), 609–618.
- Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*. John Wiley & Sons, New York-Chichester-Brisbane. Wiley Series in Probability and Mathematical Statistics.
- Rockafellar, R. T. (1968). Integrals which are convex functionals. *Pacific J. Math.*, **24**, 525–539.
- Rockafellar, R. T. (1970). *Convex analysis*. Princeton University Press, Princeton, N.J.
- Rockafellar, R. T. (1974). *Conjugate duality and optimization*. Society for Industrial and Applied Mathematics, Philadelphia, Pa. Lectures given at the Johns Hopkins University, Baltimore, Md., June, 1973, Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 16.
- Rüschendorf, L. (1976). Asymptotic distributions of multivariate rank order statistics. *Ann. Statist.*, **4**(5), 912–923.
- Rüschendorf, L. (1984). On the minimum discrimination information theorem. *Statist. Decisions*, (suppl. 1), 263–283. Recent results in estimation theory and related topics.
- Rüschendorf, L. (1987). Projections of probability measures. *Statistics*, **18**(1), 123–129.
- Ruymgaart, F. H. (1974). Asymptotic normality of nonparametric tests for independence. *Ann. Statist.*, **2**, 892–910.
- Ruymgaart, F. H., Shorack, G. R., and van Zwet, W. R. (1972). Asymptotic normality of nonparametric tests for independence. *Ann. Math. Statist.*, **43**, 1122–1135.
- Sahmoudi, M., Snoussi, H., and Amin, M. G. (2006). Robust approach for blind source separation in non-gaussian noise environments. *Proceedings of ISCCSP, Marrakesh, Morocco, IEEE/EURASIP*.
- Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *Ann. Statist.*, **35**(2), 634–672.
- Schweizer, B. (1991). Thirty years of copulas. In *Advances in probability distributions with given marginals (Rome, 1990)*, volume 67 of *Math. Appl.*, pages 13–50. Kluwer Acad. Publ., Dordrecht.
- Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.*, **82**(398), 605–610.
- Shao, J. and Tu, D. S. (1995). *The jackknife and bootstrap*. Springer Series in Statistics. Springer-Verlag, New York.
- Sheehy, A. (1987). Kullback-Leibler constrained estimation of probability measures. *Report, Dept. Statistics, Stanford Univ.*
- Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, **51**(4), 1384–1399.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.
- Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8**, 229–231.

- Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika (Prague)*, **9**, 449–460.
- Smith, R. J. (1997). Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Economic Journal*, **107**, 503–519.
- Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Büna, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Ann. Inst. Statist. Math.*, **60**(4), 699–746.
- Sugiyama, M., Suzuki, T., and Kanamori, T. (2012a). *Density ratio estimation in machine learning*. Cambridge University Press, Cambridge. With a foreword by Thomas G. Dietterich.
- Sugiyama, M., Suzuki, T., and Kanamori, T. (2012b). Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Ann. Inst. Statist. Math.*, **64**(5), 1009–1044.
- Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation. *Proceedings of ECML-PKDD2008*, **4**, 5–20.
- Taleb, A. and Jutten, C. (1997). Entropy optimization. *Artificial Neural Networks - ICANN'97*, pages 529–534.
- Teboulle, M. and Vajda, I. (1993). Convergence of best ϕ -entropy estimates. *IEEE Trans. Inform. Theory*, **39**(1), 297–301.
- Titterton, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester.
- Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canad. J. Statist.*, **33**(3), 357–375.
- Tsybakov, A. B. and van der Meulen, E. C. (1996). Root- n consistent estimators of entropy for densities with unbounded support. *Scand. J. Statist.*, **23**(1), 75–83.
- Valiant, G. and Valiant, P. (2011a). Estimating the unseen: an $n/\log(n)$ -sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC'11—Proceedings of the 43rd ACM Symposium on Theory of Computing*, pages 685–694. ACM, New York.
- Valiant, G. and Valiant, P. (2011b). The power of linear estimators. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science—FOCS 2011*, pages 403–412. IEEE Computer Soc., Los Alamitos, CA.
- van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.
- Van Hulle, M. M. (2005). Edgeworth approximation of multivariate differential entropy. *Neural Computation*, **17**, 1903–1910.
- Vostrikova, L. J. (1983). Functional limit theorems for the “disorder” problem. *Stochastics*, **9**(1-2), 103–124.
- Vostrikova, L. Y. (1981). Detection of “discordance” of a Wiener process. *Teor. Veroyatnost. i Primenen.*, **26**(2), 362–368.
- Vrins, F., Pham, D.-T., and Verleysen, M. (2007). *Is the general form of Renyi's entropy a contrast for source separation?*, volume 4666/2007 of *Lecture Notes in Computer Science*. Springer, Berlin.
- Wang, Q., Kulkarni, S. R., and Verdú, S. (2005). Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Trans. Inform. Theory*, **51**(9), 3064–3074.

- Wang, W. and Ding, A. A. (2000). On assessing the association for bivariate current status data. *Biometrika*, **87**(4), 879–893.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**(1), 1–25.
- Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, **62**(6), 3702–3720.
- Zografos, K. (1994). Asymptotic distributions of estimated f -dissimilarity between populations in stratified random sampling. *Statist. Probab. Lett.*, **21**(2), 147–151.
- Zografos, K. (1998). f -dissimilarity of several distributions in testing statistical hypotheses. *Ann. Inst. Statist. Math.*, **50**(2), 295–310.
- Zografos, K., Ferentinos, K., and Papaioannou, T. (1990). ϕ -divergence statistics: sampling properties and multinomial goodness of fit and divergence tests. *Comm. Statist. Theory Methods*, **19**(5), 1785–1802.
- Zou, F., Fine, J. P., and Yandell, B. S. (2002). On empirical likelihood for a semiparametric mixture model. *Biometrika*, **89**(1), 61–75.