Estimation of parametric and semiparametric mixture models using phi-divergences

Diaa Al-Mohamad

École Doctorale de Science Mathématiques de Paris Centre

Thèse de Doctorat

Discipline: Mathématiques
Spécialité: Statistiques

présentée par
Diaa AL MOHAMAD

Esimation d'un Modèle de Mélange paramétrique et Semiparamétrique par des $\varphi-$divergences

---

dirigée par Pr. Michel BRONIATOWSKI

Soutenance le 17 novembre devant le jury composé de

Michel BRONIATOWSK	LSTA - Université Paris VI	Directeur de thèse
Stéphane CHRETIEN	National Physical Laboratory - UK	Rapporteur
Pierre VANDEKERKHOVE	LAMA - Université Marne-la-Vallée	Rapporteur
Laurent BORDES	LMAP - Université de Pau et des Pays de l'Adour	Examinateur
Catherine MATIAS	LPMA - Université Paris VI	Examinateur
Stéphane ROBIN	AgroParisTech / INRA	Examinateur

Remerciements

Arrivant à la fin de cette thèse, je suis reconnaissant pour certaines personnes qui avaient un rôle important sur les travaux que j'ai pu mener dans mon travail de thèse. Il y a certaines personnes sans qui je n'aurai pas eu la chance de revenir ici en France afin de reprendre mon travail de thèse après deux ans et demi d'interruption.
Je vaudrais remercier tout d'abord ma mère Alia qui m'a supporté pendant toute la durée de mon séjour en France et qui m'accorde sa confiance et son support à tout moment. Je vaudrais remercier mon père Hayel qui n'est plus avec nous dans ce monde depuis longtemps. Il m'encourageait toujours pour poursuivre mes études et il a consacré sa vie pour me faire agrandir. Je remercie mes parents qui m'ont appris comment apprécier le temps. Je vaudrais remercier mon épouse Tamader qui m'a accompagné pendant toute la durée de ma thèse et qui était patiente et a compris la pression que je subissais à cause de la courte durée de ma bourse d"etude (2 ans). Je lui remercie surtout parce qu'elle a supporté ma décision de quitter mon pays la Syrie pour revenir en France, et puis d'avoir accepté d'être avec moi ici en France sans qu'elle ait d'amie ni de famille. Il faudrait que je remercie ma fille Leen qui a réjouit ma vie avec ses beaux sourires. Je suis très reconnaissant pour mon beau-père Moafk et ma belle-mère Anam qui m'ont hébergé avec ma petite famille pendant la période (deux ans environs) que je préparais mes papiers pour revenir en France. Je leur remercie pour leur support qui m'accompagne tous les jours. Je vaudrais remercie également mes frères, Nouaf, Yasser et Fayez qui m'ont accordé leur confiance et m'ont aidé pour revenir en France. Je vaudrais remercier mes beaux-frères, Maaen, Mauthana, Omar et Mohamad que je considère comme mes frères et amis.

Pour l'avancement dans cette thèse, je remercie Michel Broniatowski qui a tout d'abord accepté de me prendre à nouveau après deux ans et demi d'interruption à cause de la stituation terrible dans mon pays. Ses idées, ses conseils ainsi que sa vision de mon travail ont été des facteurs importants pour achever et finir cette recherche. Je lui remercie pour plusieurs cafés auxquels il m'a invités. Je lui remercie d'avoir compris ma situation familiale particulière. Je lui remercie de la confiance qu'il m'a accordée pendant toute la durée de la thèse. Je remercie Alexis pour quelques courtes discussions sur les L-moments. Je remercie Assia qui m'a parlé des modèles de mélanges semiparamétriques et des questions qui y sont liées. Je remercie Matthias Kohl de l'Université de Furwagen en Allemagne pour des discussions sur l'intégration numérique. Je remercie Gilles Celeux, Jean-Patrick Baudrey et Olivier Schwander pour de vives discussions sur le modèle de mélange semiparamétrique durant un groupe de travail. Je remercie mes camarades du laboratoire, Dimby, Moukhtar, Matthiau et les autres pour plusieurs discussions scientifiques ou non. Je remercie les rapporteurs de ma thèse, Mr. Vandekerkhove et Mr. Chrétien, pour leurs commentaires et leurs avis très positif et encourageant.

Il y a plusieurs personnes à qui je n'aurais peut-être pas l'occasion de dire merci. Je

remercie Khaled Halawa mon ancien prof de mathématiques et qui est devenu un cher collègue en Syrie après. Je remercie Said Dsouki et Abd Allah Aboshahein qui ont convaincu l'administration de mon institut en Syrie de m'raccorder la bourse de la thèse après sa supension, Je remercie finalement Bassam Alzneika qui m'a aidé, pendant une période très difficile dans la région où je vivais en Syrie, à finir les papiers administratifs nécessaires pour l'obtention de la bourse.

MERCI A TOUS !

Diaa. October 12, 2016.

Summary

Résumé

L'étude des modèles de mélanges est un champ d'étude très vaste. D'autre part, les $\varphi-$divergences sont des outils statistiques qui de plus en plus attirent l'attention des statisticiens et les praticiens. Dans cette thèse, nous présentons et étudions quelques aspects et propriétés des $\varphi-$divergences et les estimateurs qui sont construits à base d'une $\varphi-$divergence. Nous employons ces estimateurs à l'estimation des modèles de mélanges. Dans une seconde partie de cette thèse, nous construisons et développons une nouvelle structure pour les modèles de mélanges semiparamétriques. L'estimation dans ce nouveau modèle est basée sur les $\varphi-$divergences qui offrent de bons outils pour le traitement de notre nouvelle approche.

Nous présentons dans la première partie de cette thèse les $\varphi-$divergences, et nous en faisons un rappel des principales propriétés. Nous présentons les méthodes existantes dont l'objectif est de produire des estimateurs pour des modèles paramétriques basés sur des $\varphi-$divergences. Nous nous intéressons à l'étude de la méthode de Beran, de l'approche de Basu-Lindsay et de la forme dual des $\varphi-$divergences. Nous nous intéressons en particulier à la dernière approche. Nous montrons que les estimateurs basés sur la forme duale des $\varphi-$divergences dans un contexte paramétrique ne sont pas robustes. Ceci est exploré théoriquement et expérimentalement avec des simulations numériques. Nous proposons ensuite une modification qui rend ces estimateurs robustes. Le nouvel estimateur est alors comparé avec les autres méthodes existantes qui produisent des estimateurs robustes à base des $\varphi-$divergences. La comparaison est également menée par rapport à un estimateur considéré comme $\ll$très performant$\gg$ pour l'estimation dans des modèles paramétriques; le minimum density power divergence. Notre nouvel estimateur montre de bonnes propriétés par rapport aux autres estimateurs en compétition.
Dans un second temps, nous présentons un algorithme d'optimisation dont l'objectif est de calculer les estimateurs à base de divergences. Notre algorithme est un algorithme proximal qui perturbe la fonction objective à chaque itération par une autre fonction convenablement choisie. La convergence des séquences générées par l'algorithme est étudiée. Nous montrons que les points limites des séquences générées par l'algorithme sont des points stationnaires de la fonction objective. D'autres propriétés de convergence globale de la séquence vers un optimum local de la fonction objective sont explorées mais en imposant des hypothèses plus restrictives. Nous étudions la convergence de la séquence générée par l'algorithme proximal dans plusieurs exemples. La convergence de l'algorithme EM est étudiée à nouveau sur quelques exemples mais dans l'esprit de notre approche.

Dans la deuxième partie de cette thèse, nous construisons une nouvelle structure pour les modèles de mélanges semiparamétriques à deux composantes dont l'une est incon-

nue. La nouvelle approche permet l'incorporation d'une information a priori linéaire sur la composante inconnue; par exemple des contraintes de moments ou de L-moments. Nous développons deux approches pour l'estimation de ce modèle; une approche pour les contraintes de type moments et une approche pour les contraintes de L-moments. Les propriétés asymptotiques des estimateurs résultant sont étudiées et prouvées sous des hypothèses standards. Nous illustrons par des simulations numériques les avantages de la nouvelle approche et de l'incorporation d'une information a priori par rapport aux méthodes existantes d'estimation d'un modèle semiparamétrique sans aucune information préliminaire à part une hypothèse de symétrie.

Mots clés: Modèle de mélange, $\varphi-$divergence, estimateur à noyau symétrique et asymétrique, algorithme proximal, dualité de Fenche-Legendre, modèle semiparamétrique, modèle semiparamétrique de quantile, L-moments.

Abstract

The study of mixture models constitutes a large domain of research. On the other hand, $\varphi-$divergences attract more and more the attention of statisticians and practitioners. In this work, we show some of the aspects and properties of $\varphi-$divergence-based estimators. We employ these estimators in mixture models. We build and develop in a second part a new structure for semiparametric mixture models and estimate the new model using $\varphi-$divergences efficiently.

In the first part of this work, we present $\varphi-$divergences and recall some of their basic properties. We present some of the existing methods in the literature which produce parametric estimation using $\varphi-$divergences; namely Beran's approach, the Basu-Lindsay approach and the dual formula of $\varphi-$divergences. We are interested in the later method. We show that the estimator based on the existing dual formula of $\varphi-$divergences in the parametric settings is not robust against outliers in several models. The problem is explored theoretically and experimentally. We propose a modification to this estimation method in order to robustify it. The new estimator is then compared to existing methods based on $\varphi-$divergences. The comparison is also done with respect to a powerful estimator in parametric estimation; the minimum density power divergence estimator. Our new estimator shows encouraging performances.
We present after that an optimization algorithm in order to calculate estimators based on a divergence criterion. The algorithm is a proximal-point algorithm which optimizes a modified version of the objective function by adding a suitable regularization term. Convergence properties of the presented algorithm are studied. We prove the convergence of the limiting points of the sequence generated by the algorithm to stationary points of the objective function. More properties are explored but with further assumptions in order to prove the convergence of the whole sequence towards a local optimum of the objective. Several examples are discussed, and another proof of convergence of the EM algorithm is given in several mixtures in the light of our approach.

In a second part of this work, we construct a new structure for semiparametric two-component mixture models where one component is unknown. The new structure permits to incorporate some prior linear information about the unknown component such as moments or L-moments constraints. Two different approaches are developed using $\varphi-$divergences in order to estimate the semiparametric mixture model; an approach for moment-type constraints and an approach for L-moments constraints. The asymptotic properties of the resulting estimators are studied and proved under standard conditions. The new structure is demonstrated by simulations to produce better estimates using the prior information than existing methods in the literature which do not consider in general any prior information except for a symmetric assumption.

Keywords: Mixture model, $\varphi-$divergence, symmetric and asymmetric kernel density estimator, proximal-point algorithm, Fenchel-Legendre duality, semiparametric model, semiparametric linear quantile model, L-moment.

Contents

List of Figures

List of Tables

Introduction

Le but de cette thèse est l'étude et l'estimation d'un modèle de mélanges de lois de la forme:

$$P(.|\phi) = \sum_{i=1}^{j} \lambda_i P_i(.|\theta_i), \qquad \text{t.q.} \quad \sum_{i=1}^{j} \lambda_i = 1.$$

Dans la première partie de la thèse (Chap.1 et 2), le modèle de mélange est paramétrique. C'est à dire que la distribution de chaque composante, indexée par $i = 1, \cdots, j$, correspond a une loi connue paramétrée par θ_i. Dans la deuxième partie de la thèse (Chap.3 et 4), nous aborderons le cas particulier où $j = 2$ en supposant que la première composante du mélange est paramétrique alors que la deuxième est nonparamétrique. Nous proposons ensuite une structure où la deuxième composante est semiparamétrique au sens où elle appartient à une famille de lois définies par des contraintes linéaires, par exemple l'ensemble des lois de probabilité ayant une variance égale au carré de l'espérance. Les modèles de mélanges paramétriques ont des applications diverses en biologie, en machine learning etc., voir Titterington et al. [1985] ou McLachlan and Peel [2005] pour plus de détails. Les modèles de mélanges semiparamétriques ont été employés dans différents contextes en génétique (Ma et al. [2011]), biologie (Bordes et al. [2006]) en machine learning (Song et al. [2010]) pour des algorithmes de clustering, etc. Le modèle semiparamétrique pourrait être appliqué dans d'autres situations et sur plus d'applications comme en traitement du signal.

L'estimation d'un modèle de mélange paramétrique se fait en général avec l'algorithme EM de Dempster et al. [1977]. L'algorithme EM offre une procédure facile à programmer et dont la complexité est faible. En effet, l'algorithme EM maximise la log-vraisemblance du modèle de mélange itérativement. A chaque itération, nous maximisons la vraisemblance à l'intérieur de chaque classe (composante) en attribuant des poids $h_{i,k}$ à chaque observation (numéro i) mesurant son appartenance à la classe k. Cependant, l'algorithme EM produit des estimateurs non-robustes parce que nous calculons le maximum de vraisemblance en fin de compte. Le maximum de vraisemblance est un estimateur qui est connu d'être sensible aux points aberrants *(outliers)* et au fait que le modèle ne contient pas la vraie distribution des données *(misspecification)*. L'objectif de la première partie est d'appliquer un autre outil d'estimation qui produit des estimateurs robustes. Nous formulons également un algorithme qui ressemble à l'algorithme EM au sens où l'optimisation n'est pas menée sur tous les paramètres en même temps, mais sur les proportions dans une étape et sur les paramètres décrivant les classes dans une autre étape.
L'estimation d'un modèle de mélange semiparamétrique à deux composantes est un sujet très récent. Plusieurs méthodes existent pour estimer la proportion et/ou les paramètres de la composante paramétrique sans qu'il y ait de contraintes sur la composante inconnue. Une hypothèse de symmétrie sur la composante inconnue a été employée afin de mieux estimer le modèle de mélange; voir Bordes and Vandekerkhove [2010] et Maiboroda and Sugakova [2012]. L'estimation d'un modèle de mélange semiparamétrique sans qu'il y ait

de contraintes sur la composante inconnue est difficile surtout lorsque nous avons à estimer des paramètres inconnus de la composante paramétrique. Nous proposons une méthode pour estimer un modèle de mélange semiparamétrique lorsque la composante inconnue est définie par des contraintes linéaires de type moments ou L-moments. Nous étudions les propriétés asymptotiques des estimateurs obtenus. Plusieurs simulations numériques sont présentées afin d'illustrer l'avantage de la nouvelle approche.

0.1 Première partie: Estimation robuste basée sur des $\varphi-$divergences avec application aux modèles de mélanges paramétriques

0.1.1 Chapitre 1: Estimation basée sur des $\varphi-$divergences

Les $\varphi-$divergences sont des mesures de distance ou dissimilarité entre des distributions de probabilité ou plus généralement entre des measures $\sigma-$finies. Elles ont été introduites indépendamment par Csiszár [1963] et Ali and Silvey [1966]. Pour deux mesures P et Q $\sigma-$finies telles que Q est absolument continue par rapport à P, nous définissons la $\varphi-$divergence entre Q et P par:

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

où φ est une fonction positive convexe telle que $\varphi(1) = 0$. Si φ est strictement convexe alors:

$$D_\varphi(Q,P) = 0 \quad \text{si et seulement si} \quad P = Q.$$

L'estimation par une $\varphi-$divergence se fait en minimisant celle-ci entre une famille de lois et une mesure de probabilité P_T inconnue. La loi P_T n'est connue en générale que par un échantillon $Y_1, \cdots, Y_n$ observée. La famille de lois est un modèle P_ϕ avec $\phi \in \Phi \subset \mathbb{R}^d$ paramétré par le paramètre ϕ. Le but est de trouver le meilleur vecteur de paramètres ϕ^T tel que P_{ϕ^T} soit le plus proche possible de P_T d'un point de vue d'une $\varphi-$divergence. En particulier, si P_T est un membre du modèle $(P_\phi)_\phi$, alors il existe ϕ^T tel que $P_T = P_{\phi^T}$, et

$$\phi^T = \arg\min_{\phi\in\Phi} D_\varphi(P_\phi, P_T).$$

Comme P_T est inconnue, nous avons besoin de la remplacer par un estimateur afin d'estimer ϕ^T. Dans le cas où les mesures de probabilité sont définies sur des espaces discrets, P_T est remplacée par la mesure empirique, voir Lindsay [1994]. Dans le cas des modèles continus, remplacer P_T par sa version empirique n'est pas convenable, car le modèle ne serait pas absolument continu par rapport à la mesure empirique pour n'importe quel n, et aucune procédure d'estimation ne pourrait être produite, voir Broniatowski and Vajda [2012]. Plusieurs approches ont été proposées afin d'approximer la $\varphi-$divergence lorsque le modèle est continu.

- L'approche de Beran [1977]. Cette approche consiste à remplacer directement P_T par un estimateur à noyau. L'approche de Beran [1977] a été proposée dans le cas de la divergence de Hellinger. Cette approche a été plus tard généralisée à la classe des $\varphi-$divergences par Park and Basu [2004] et Kuchibhotla and Basu [2015].

- L'approche de Basu and Lindsay [1994]. Cette approche consiste à remplacer P_T par un estimateur à noyau et convoler le modèle avec le même noyau afin de réduire le

rôle de la fenêtre. Les auteurs démontrent que sous une certaine condition (difficile) sur le noyau (noyau transparent), l'estimateur basé sur leur approche est consistant sans avoir besoin que l'estimateur à noyau soit consistant.

- La forme dual des $\varphi-$divergences. Cette approche a été développée indépendamment par Broniatowski and Keziou [2006] et Liese and Vajda [2006]. On peut montrer que pour trois densités de probabilités p_α, p_ϕ et p_T, nous avons:

$$D_\varphi(p_\phi, p_T) \geq \int \varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x)p_\phi(x)dx - \int \varphi^\#\left(\frac{p_\phi}{p_\alpha}\right)(y)p_T(y)dy,$$

où $\varphi^\#(t) = t\varphi'(t) - \varphi(t)$, et que l'égalité est atteinte lorsque $p_\alpha = p_T$. Alors:

$$D_\varphi(p_\phi, p_T) = \sup_{\alpha\in\Phi}\left\{\int \varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x)p_\phi(x)dx - \int \varphi^\#\left(\frac{p_\phi}{p_\alpha}\right)(y)p_T(y)dy\right\}. \quad (0.1.1)$$

Il suffit maintenant de remplacer $p_T(y)dy$ par la mesure empirique dP_n afin d'avoir un estimateur dit *dual* de la $\varphi-$divergence. L'intérêt de cette approche est que, contrairement aux autres approches, nous n'avons pas besoin d'un estimateur à noyau, et donc nous n'avons pas à chercher un *bon* noyau et une *bonne* fenêtre.

Dans le premier chapitre, nous nous intéressons à la forme duale des $\varphi-$divergences. Broniatowski and Keziou [2009b] et Liese and Vajda [2006] proposent le minimum dual $\varphi-$divergence estimateur (MDφDE) de ϕ^T:

$$\hat{\phi} = \arg\inf_{\phi\in\Phi}\sup_{\alpha\in\Phi}\left\{\int \varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x)p_\phi(x)dx - \frac{1}{n}\sum_{i=1}^{n}\varphi^\#\left(\frac{p_\phi}{p_\alpha}\right)(y_i)\right\}.$$

Lorsque P_T appartient au modèle P_ϕ, la forme duale estime bien la $\varphi-$divergence parce que le supremum sur α sera atteint à $\alpha = \phi^T$. Cependant, si P_T n'est pas dans le modèle, ce n'est plus le cas. Par exemple, le cas des données contenant des outliers. En effet, si P_T n'est pas dans le modèle, alors nous aurons:

$$D_\varphi(p_\phi, p_T) \geq \sup_{\alpha\in\Phi}\left\{\int \varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x)p_\phi(x)dx - \int \varphi^\#\left(\frac{p_\phi}{p_\alpha}\right)(y)p_T(y)dy\right\}.$$

Nous illustrons à la figure (1) un exemple simple qui montre l'impact de ce problème. Nous illustrons également la solution que nous allons proposer ensuite. Notre solution consiste à utiliser un estimateur à noyau au lieu de p_α. Ceci permet de s'adapter à p_T que ce soit sous le modèle ou non, et en même temps permet de se débarasser de la forme suprémale. Notre nouvel estimateur est défini par:

$$D_\varphi(p_\phi, p_T) = \sup_{w>0}\left\{\int \varphi'\left(\frac{p_\phi}{K_{n,w}}\right)(x)p_\phi(x)dx - \int \varphi^\#\left(\frac{p_\phi}{K_{n,w}}\right)(y)p_T(y)dy\right\}. \quad (0.1.2)$$

En effet, un "bon" choix de la fenêtre permet d'introduire le nouvel estimateur:

$$\hat{\phi}_n = \arg\inf_{\phi\in\Phi}\int \varphi'\left(\frac{p_\phi}{K_{n,w_{\text{opt}}}}\right)(x)p_\phi(x)dx - \frac{1}{n}\sum_{i=1}^{n}\varphi^\#\left(\frac{p_\phi}{K_{n,w_{\text{opt}}}}\right)(y_i).$$

Nous démontrons que ce nouvel estimateur est consistant et asymptotiquement Gaussian sous des hypothèses standards. Les simulations numériques montrent que le nouvel estimateur performe mieux que les autres estimateurs présentés dans ce chapitre. Nous comparons la performance de cet estimateur avec le minimum density power divergence (MDPD) de Basu et al. [1998] qui est un estimateur de Bregman. Notre estimateur performe aussi bon que le MDPD dans plusieurs simulations et performe mieux dans un modèle à queue lourde qui est le GPD.

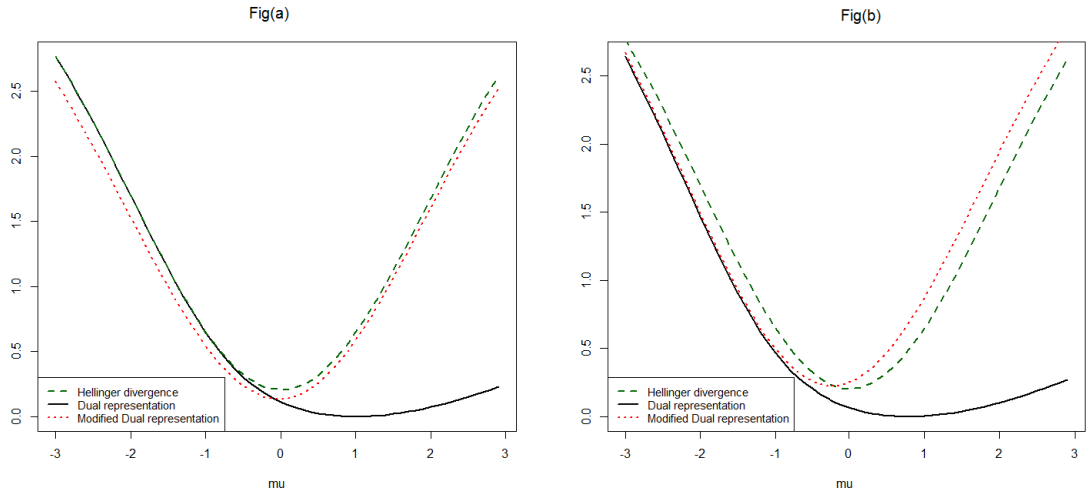

Figure 1: Sous-estimation causée par la forme duale classique en comparaison avec notre alternative. La vraie distribution P_T est $0.9\mathcal{N}(\mu = 0, \sigma = 1)+0.1\mathcal{N}(\mu = 10, \sigma = 2)$. Figure (a) montre la forme duale classique (0.1.1) en comparaison avec la nouvelle formulation duale définie par (0.1.2). Figure (b) montre les approximations correspondantes après avoir remplacé la vraie distribution par sa version empirique.

0.1.2 Chapitre 2:

Les procédures d'estimation présentées dans le chapitre précédent sont en général non-convexes. Les méthodes d'optimisation convexe ne garantissent que la convergence vers un optimum local de la fonction objective si celle-ci n'est pas convexe. Nous introduisons dans ce chapitre un algorithme d'optimisation proximale. Un algorithme proximal est un algorithme itératif qui à chaque itération optimise une version régularisée de la fonction objective. Les algorithmes proximaux ont été prouvés de produire de meilleurs résultats que les algorithmes d'optimisation classiques, voir Goldstein and Russak [1987].
Soient $(X_1, Y_1), \cdots, (X_n, Y_n)$ un échantillon de couples de variables aléatoires i.i.d. distribuées selon la densité $f(x, y|\phi^T)$ pour un $\phi^T \in \Phi$. Soient $(x_1, y_1), \cdots, (x_n, y_n)$ des réalisations de ces couples. Les données $y_1, \cdots, y_n$ sont les données observées et les données $x_1, \cdots, x_n$ sont les données inobservées ou les étiquettes. Par exemple, les données $x_1, \cdots, x_n$ sont les classes correspondant aux points $y_1, \cdots, y_n$.
L'algorithme EM est une procédure itérative qui estime le vecteur de paramètres ϕ^T en maximisant l'espérance de la log-vraisemblance complétée sachant les données observées, c.à.d.

$$\begin{aligned} \phi^{k+1} &= \arg\max_{\Phi} Q(\phi, \phi^k) \\ &= \arg\max_{\Phi} \mathbb{E}\left[\log(f(\mathbf{X}, \mathbf{Y}|\phi)) \,\middle|\, \mathbf{Y} = \mathbf{y}, \phi^k\right], \end{aligned}$$

où $\mathbf{X} = (X_1, \cdots, X_n)$, $\mathbf{Y} = (Y_1, \cdots, Y_n)$ et $\mathbf{y} = (y_1, \cdots, y_n)$. On peut démontrer que ces itérations s'écrivent de la façon suivante:

$$\phi^{k+1} = \arg\max_{\Phi} \sum_{i=1}^{n} \log\left(p_\phi(y_i)\right) + \sum_{i=1}^{n} \int_{\mathcal{X}} \log\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k) dx. \tag{0.1.3}$$

où $h_i(x|\phi^k) = \frac{f(x,y_i|\phi^k)}{p_{\phi^k}(y_i)}$ est la densité conditionnelle des étiquettes sachant une donnée y_i, et p_ϕ est la loi marginale des données observées. L'algorithme EM (0.1.3) s'écrit comme un algorithme proximal, car nous sommes en train de maximiser la log-vraisemblance en la perturbant à chaque itération de l'algorithme par une fonction positive qui ressemble à une distance de Kullback-Lebiler mais entre les densités conditionnelles des étiquettes.
Tseng [2004] propose de généraliser (0.1.3) en permettant au terme proximal à prendre d'autres formes plus générales guidées par une fonction génératrice ψ. Tseng propose l'algorithme suivant:

$$\phi^{k+1} = \arg\sup_{\phi} J(\phi) - D_\psi(\phi, \phi^k), \tag{0.1.4}$$

où $J(\phi)$ est la log-vraisemblance et

$$D_\psi(\phi, \phi^k) = \sum_{i=1}^{n} \int_{\mathcal{X}} \psi\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k) dx. \tag{0.1.5}$$

La fonction ψ est prise comme une fonction convexe positive qui vérifie $\psi(1) = \psi'(1) = 0$. Pour l'algorithme (0.1.3), $\psi(t) = -\log(t) + t - 1$.
L'algorithme EM ainsi que la généralisation faite par Tseng ont une objective de maximiser la log-vraisemblance. Par conséquent, les estimateurs issus de ces algorithmes ne sont pas robustes contre les outliers ou une perturbation du modèle autour de la vraie distribution des données. Pour cela, nous proposons de généraliser l'algorithme de Tseng en utilisant le lien entre la maximisation de la log-vraisemblance et la minimisation de la distance de Kullback-Leibler entre la mesure empirique et le modèle dans les modèles discrets. Bien évidemment, pour les modèles continus, ce lien est atteint de manière différente, car la distance entre le modèle et la mesure empirique n'est pas bien définie[1]. Nous proposons de remplacer la log-vraisemblance par un estimateur d'une $\varphi-$divergence.

$$\phi^{k+1} = \arg\inf_{\phi} \hat{D}_\varphi(p_\phi, p_T) + \frac{1}{n} D_\psi(\phi, \phi^k). \tag{0.1.6}$$

En prenant $\hat{D}_\varphi(p_\phi, p_T)$ l'estimateur induit par la forme duale (0.1.1) après avoir replacé $p_T(y)dy$ par dP_n, et pour $\varphi(t) = -\log(t) + t - 1$, nous avons:

$$\begin{aligned}
\phi^{k+1} &= \arg\inf_{\phi} \left\{ \sup_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \log(p_\alpha(y_i)) - \frac{1}{n} \sum_{i=1}^{n} \log(p_\phi(y_i)) + \frac{1}{n} D_\psi(\phi, \phi^k) \right\} \\
&= \arg\inf_{\phi} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(p_\phi(y_i)) + \frac{1}{n} D_\psi(\phi, \phi^k) \right\} \\
&= \arg\sup_{\phi} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log(p_\phi(y_i)) - \frac{1}{n} D_\psi(\phi, \phi^k) \right\} \\
&= \arg\sup_{\phi} J(\phi) - \frac{1}{n} D_\psi(\phi, \phi^k).
\end{aligned}$$

Donc, notre algorithme contient la généralisation de Tseng. De plus, pour $\psi(t) = -\log(t) + t - 1$, on se trouve avec l'algorithme EM. Nous proposons également dans le cas d'un modèle de mélange

$$p_\phi(y) = \sum_{i=1}^{s} \lambda_i p_i(y|\theta_i)$$

[1] Dans le monde des $\varphi-$divergences, cette distance est considérée infinie.

un algorithme proximal à deux niveaux; une sous-étape qui optimise sur les proportions λ_i et une sous-étape qui optimise sur les paramètres décrivant les classes θ_i. En d'autres termes:

$$\lambda^{k+1} = \underset{\lambda\in[0,1]^s, s.t.(\lambda,\theta^k)\in\Phi}{\arg\inf} \hat{D}_\varphi(p_{\lambda,\theta^k}, p_{\phi^*}) + D_\psi((\lambda,\theta^k),\phi^k); \tag{0.1.7}$$

$$\theta^{k+1} = \underset{\theta\in\Theta, s.t.(\lambda^{k+1},\theta)\in\Phi}{\arg\inf} \hat{D}_\varphi(p_{\lambda^{k+1},\theta}, p_{\phi^*}) + D_\psi((\lambda^{k+1},\theta),\phi^k). \tag{0.1.8}$$

Nous démontrons sous des hypothèses standards que les séquences $(\phi^k)_k$ générées par l'un des algorithmes (0.1.6) ou (0.1.7, 0.1.8) convergent vers un point stationnaire de l'estimateur de la divergence $\phi \mapsto \hat{D}_\varphi(p_\phi, p_T)$.
Définissons l'ensemble Φ^0 par:

$$\Phi^0 = \{\phi \in \Phi : \hat{D}_\varphi(p_\phi, p_T) \leq \hat{D}_\varphi(p_{\phi^0}, p_T)\}. \tag{0.1.9}$$

Nous citons ici les deux principaux résultats théoriques concernant la convergence de la séquence Φ^k générée par les algorithmes (0.1.6) ou (0.1.7, 0.1.8).

Proposition 0.1.1. *Supposons que les séquences (0.1.6) et (0.1.7, 0.1.8) sont bien définies dans Φ. Pour les deux algorithmes, la séquence $(\phi^k)_k$ vérifie les propriétés suivantes:*

(a) $D_\varphi(p_{\phi^{k+1}}|p_T) \leq D_\varphi(p_{\phi^k}|p_T)$;

(b) $\forall k, \phi^k \in \Phi^0$;

(c) Supposons que les fonctions $\phi \mapsto \hat{D}_\varphi(p_\phi|p_T), D_\psi$ sont semicontinues inférieurement et que l'ensemble Φ^0 est compact, alors la séquence $(\phi^k)_k$ est bien définie et bornée. De plus, la séquence $\left(\hat{D}_\varphi(p_{\phi^k}|p_T)\right)_k$ converge.

Cette proposition annonce une propriété essentielle de l'algorithme. Sous des conditions simples, nous avons que la fonction objective (l'estimateur de la divergence) converge le long de la séquence ϕ^k, voir figure (2). Cette propriété peut être utilisée comme un critère d'arrêt de l'algorithme au cas où la séquence de vecteurs ϕ^k ne converge pas. Un deuxième résultat principal dans ce travail prouve la convergence des sous-suites vers un point stationnaire de la divergence estimée. Ce résultat est nouveau, car les résultats existants supposent que le terme proximal D_ψ vérifie une hypothèse d'identifiabilité $D_\psi(\phi,\phi') = 0$ ssi $\phi = \phi'$. Dans ce résultat, nous ne demandons pas cette hypothèse.

Proposition 0.1.2. *Supposons que*

1. *les fonctions $\phi \mapsto \hat{D}_\varphi(p_\phi|p_T), D_\psi$ et $\nabla_1 D_\psi$ sont définies et continues sur, respectivement, $\Phi, \Phi\times\Phi$ et $\Phi\times\Phi$;*
2. *$\nabla\hat{D}_\varphi(p_\phi|p_T)$ est définie et continue sur Φ;*
3. *Φ^0 est compact,*

alors pour l'algorithme défini par (0.1.6), toute sous-suite convergente converge vers un point stationnaire de la fonction objective $\phi \to \hat{D}(p_\phi, p_T)$. De plus, si l'hypothèse 2 n'est pas vérifiée, alors 0 appartient au sous-gradient de $\phi \mapsto \hat{D}(p_\phi, p_T)$ calculé au point limite. En d'autres termes, les sous-suites convergentes convergent vers des points stationnaires généralisés.

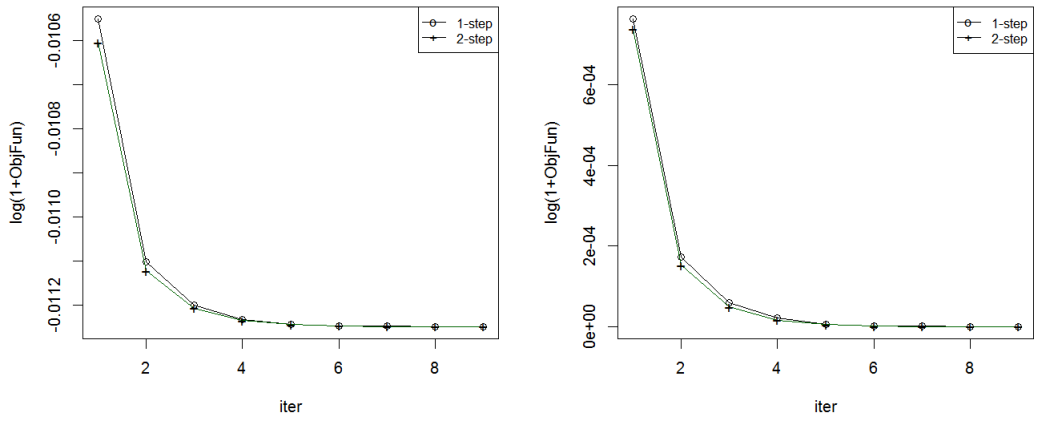

Figure 2: Décroissance de la (estimateur de la) distance de Hellinger entre la vraie distribution des données et le modèle estimé à chaque itération de l'algorithme proximal dans le cas d'un modèle de mélange à 2 composantes Gaussiennes. La figure de gauche illustre la courbe des valeurs pour le nouvel estimateur dual (0.1.2) estimé. La figure de droite illustre la courbe des valeurs de l'estimateur dual classique (0.1.1) estimé. Les valeurs sont représentées sur une échelle logarithmique $\log(1+x)$. Le 1-step représente l'algorithme (0.1.6), et le 2-step représente l'algorithme (0.1.7, 0.1.8)

Example 0.1.1. Dans un mélange de deux composantes Gaussiennes:

$$p_{\lambda,\mu}(x) = \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_1)^2} + \frac{1-\lambda}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_2)^2},$$

on peut démontrer avec notre approche que les points limites générés par l'algorithme EM sont des points stationnaires de la log-vraisemblance dès que le vecteur initial ϕ^0 vérifie la condition suivante:

$$J(\phi^0) > \max\left[J\left(0, \infty, \frac{1}{n}\sum_{i=1}^{n} y_i\right),\ J\left(1, \frac{1}{n}\sum_{i=1}^{n} y_i, \infty\right)\right],$$

où J est la log-vraisemblance. D'autres exemples et discussions sont présentés dans le chapitre ci-dessous. Les simulations numériques menées sur des mélanges de Gaussiens et des mélanges de Weibull montrent que l'algorithme proximal marche, et nous arrivons à calculer les estimateurs basés sur les formes duales présentées dans le chapitre précédent. Une application de notre algorithme proximal sur le density power divergence de Basu et al. [1998] est également présentée.

0.2 Deuxième partie: Modèles de mélanges semiparamétriques à deux composantes dont l'une est inconnue

Un modèle de mélange semiparamétrique à deux composantes est définie par:

$$f(x) = \lambda f_1(x|\theta) + (1-\lambda) f_0(x), \qquad \text{for } x \in \mathbb{R}^r, \tag{0.2.1}$$

où λ et θ sont deux paramètres à estimer. f_0 est considérée inconnue durant l'estimation.

0.2.1 Chapitre 3: Modèles de mélanges semiparamétriques à deux composantes dont l'une est définie par des contraintes linéaires sur sa distribution

Le modèle de mélange semiparamétrique (0.2.1) a été étudié et employé dans certaines applications récemment par plusieurs auteurs, Bordes et al. [2006], Robin et al. [2007], Song et al. [2010], Ma et al. [2011] et Xiang et al. [2014]. Dans ces papiers, le modèle n'a pas été traité avec la définition générale donnée ci-dessus. Certains auteurs ont considéré que θ est connu de manière à ce que la composante paramétrique soit entièrement connue. D'autres auteurs ont considéré une distribution précise comme le Gaussien pour f_1 sans aucune généralisation à d'autres familles de distributions.
Plusieurs méthodes d'estimation du modèle de mélanges semiparamétrique (0.2.1) ont été introduites. Bordes et al. [2006] proposent d'étudier le modèle (0.2.1) lorsque $r = 1$, f_0 est symétrique par rapport à une valeur inconnue μ_0 et f_1 est entièrement connue, c.à.d. θ est connu. Song et al. [2010] proposent d'étudier le modèle (0.2.1) lorsque f_1 est un Gaussien centré en 0 avec une variance inconnue. Ils supposent en plus que $f_0(0) = 0$. D'autres auteurs ont proposé des méthodes de type EM, voir Robin et al. [2007], Song et al. [2010] et Ma et al. [2011]. Une méthode basée sur la distance de Hellinger a été proposée par Xiang et al. [2014], mais nous n'en parlons pas, car l'algorithme présenté dans leur article n'est pas claire et contient des calculs intégrales qui ne se font pas avec des méthodes numériques.
Ces méthodes d'estimation ne sont pas basées sur une théorie solide exceptée la méthode de Bordes et al. [2006]. De plus, le comportement asymptotique des algorithmes proposés n'a pas été étudié. La convergence des méthodes itératives de type EM n'a pas été établie non plus. La méthode de Bordes et al. [2006] exploite la structure de symétrie imposée sur f_0 pour construire un estimateur consistant et asymptotiquement Gaussien, voir Bordes and Vandekerkhove [2010].
L'article de Xiang et al. [2014] illustre une comparaison entre plusieurs méthodes d'estimation pour le modèle semiparamétrique (0.2.1) lorsque θ est connu. Les méthodes donnent de bonnes performances sans qu'il y ait une méthode gagnante. Les données ont été générées par des mélanges de deux Gaussiens. Nous avons refait des simulations similaires mais en considérant θ inconnu. Les résultats n'ont pas été satisfaisants. La méthode de Bordes and Vandekerkhove [2010] a tendance à donner de bonnes résultats pour une proportion basse de la partie paramétrique, mais non pas très proche de zéro. Les autres méthodes ont tendance à donner de bonnes performances lorsque la proportion de la partie paramétrique est élevée. Nous croyons que le problème vient du degré de difficulté du modèle de mélange semiparamétrique. L'ajout d'une information a priori comme la symétrie de f_0 a permis d'améliorer l'estimation et de mieux étudier la théorie liée à la méthode.
Suivant l'idée de Bordes et al. [2006], nous proposons d'ajouter une information a priori relativement générale de manière à ce que nous puissions bien estimer le modèle de mélange semiparamétrique. Nous proposons d'ajouter une information linéaire comme les contraintes de moments. Les contraintes linéaires peuvent être traitées avec des outils d'analyse convexe. Nous définissons ainsi notre modèle de mélange semiparamétrique sous des contraintes linéaires sur la composante inconnue par:

$$\begin{aligned} P(.|\phi) &= \lambda P_1(.|\theta) + (1-\lambda)P_0 \quad \text{s.t.} \\ P_0 \in \mathcal{M}_\alpha &= \left\{ Q \in M \text{ s.t. } \int_{\mathbb{R}^r} dQ(x) = 1, \int_{\mathbb{R}^r} g(x) dQ(x) = m(\alpha) \right\} \end{aligned} \tag{0.2.2}$$

où $g(x) = (g_1(x), \cdots, g_\ell(x))$ et $m(\alpha) = (m_1(\alpha), \cdots, m_\ell(\alpha))$.
L'estimation d'un modèle semiparamétrique défini par des contraintes linéaires a été

étudiée par Broniatowski and Keziou [2012]. Les auteurs proposent d'estimer un modèle semiparamétrique par des $\varphi-$divergences. Broniatowski and Decurninge [2016] ont travaillé avec des modèles semiparamétriques sous des contraintes de L-moments. Afin d'exploiter leurs méthodologies, nous devons travailler plutôt sur un modèle défini par P_0. En effet, supposons avoir un échantillon $X_1, \cdots, X_n$ distribué selon P_T un mélange de deux lois $P_1(.|\theta^*)$ et P_0^*. Nous avons:

$$P_0^* = \frac{1}{1-\lambda^*} P_T - \frac{\lambda^*}{1-\lambda^*} P_1(.|\theta^*). \tag{0.2.3}$$

Définissons l'ensemble:

$$\mathcal{N} = \left\{ Q = \frac{1}{1-\lambda} P_T - \frac{\lambda}{1-\lambda} P_1(.|\theta), \quad \lambda \in (0,1), \theta \in \Theta \right\}. \tag{0.2.4}$$

Notons que P_0^* appartient à $\mathcal{N}$ pour $\lambda = \lambda^*$ et $\theta = \theta^*$. Par ailleurs, comme P_0^* vérifie l'ensemble de contraintes de $\mathcal{M}_{\alpha^*}$ pour un α^* inconnu, nous pouvons écrire:

$$P_0^* \in \mathcal{N} \bigcap \cup_{\alpha \in \mathcal{A}} \mathcal{M}_\alpha.$$

Alors, il est raisonnable de définir une procédure d'estimation qui minimise la distance entre $\mathcal{N}$ et $\cup_{\alpha \in \mathcal{A}} \mathcal{M}_\alpha$. Cette distance est atteinte en P_0^*. Nous avons alors:

$$(\lambda^*, \theta^*, \alpha^*) \in \arg\inf_{\lambda,\theta,\alpha} \inf_{P_0 \in \mathcal{M}_\alpha} D_\varphi \left(P_0, \frac{1}{1-\lambda} P_T - \frac{\lambda}{1-\lambda} P_1(.|\theta) \right). \tag{0.2.5}$$

Ceci est un problème d'optimisation sur un espace de dimension infinie. Pour le résoudre, nous utilisons un résultat de dualité de Fenchel-Legendre, voir Proposition 1.4 de Decurninge [2015] (voir également Proposition 4.2 de Broniatowski and Keziou [2012]).

$$\begin{aligned}
(\lambda^*, \theta^*, \alpha^*) &= \arg\inf_{\phi} \inf_{Q \in \mathcal{M}_\alpha} D_\varphi \left(Q, \frac{1}{\lambda - 1} P_T - \frac{\lambda}{1-\lambda} P_1(.|\theta) \right) \\
&= \arg\inf_{\phi} \sup_{\xi \in \mathbb{R}^{l+1}} \xi^t m(\alpha) - \frac{1}{1-\lambda} \int \psi\left(\xi^t g(x)\right) dP_T(x) \\
&\quad + \frac{\lambda}{1-\lambda} \int \psi\left(\xi^t g(x)\right) dP_1(x|\theta).
\end{aligned} \tag{0.2.6}$$

où $\psi(t) = \sup_x tx - \varphi(x)$. Dans cette formule, nous avons $m(\alpha) = (1, m_1(\alpha), \cdots, m_\ell(\alpha))$. Il est possible maintenant d'estimer le triplet $(\lambda^*, \theta^*, \alpha^*)$ à la base d'un échantillon $X_1, \cdots, X_n$ par:

$$\begin{aligned}
(\hat{\lambda}, \hat{\theta}, \hat{\alpha}) &= \arg\inf_{\lambda,\theta,\alpha} \sup_{\xi \in \mathbb{R}^{l+1}} \xi^t m(\alpha) - \frac{1}{1-\lambda} \frac{1}{n} \sum_{i=1}^{n} \psi\left(\xi^t g(X_i)\right) \\
&\quad + \frac{\lambda}{1-\lambda} \int \psi\left(\xi^t g(x)\right) dP_1(x|\theta).
\end{aligned} \tag{0.2.7}$$

Nous prouvons que cet estimateur est consistant et asymptotiquement Gaussien sous des hypothèses standards.

Example 0.2.1. Prenons l'exemple d'un modèle de mélange à deux composantes dont l'une est définie par trois contraintes de moments (les trois premiers moments). L'ensemble $\mathcal{M}_\alpha$ est défini par:

$$\mathcal{M}_\alpha = \left\{ Q : \int dQ(x) = 1, \int x dQ(x) = m_1(\alpha), \int x^2 dQ(x) = m_2(\alpha), \int x^3 dQ(x) = m_3(\alpha) \right\}.$$

Si $\varphi(t) = (t-1)^2/2$, alors $\psi(t) = \frac{1}{2}t^2 + t$ et l'optimum sur ξ est donné par:

$$\xi(\phi) = \Omega^{-1}\left(m(\alpha) - \int g(x)\left(\frac{1}{1-\lambda}dP(x) - \frac{\lambda}{1-\lambda}dP_1(x|\theta)\right)\right), \text{ pour } \phi \in \Phi^+.$$

où

$$\Omega = \int g(x)g(x)^t\left(\frac{1}{1-\lambda}dP(x) - \frac{\lambda}{1-\lambda}dP_1(x|\theta)\right).$$

Φ^+ est l'ensemble de paramètres pour lesquels la fonction objective est concave par rapport à ξ. En d'autres termes, $\Phi^+ = \{\phi : \Omega \text{ est symétrique définie positive}\}$. Soit M_i le moment d'ordre i de P_T. Dénotons également $M_i^{(1)}(\theta)$ le moment d'ordre i de la composante paramétrique $P_1(.|\theta)$.

$$M_i = \mathbb{E}_{P_T}[X^i], \qquad M_i^{(1)}(\theta) = \mathbb{E}_{P_1(.|\theta)}[X^i].$$

Un calcul simple montre que:

$$\begin{aligned}\Omega &= \int g(x)g(x)^t\left(\frac{1}{1-\lambda}dP(x) - \frac{\lambda}{1-\lambda}dP_1(x|\theta)\right)\\ &= \left[\frac{1}{1-\lambda}M_{i+j-2} - \frac{\lambda}{1-\lambda}M_{i+j-2}^{(1)}(\theta)\right]_{i,j\in\{1,\cdots,4\}}.\end{aligned}$$

et la fonction objective dans (0.2.6) est donnée par:

$$\begin{aligned}H(\phi,\xi) = \xi^t m(\alpha) - \Bigg[&\frac{1}{2}\xi_1^2 + \xi_1 + (\xi_1\xi_2+\xi_2)\left(\frac{1}{1-\lambda}M_1 - \frac{\lambda}{1-\lambda}M_1^{(1)}(\theta)\right)\\ +(\xi_2^2/2+\xi_1\xi_2+\xi_3)&\left(\frac{1}{1-\lambda}M_2 - \frac{\lambda}{1-\lambda}M_2^{(1)}(\theta)\right)+(\xi_1\xi_4+\xi_2\xi_3+\xi_4)\left(\frac{1}{1-\lambda}M_3 - \frac{\lambda}{1-\lambda}M_3^{(1)}(\theta)\right)\\ &+(\xi_3^2/2+\xi_2\xi_4)\left(\frac{1}{1-\lambda}M_4 - \frac{\lambda}{1-\lambda}M_4^{(1)}(\theta)\right) + \xi_3\xi_4\left(\frac{1}{1-\lambda}M_5 - \frac{\lambda}{1-\lambda}M_5^{(1)}(\theta)\right)\\ &+\xi_4^2/2\left(\frac{1}{1-\lambda}M_6 - \frac{\lambda}{1-\lambda}M_6^{(1)}(\theta)\right)\Bigg].\end{aligned}$$

Cet exemple montre que notre estimateur peut être calculé de manière efficace et avec une complexité linéaire sans que la dimension des données intervienne.

Des simulations numériques ont été menées afin de tester la validité de notre approche et de comparer sa performance aux méthdes existantes.

0.2.2 Chapitre 4: Modèles de mélanges semiparamétrique à deux composantes dont l'une est définie par des contraintes de L-moments

Le sujet de ce chapitre est considéré comme la suite du chapitre précédent. Nous avons proposé une structure pour un modèle de mélange semiparamétrique à deux composantes dont l'une est définie par des contraintes linéaires. En prenant des contraintes de moments, on est aperçu que les chiffres de calculs explosent facilement. Exemple 0.2.1 montre le cas des trois premiers moments imposés sur la composante inconnue P_0. Le calcul de la matrice Ω ainsi que la fonction objective ensuite $H(\phi,\xi)$ est une arithmétique entre les moments de la composante paramétrique et les moments du mélange. Avec les trois premiers moments, nous avons déjà à calculer les moments jusqu'à l'ordre 6. Si l'on travaille avec des distributions à queues lourdes, le moment d'ordre 6 prendra des valeurs d'ordre 10^{10}

voire plus. Les moments du mélange eux-mêmes seront remplacés durant l'estimation par des moments empiriques. Ceux-ci explosent rapidement pour un échantillon donné même pour des distributions à queues légères. En effet, le calcul de l'inverse de la matrice Ω devient délicat. Par exemple, durant nos simulations numériques, il n'a pas été possible d'utiliser une méthode numérique pour inverser la matrice Ω car elle a eu une sensibilité élevée dans certains voisinages des paramètres ϕ, et par conséquent, nous avons calculé l'inverse avec des méthodes d'inversion par bloc directes.
Récemment, les L-moments ont été proposés par Hosking [1990] et ils sont de plus en plus utilisés comme des alternatives des moments standards. Les L-moments sont représentatifs et caractérisent la loi de probabilité dès que son espérance existe, voir Théorème 1 de Hosking [1990]. De plus, les quatre premiers L-moments sont des indicateurs de la moyenne, l'échelle, le skewness et le kurtosis. Ces premières propriétés sont déjà intéressantes pour considérer les L-moments. Nous citons aussi le fait que les calculs numériques des L-moments ne s'explosent pas facilement et dans nos simulations, les valeurs numériques ont été toujours proches de 1 et aucun problème de sensibilité de matrices n'a été rencontré.
Avant de procéder à l'introduction du nouveau modèle, nous rappelons la définition des L-moments. Soient $X_{1:n} < \ldots < X_{n:n}$ les statistiques d'ordre associées à un échantillon $X_1, \cdots, X_n$ i.i.d. ayant une fonction de répartition $\mathbb{F}_T$.

Definition 0.2.1. *Le L-moment d'ordre r, noté λ_r, $r = 1, 2, \ldots$ est défini par:*

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} \mathbb{E}\left(X_{r-k:r}\right).$$

Une propriété très intéressante et essentielle des L-moments est qu'ils sont linéaires en les mesures de quantile. Une mesure de quantile est définie par le moyen de la fonction quantile de la manière suivante. Pour tout borélien de $\mathcal{B}([0,1])$

$$\mathbf{F}^{-1}(B) = \int_0^1 \mathbb{1}_{x\in B} d\mathbb{F}^{-1}(x) \in \mathbb{R} \cup \{-\infty, +\infty\}.$$

Les L-moments peuvent être réécrits par

$$\lambda_r = -\int_{\mathbb{R}} K_r(t) d\mathbf{F}^{-1}(t), \qquad r \geq 2 \tag{0.2.8}$$

où

$$K_r(t) = \int_0^t L_{r-1}(u)du = \sum_{k=0}^{r-1} \frac{(-1)^{r-k}}{k+1} \binom{r}{k} \binom{r+k}{k} t^{k+1} \tag{0.2.9}$$

sont les polynômes de Legendre translatés et intégrés. La linéarité des L-moments par rapport aux quantiles est la clé essentielle pour construire notre méthode d'estimation dans un mélange semiparamétrique basée sur le résultat de dualité de Fenchel-Legendre. Nous définissons notre modèle de mélange semiparamétrique à deux composantes dont l'une est définie par des contraintes de L-moments par:

$$\begin{aligned} P(.|\phi) &= \lambda P_1(.|\theta) + (1-\lambda)P_0 \quad \text{s.t.} \\ \mathbf{F}_0^{-1} \in \mathcal{M}_\alpha &= \left\{ \mathbf{Q}^{-1} \in M^{-1}, \mathbf{Q}^{-1} \ll \mathbf{F}_0^{-1} \text{ s.t. } \int_0^1 K(u) d\mathbf{Q}^{-1} = m(\alpha) \right\} \end{aligned} \tag{0.2.10}$$

où $m(\alpha) = (m_2(\alpha), \cdots, m_{\ell-1}(\alpha))$. Nous définissons de manière similaire au chapitre précédent un modèle par le moyen de P_0. Définissons les ensembles:

$$\Phi^+ = \left\{(\lambda,\theta) \in (0,1)\times\Theta : \frac{1}{1-\lambda}\mathbb{F}_T - \frac{\lambda}{1-\lambda}\mathbb{F}_1(.|\theta) \text{ est une fonction de répartition}\right\},$$

$$\mathcal{N}^{-1} = \left\{\mathbf{Q}^{-1} \in M^{-1} \ : \ \exists(\lambda,\theta)\in\Phi^+ \text{ s.t. } \mathbb{Q}^{-1} = \left(\frac{1}{1-\lambda}\mathbb{F}_T - \frac{\lambda}{1-\lambda}\mathbb{F}_1(.|\theta)\right)^{-1}\right\}.$$

L'ensemble Φ^+ représente l'ensemble effectif des paramètres concernés dans le nouveau modèle écrit avec P_0. En effet, un couple (λ,θ) ne définit pas en général une fonction de répartition ayant la forme $\frac{1}{1-\lambda}\mathbb{F}_T - \frac{\lambda}{1-\lambda}\mathbb{F}_1(.|\theta)$. Pour cela, il est important pour le moment de ne garder que les paramètres qui rendent cette fonction une fonction de répartition. Nous aurons la possibilité plus tard d'ignorer ce problème et de travailler sur tout l'ensemble Φ.
Nous avons:

$$\mathbb{F}_0^{*-1} \in \mathcal{N}^{-1} \bigcap \cup_\alpha \mathcal{M}_\alpha.$$

Donc, il est raisonnable de définir une procédure d'estimation qui minimise une distance entre les deux ensembles $\mathcal{N}^{-1}$ et $\cup_\alpha\mathcal{M}_\alpha$.

$$(\lambda^*,\theta^*,\alpha^*) \in \underset{(\lambda,\theta,\alpha)\in\Phi^+}{\arg\inf} \inf_{\mathbf{F}_0^{-1}\in\mathcal{M}_\alpha} D_\varphi\left(\mathbf{F}_0^{-1}, \left(\frac{1}{1-\lambda}\mathbf{F}_T - \frac{\lambda}{1-\lambda}\mathbf{F}_1(.|\theta)\right)^{-1}\right). \quad (0.2.11)$$

Afin de résoudre ce problème d'optimisation qui est mené sur un espace de dimension infinie, nous utilisons à nouveau la Proposition 1.4 de Decurninge [2015] (voir également Proposition 4.2 de Broniatowski and Keziou [2012]) pour écrire:

$$(\lambda^*,\theta^*,\alpha^*) \in \underset{(\lambda,\theta,\alpha)\in\Phi^+}{\arg\inf} \sup_{\xi\in\mathbb{R}^{\ell-1}} \xi^t m(\alpha) - \int_0^1 \psi\left(\xi^t K(u)\right) d\left(\frac{1}{1-\lambda}\mathbf{F}_T - \frac{\lambda}{1-\lambda}\mathbf{F}_1(.|\theta)\right)^{-1}(u).$$

En utilisant le Lemme 1.2 de Decurninge [2015], nous pouvons écrire:

$$(\lambda^*,\theta^*,\alpha^*) \in \underset{(\lambda,\theta,\alpha)\in\Phi^+}{\arg\inf} \sup_{\xi\in\mathbb{R}^{\ell-1}} \xi^t m(\alpha) - \int_{\mathbb{R}} \psi\left(\xi^t K\left(\frac{1}{1-\lambda}\mathbb{F}_T(x) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(x|\theta)\right)\right) dx. \quad (0.2.12)$$

Ceci est une procédure d'estimation où la fonction de répartition générant les données $\mathbb{F}_T$ peut être approximée par sa version empirique afin d'estimer le triple $(\lambda^*,\theta^*,\alpha^*)$ à la base d'un échantillon donné. Cependant, la caractérisation de l'ensemble Φ^+ ici n'est pas évidente et très coûteuse numériquement. De plus, l'ensemble Φ^+ pourrait prendre des formes qui ne sont pas adéquates pour les algorithmes d'optimisation numériques, voir figure (3). Afin de résoudre ce problème, nous montrons que $\phi^* = (\lambda^*,\theta^*,\alpha^*)$ est un infimum global de la fonction $H(\phi,\xi(\phi))$ où:

$$H(\phi,\xi) = \xi^t m(\alpha) - \int \psi\left[\xi^t K\left(\frac{1}{1-\lambda}\mathbb{F}_T(y) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(y|\theta)\right)\right] dy,$$

et

$$\xi(\phi) = \underset{\xi\in\mathbb{R}^{\ell-1}}{\arg\sup}\ H(\phi,\xi).$$

Donc, si la fonction $H(\phi,\xi(\phi))$ n'a qu'un seul infimum global, ceci ne sera autre que ϕ^*. Cela justifie notre procédure d'estimation:

$$\phi^* = \underset{(\alpha,\theta,\lambda)\in\Phi}{\arg\inf} \sup_{\xi\in\mathbb{R}^{\ell-1}} \xi^t m(\alpha) - \int \psi\left[\xi^t K\left(\frac{1}{1-\lambda}\mathbb{F}_T(x) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(x|\theta)\right)\right] dx. \quad (0.2.13)$$

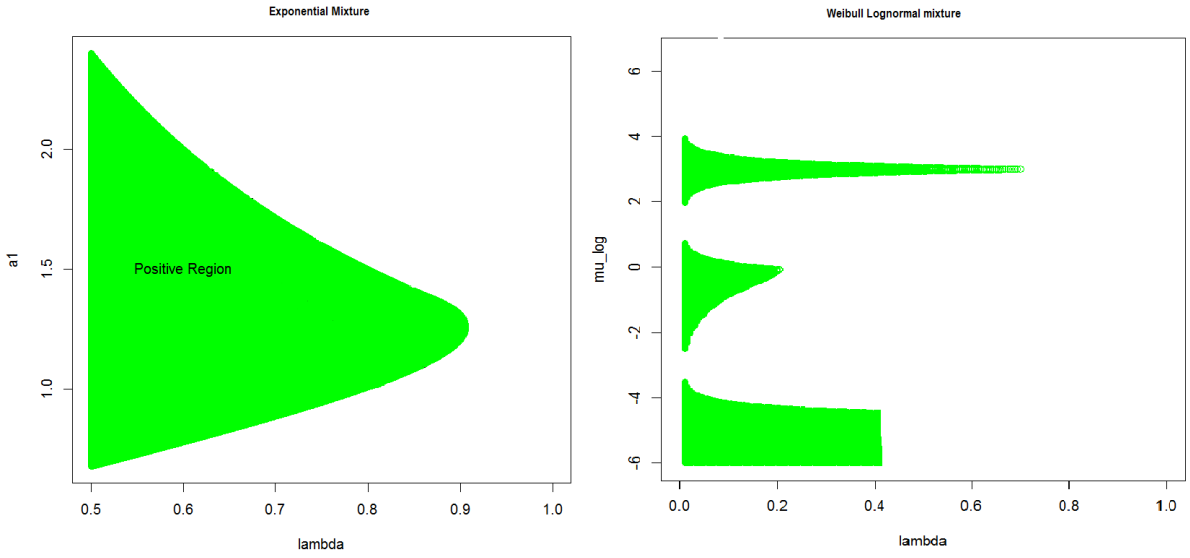

Figure 3: Differentes formes de l'ensemble Φ^+. Pour le mélange Weibull-Lognormal, c'est le Weibull qui est la composante semiparametrique.

Avec un échantillon $X_1, \cdots, X_n$ distribué selon $\mathbb{F}_T$, nous estimons ϕ^* par:

$$\hat{\phi} = \underset{(\alpha,\theta,\lambda)\in\Phi}{\arg\inf} \sup_{\xi\in\mathbb{R}^{\ell-1}} \xi^t m(\alpha) - \int \psi\left[\xi^t K\left(\frac{1}{1-\lambda}\mathbb{F}_n(x) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(x|\theta)\right)\right] dx. \quad (0.2.14)$$

Nous montrons que cet estimateur est consistant et asymptotiquement Gaussien sous des hypothèses standards. Les simulations numériques montrent le gain de l'utilisation des contraintes de L-moments par rapport aux moments standard surtout lorsque la proportion de la composante paramétrique est très basse; 0.05 voire 0.01.

Example 0.2.2. Prenons le cas d'une divergence de χ^2 pour $\varphi(t) = (t-1)^2/2$ et $\psi(t) = t^2/2 + t$. La fonction objective $H(\phi, \xi)$ est donnée par:

$$H(\phi,\xi) = \xi^t m(\alpha) - \int \frac{1}{2}\left(\xi^t K\left(\mathbb{F}_0(y|\phi)\right)\right)^2 + \xi^t K\left(\mathbb{F}_0(y|\phi)\right) dy,$$

où

$$\mathbb{F}_0(y|\phi) = \frac{1}{1-\lambda}\mathbb{F}_T(y) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(y|\theta).$$

La fonction H est un polynôme de degrée 2 en ξ. Pour tout $\phi \in \Phi$, le supremum par rapport à ξ est donné par:

$$\xi(\phi) = \Omega^{-1}\left(m(\alpha) - \int K(\mathbb{F}_0(y|\phi))dy\right),$$

où

$$\Omega = \int K\left(\mathbb{F}_0(y|\phi)\right) K\left(\mathbb{F}_0(y|\phi)\right)^t dy.$$

La matrice Hessienne de $\xi \mapsto H(\phi, \xi)$ est égale à $-\Omega$, et donc c'est une matrice symétrique définie négative pour tout ϕ dans Φ. Par conséquent, $\xi(\phi)$ est un supremum global de $\xi \mapsto H(\phi, \xi)$.

Part I

Robust Estimation Using $\varphi-$Divergences with Application to Parametric Mixture Models

Chapter 1

Estimation using a phi-divergence

The maximum likelihood estimation is a simple and efficient method to estimate unknown parameters of a given model. The most common drawback of this method is its sensibility to contamination and misspecification. From the first years of the twentieth century, many researchers such as Pearson, Hellinger, Kullback, Neymann and others started developing different approaches using distance-like functions between probability density functions called divergences. Several divergence-based techniques permit to construct robust estimators, such as $\varphi-$divergences (Csiszár [1963], Ali and Silvey [1966]), $S-$divergences (Ghosh et al. [2013]), Rényi pseudodistances (see for example Toma and Leoni-Aubin [2013]), Bregman divergences and many others. In this work, we are particularly interested in $\varphi-$divergences on the one hand, and on the other hand, in comparing the resulting estimators and existing approaches with the maximum likelihood estimator (MLE) and some well-known divergences.
Estimation using $\varphi-$divergences is based on the idea of minimizing a distance between the true distribution and a given model. In practice, the true distribution is replaced by its empirical version calculated on the basis of an $n-$sample. When working with discrete models, everything goes well and a simple plug-in of the empirical distribution results in a plausible and good estimation procedure, see Lindsay [1994]. The true challenge appears when we work with continuous models and smoothing techniques are apparently necessary tools. Several techniques were proposed in the literature. We give a brief summary of some of these approaches and present a new method which has very encouraging performances and properties. A comparison of several methods based on $\varphi-$divergences will be presented at the end of this chapter with an extensive simulation study on several distributions. The comparison is held with respect to the maximum likelihood estimator (MLE) and a powerful estimator called the minimum density power divergence (MDPD) introduced by Basu et al. [1998].

1.1 A brief introduction about φ-divergences

1.1.1 Definition, useful properties and standard examples

φ-divergences were introduced independently by Csiszár [1963] (as "f-divergences") and Ali and Silvey [1966]. Let P and Q be two $\sigma-$finite measures defined on $(\mathbb{R}^r, \mathscr{B}(\mathbb{R}^r))$ such that Q is absolutely continuous (a.c.) with respect to (w.r.t.) P. Let $\varphi : \mathbb{R} \mapsto [0, +\infty]$ be a proper convex function with $\varphi(1) = 0$ and such that its domain $\text{dom}\varphi = \{x \in \mathbb{R} \text{ such that } \varphi(x) < \infty\} := (a_\varphi, b_\varphi)$ with $a_\varphi < 1 < b_\varphi$. The φ-divergence between Q

and P is defined by:

$$D_\varphi(Q,P) = \int_{\mathbb{R}^r} \varphi\left(\frac{dQ}{dP}(x)\right) dP(x), \tag{1.1.1}$$

where $\frac{dQ}{dP}$ is the Radon-Nikodym derivative. When Q is not a.c.w.r.t. P, we set $D_\varphi(Q,P) = +\infty$. When, $P = Q$ then $D_\varphi(Q,P) = 0$. Furthermore, if the function $x \mapsto \varphi(x)$ is strictly convex on a neighborhood of $x = 1$, then

$$D_\varphi(Q,P) = 0 \quad \text{if and only if} \quad P = Q. \tag{1.1.2}$$

In the definition of $\varphi-$divergences, we have considered the general case of $\sigma-$finite measures. In the whole Part I (Chapters 1 and 2) of this work, we will only be interested in $\varphi-$divergences between probability measures. In Chapter 3, we will be working with finite signed measures, and finally in Chapter 4, we will be working in the frame frame of the general case of $\sigma-$finite measures.
Several standard statistical divergences can be expressed as $\varphi-$divergences; the Hellinger, the Pearson's and the Neymann's χ^2, and the (modified) Kullback-Leibler. They all belong to the class of Cressie-Read (see Cressie and Read [1984]), also known as "power divergences", defined through the generator function φ_γ given by:

$$\varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)}, \tag{1.1.3}$$

for $\gamma = \frac{1}{2}, 2, -2, 0, 1$ respectively[1]. More details and properties can be found in Liese and Vajda [1987] or Pardo [2006].
Estimators based on $\varphi-$divergences were developed in the parametric (see Beran [1977],Lindsay [1994],Park and Basu [2004],Broniatowski and Keziou [2009a]) and the semiparametric setups (see Broniatowski and Keziou [2012] and Broniatowski and Decurninge [2016]). In completly nonparametric setup, we may mention the work of Karunamuni and Wu [2009] on two component mixture models when both components are unknown.

1.1.2 General estimation based on φ-divergences

Estimation based on $\varphi-$divergences consists in finding the projection of the true distribution P_T on the set $\{P_\phi, \phi \in \Phi\}$, i.e. the model. Minimum discrepancy or minimum divergence estimators are defined by:

$$\phi^T = \arg\min_{\phi\in\Phi} D_\varphi(P_\phi, P_T). \tag{1.1.4}$$

This procedure was proved to be robust in the sens that a perturbation of the model in a small neighborhood of P_T would result in a small perturbation in the resulting estimates, see Donoho and Liu [1988]. Beran [1977] has proved that $\arg\min_{\phi\in\Phi} D_\varphi(P_\phi, P)$ for the case of the Hellinger divergence is continuous as a function of P in a Hellinger neighborhood of P_T. This is also translated into an automatic robustness of the Hellinger divergence for small perturbations around the true distribution in the Hellinger topology.

In practice, the estimation procedure (1.1.4) needs to be approximated on the basis of a dataset $X_1, \cdots, X_n$ since the true distribution is unknown. When working with discrete

[1]For $\gamma \in \{0,1\}$, the limit is calculated since it is not well-defined. We denote $\varphi_0(x) = -\log x + x - 1$ for the case of the modified Kullback-Leibler and $\varphi_1(x) = x\log x - x + 1$ for the Kullback-Leibler.

models, $\varphi-$divergences are approximated using a direct plug-in of the empirical distribution P_n. This is possible because both the model and the empirical distribution are absolutely continuous with respect to each others for large n. Efficient and robust estimators were derived and extensively studied; see for example Simpson [1987] and Lindsay [1994].

For continuous models, the empirical distribution is no longer suitable to replace directly the true distribution since the model has a continuous support. Thus, the model is not absolutely continuous with respect to P_n for any n and no estimation procedure can be produced, see Broniatowski and Vajda [2012] for a discussion about this point. Authors such as Beran [1977], Park and Basu [2004] and Kuchibhotla and Basu [2015] proposed to simply smooth the empirical distribution using kernels, see paragraph 1.2.1. Basu and Lindsay [1994] proposed to smooth both the model and the empirical distribution; see paragraph 1.2.2. Although smoothing the model may result in a loss of information, Basu and Lindsay show, in simple models, that this loss is rather small. They also notice that there is still a difficulty in the choice of the window and the kernel for the smoothing.
Recently, an approach based on some convexity arguments has been proposed independently by Liese and Vajda [2006] and Broniatowski and Keziou [2006], see paragraph 1.3.1. In both articles, the authors provide similar "supremal" representations of $\varphi-$divergences where a simple plug-in of the empirical distribution is possible without any smoothing techniques. The resulting estimators were called as minimum dual $\varphi-$divergence estimators (MDφDE). Another estimator based on the dual formula called the dual $\varphi-$divergence estimator (DφDE) was proposed. This estimator is proved to be consistent by Broniatowski and Keziou [2009b], see paragraph 1.3.2 for more details. Since the introduction of the MDφDE, no complete study about its robustness was proposed except for the calculus of the influence function in Toma and Broniatowski [2011] and Broniatowski and Vajda [2012]. There were no simulation studies either, except for the paper of Frýdlovà et al. [2012]. However, in the later, the authors have considered only the case of Gaussian model where the MDφDE coincides with the maximum likelihood estimator. Broniatowski [2014] has proved that the MDφDE coincides with the MLE on any regular exponential family, hence on a Gaussian model. Hence, the simulation results of Frýdlovà et al. [2012] shows only that known fact that the MLE is not robust.
The dual representation proposed by both Liese and Vajda [2006] and Broniatowski and Keziou [2006] yields estimators which perform well under the model and have efficiency comparable to the MLE[2]. Weak and strong consistency is reached under classical conditions (see Broniatowski and Keziou [2009b]). Limit laws of the MDφDE and the estimated divergence are simple and were exploited to build statistical tests. However, when we are not under the model, this approach suffers from lack of robustness. Under contamination or under misspecification, this approach does not approximate well the $\varphi-$divergence between the true distribution and the model. It even remarkably underestimates its value. We propose in the sequel a brief explanation of this problem and provide a general solution, see paragraph 1.4.2. A new robust estimator called kernel-based MDφDE is introduced. Our estimator avoids the supremal form of the MDφDE, see paragraph (1.5.1). We study asymptotic properties of this estimator in Section 1.6.

[2]The MDφDE is even as efficient as the MLE in regular exponential families.

1.2 Estimation based on $\varphi-$divergences in continuous models

In what follows, we suppose to have an i.i.d. sample $Y_1, \cdots, Y_n$ drawn from the probability distribution P_T. The function K will denote a kernel function defined on $\mathbb{R}^r$ not necessarily symmetric. In this section, we present two general approaches to approximate a $\varphi-$divergence on the basis of a given sample.

1.2.1 Beran's approach: Smoothing of the empirical distribution

A simple and natural approach to approximate the $\varphi-$divergence between the true distribution of the data P_T and the model is to replace P_T by a smoothed version of the empirical distribution P_n, say $K_{n,w}$ with w a smoothing parameter. An estimator of ϕ^T is then given by:

$$\hat{\phi} = \arg\min_{\phi\in\Phi} \int \varphi\left(\frac{p_\phi}{K_{n,w}}\right)(y) K_{n,w}(y) dy. \tag{1.2.1}$$

This method was first introduced in the context of $\varphi-$divergences by Beran [1977] who studied the Hellinger divergence in a univariate context and proved that it is robust and asymptotically efficient in the same time. It was then generalized to the class of $\varphi-$divergences in the univariate context, see Park and Basu [2004] and Kuchibhotla and Basu [2015]. The Hellinger divergence has very favorable properties. Indeed, Jiménez and Shao [2001] proved that no minimum power-divergence estimator performs better than the minimum Hellinger in terms of both second order efficiency and robustness. The idea of Beran was also employed in the estimation of the proportion of a nonparametric mixture model, see Karunamuni and Wu [2009]. In their approach, however, we suppose to have three i.i.d. samples; a sample drawn according to each component and a sample drawn from the whole mixture. See also Tang and Karunamuni [2013] for an application on finite mixture regression models and the references therein.
In the multivariate context, Tamura and Boos [1986] have studied the asymptotic properties of the Hellinger divergence. Surprisingly, the estimator needs a correction term in order to converge to a multivariate Gaussian distribution at a $\sqrt{n}$ speed. In the univariate case, this correction term does not exist since it converges to zero in probability when multiplied by $\sqrt{n}$.

Asymptotic properties of the resulting estimators were only studied in the previous references when $K_{n,w}$ is the Parzen-Rosenblatt kernel density estimator, i.e. a symmetric kernel density estimator. In the context of nonnegative supported distributions, the use of symmetric kernels is not advised especially if there is a considerable mass near zero which is the case for example of the exponential distribution. Several techniques for bias correction were proposed, see Karunamuni and Alberts [2005] for a survey. We mention also asymmetric kernels, see for example Libengue Dobele-kpoka [2013] for a general approach. We give in the next paragraph more details especially about asymmetric kernels and provide some examples. These two solutions provided considerable improvement in the estimation of densities defined on the half real line. To the best of our knowledge, the use of asymmetric kernels in parametric estimation has not been considered in the literature. This is may be because asymmetric kernels is still a recent topic and the first paper goes back to Chen [1999]. Moreover, the theory is not sufficiently developed yet. Indeed, consistency of asymmetric kernel density estimators is only proved on every compact subset of the domain of definition of the true density and not on the whole domain.

Consistency becomes more difficult to prove when the density explodes for example near zero. Furthermore, the rules for the choice of the window are not very efficient as we will see in the simulations in Section 1.7.

Remark 1.2.1. Unlike symmetric kernels, generalization of asymmetric kernels to the multivariate case is not simple. So far, and to the best of our knowledge, there is only two recent papers which treat the multivariate case Bouezmarni and Rombouts [2010] and Funke and Kawka [2015]. The two papers suppose that the data is bounded. Both methods can be applied in our kernel-based MDφDE and in Beran's method (and its generalization), but more investigations are needed in order to employ them in the Basu-Lindsay approach.

1.2.2 The Basu-Lindsay approach: Smoothing the model

The idea of smoothing the empirical distribution was applied to avoid the problem of absolute continuity of the model with respect to P_n when we use the later to replace the true distribution in (1.1.1). Basu and Lindsay [1994] argue that the use of this method requires consistency and rates of convergence for the kernel estimator. Thus, they propose to smooth not only the empirical distribution, but also the model. Indeed, smoothing equally the model p_ϕ and the empirical measure P_n, as in (1.2.2) here below, may reduce the influence of the choice of the *window* on the resulting estimator. For example, if the smoothing is by convolution with a symmetric kernel K such as the Gaussian kernel, the Basu-Lindsay approach is summarized in the following two lines:

$$\begin{aligned} p^*_\phi(x) &= \frac{1}{w}\int_{\mathbb{R}} p_\phi(y) K\left(\frac{x-y}{w}\right) dy; \\ \hat{\phi} &= \arg\inf_{\phi\in\Phi} \int_{\mathbb{R}} \varphi\left(\frac{p^*_\phi(x)}{K_{n,w}(x)}\right) K_{n,w}(x)dx, \end{aligned} \tag{1.2.2}$$

where $K_{n,w}(x) = \frac{1}{nw}\sum K\left(\frac{x-y_i}{w}\right)$ is the Parzen-Rosenblatt symmetric-kernel estimator. For example, in the Gaussian model $\mathcal{N}(\mu, \sigma^2)$, the smoothed model is merely a Gaussian density with variance equal to $\sigma^2 + h^2$. Thus, the Basu-Lindsay approach appears as if we are calculating a divergence between a *weighted* version of the model and the kernel estimator.

The authors prove the robustness of (1.2.2) using the residual adjustment function (RAF), see Lindsay [1994], since the corresponding influence function is generally unbounded, keeping first order efficiency in hand. The basic problem from a theoretical point of view is that in order to eliminate the role of the smoothing window, one needs to find what the authors call a *transparent* kernel[3]. This is a very hard task in general as has already been mentioned in Kuchibhotla and Basu [2015] for example. Basu and Lindsay have only provided three simple examples (Gaussian, Poisson and gamma) where one can provide a transparent kernel but have not shown any leads for a general method. They have also shown in simple examples that when we use non transparent kernels, loss of information is not large. Besides, consistency and rates of convergence for the kernel estimator become necessary in order to obtain the consistency of the resulting estimator.

We will show in the following paragraph that if we are working with non classical situations

[3]The transparency assumption here means that the smoothed score function (derivative of the log-likelihood) is proportional to the non smoothed one. The proportion rate can only be a function of the parameters.

such as densities defined on $[0, \infty)$, we may encounter further difficulties in the Basu-Lindsay approach.

Smoothing-the-model's effect: symmetric versus asymmetric kernels

The Basu-Lindsay approach seems to be more sensitive to the choice of the *kernel* than standard methods. For example, let's take the case of densities defined on $(0, \infty)$ (with zero possibly included). Simple examples of such distributions are Weibull distributions and generalized Pareto distributions (GPDs). It is well-known that estimation based on symmetric kernels is biased near zero. Thus, smoothing the model with such kernels will result in similar bias near zero. Figure 1.1 shows the influence of a Gaussian kernel on a GPD model. The smoothed model has a peak near zero and decreases then towards zero, and hence largely underestimates the values of the "not smoothed" model near zero. Thus, the divergence calculates a distance between a biased estimator of the true distribution and a biased model, and there is no intuitive guarantee of what should give the minimization of such function. Standard methods which do not smooth the model would suffer less from this sort of problems since the bias is only in the kernel estimator.
Simulation results in Section 1.7 show that among the three methods which use a kernel estimator (Beran's approach, the Basu-Lindsay approach and our kernel-based MDφDE which will be introduced later on) the Basu-Lindsay approach is the most sensitive one. Under the model, all three methods do not give satisfactory results in comparison to the MLE (or the classical MDφDE which will be presented later on) when we use symmetric kernels. When outliers are present, they still give a better result than the MLE.

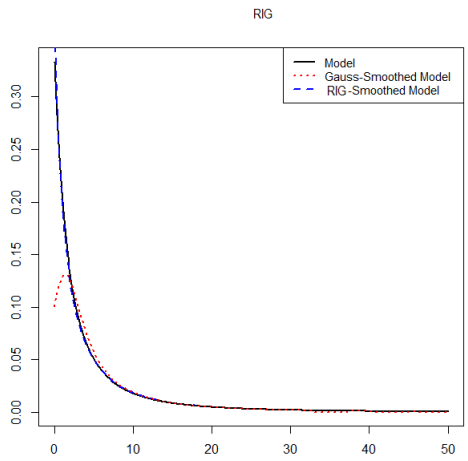

Figure 1.1: Smoothing the model with a Gaussian kernel results in a great loss in information. The use of an asymmetric kernel such as the the reciprocal inverse Gaussian (RIG) seems to be a good alternative

The solution for the previous problem is of course to either use a bias-correction method, see Karunamuni and Alberts [2005], or to use asymmetric kernels which do not suffer from the boundary bias, see Libengue Dobele-kpoka [2013]. A more intriguing example is a Weibull distribution with shape parameter in $(0, 1)$. The density function explodes to infinity as we approach from zero[4]. Cases such as GPD models can be treated efficiently

[4]Of course, if we are defining the Weibull distribution with a location parameter, the pdf explodes to

using bias-correction methods since the support needs to be semi-closed. Models which have singularities such as the Weibull model can be treated using asymmetric kernels such as gamma kernels or reciprocal inverse Gaussian kernels[5]. Such kernels could be employed to recover a good performance in the Basu-Lindsay approach.

Let's see how this kind of solution can be applied on the Basu-Lindsay approach. We discuss only the case of asymmetric kernels since similar arguments hold for bias-correction methods. Let $\hat{f}$ be the asymmetric-kernel estimator defined by:

$$\hat{f}(x) = \frac{1}{nc(y_1, \cdots, y_n)} \sum_{i=1}^{n} K_{x,w}(y_i),$$

where $K_{x,w}$ is the asymmetric kernel calculated at observation y_i, and $c(y_1, \cdots, y_n)$ is a constant which ensures integrability to 1. For example, K is the gamma kernel:

$$K_{x,w}(y) = \frac{y^{x/w}}{\Gamma(1 + x/w)h^{1+x/w}} e^{-y/w}, \qquad \text{for } y \in [0, \infty),$$

where Γ is the classical gamma function. Estimator $\hat{f}$ can no longer be defined as the convolution between the asymmetric kernel and the empirical distribution in the same way as symmetric ones. Thus, the smoothed model in the Basu-Lindsay approach can no longer be obtained by simple convolution. It is given by:

$$p^*_\phi(x) = \int_0^\infty \frac{1}{c(y)} K_{x,w}(y) p_\phi(y) dy,$$

where $c(y)$ is a function which normalizes the kernel for each value of y in order to be a density. It is given by:

$$c(y) = \int_0^\infty K_{z,w}(y) dz.$$

Unfortunately, this normalization function cannot be calculated but numerically. Taking into account the number of integrations needed to perform such a task and the calculus of the $\varphi-$divergence afterwards which also needs numerical integration, we get a high complexity and execution time. In comparison to the classical approach of smoothing only the empirical distribution (Kuchibhotla and Basu [2015]), the calculus of the smoothed model imposes two extra embedded integrals making the calculus of the $\varphi-$divergence very difficult on two levels. The first one is the execution time, and the second one is the subtlety of the whole calculus since all these integrals are carried out over slow decreasing functions on the half real line[6].

Remark 1.2.2. We were unable to use asymmetric kernels in the Basu-Lindsay approach, because integration calculus (three embedded ones) failed even when restricting the calculus of the normalizing function $c(y)$ on a finite interval. The execution time using the statistical tool R Core Team [2015] on an i7 laptop with 8G RAM took 12 minutes for a simple calculus of the smoothed model. One can imagine now the execution time of the

infinity near the value of the location parameter.

[5] Asymmetric kernels have an attractive property that they can treat both bounded and unbounded densities.

[6] The calculus of bounded integrals is far more simple than infinite integrals. Besides, a slow decreasing function (at the border of the its domain), even if it is smooth, is harder to be handled by numerical integration methods than fast decreasing ones.

$\varphi-$divergence and finally the optimization over ϕ. The method should work if one can handle efficiently the problem of numerical integrations and give close results to the case when we do not smooth the model.

Remark 1.2.3. The use of the normalization function is necessary to get a very small loss of information. If it is not used, there will be a similar underestimation near zero to the case of symmetric kernels when applied on models defined on a semi-closed intervals.

The Varying Kernel Density Estimator

Very recently, Mnatsakanov and Sarkisian [2012] have proposed a kernel-type estimator which does not contain a normalization function. Their approach is based on the so called Mellin transform to approximate the distribution function and then derive an estimate of the density function. Let $\mathbb{F}$ be a cdf and define the following operator:

$$\left(\mathcal{M}\mathbb{F}\right)(j) = \int_0^\infty t^{-j} d\mathbb{F}(t) = \mu_j, \qquad \text{for } j = 0, 1, \cdots$$

Introduce the sequence of operators $\mathcal{M}_\alpha^{-1}$:

$$\left(\mathcal{M}_\alpha^{-1}\mu\right)(x) = 1 - \sum_{k=0}^{\alpha} \frac{(\alpha x)^k}{k!} \sum_{j=k}^{\infty} \frac{(-\alpha x)^{j-k}}{(j-k)!}\mu_j, \qquad x \in \mathbb{R}_+. \tag{1.2.3}$$

Here $\mu = \{\mu_j, j = 0, 1, \cdots\}$ and $\alpha \to \infty$ at a specific rate. The transform $\mathcal{M}\mathbb{F}(1-z)$ where z is a complex variable, is known as the Mellin transform. It is possible to recover a function from its Mellin Transform, see for example Tagliani [2001]. Under some conditions, we may write:

$$\mathbb{F}_\alpha = \mathcal{M}_\alpha^{-1}\mathcal{M}\mathbb{F} \xrightarrow[\text{weakly}]{\alpha\to\infty} \mathbb{F}.$$

Inserting the empirical moments in (1.2.3), we may construct an estimator of $\mathbb{F}$ as follows:

$$\tilde{\mathbb{F}}(x) = 1 - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{\alpha}\frac{1}{k!}\left(\frac{\alpha}{X_i}x\right)^k \exp\left(-\frac{\alpha}{X_i}x\right), \qquad x \in \mathbb{R}_+.$$

Notice that the inner sum in the previous display tends to $\mathbb{1}_{X_i>x}$ as $\alpha \to \infty$ which means that $\tilde{\mathbb{F}}$ approximates indeed $\mathbb{F}$. Note also that the estimator $\tilde{\mathbb{F}}$ is derivable so that an estimator of the density can be deduced directly by derivation as follows:

$$\hat{f}_\alpha(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i}\frac{1}{\Gamma(\alpha)}\left(\frac{\alpha x}{y_i}\right)^\alpha \exp\left(-\frac{\alpha x}{y_i}\right), \tag{1.2.4}$$

for a "bandwidth" $\alpha \in \mathbb{N}^*$. This is called by Mnatsakanov and Sarkisian [2012] the varying kernel density estimator (vKDE). This estimator is different from the estimators defined based on symmetric or asymmetric kernels as explained by the authors. They provide a bias-corrected version of this estimator to reduce the bias at the boundary in practice. Nevertheless, we prefer to use (1.2.4) because it integrates to 1 and the Basu-Lindsay approach can be performed more efficiently and reasonably in comparison to the use of asymmetric kernels when working with distributions defined on $\mathbb{R}_+$. The parameter α is a natural number, and (1.2.4) is $L1-$consistent as α goes to infinity under suitable conditions. It even achieves the optimal rate of convergence for MSE and MISE.

It is important to notice that $\hat{f}_\alpha(0) = 0$ for $\alpha \geq 1$. Thus, it is preferable in the context of density estimation to be used for densities which have value equal to 0 at 0 or for densities which are defined on $(0, \infty)$. However, in kernel-based estimation procedures, the value at zero is not important because it disappears in integration calculus. Besides, no observation will have exactly the value zero. Thus, (1.2.4) can still be used in a parameter estimation procedure even if we are working with densities not well defined on zero or have a positive value at zero.

1.3 A plug-in estimate: the dual formula of $\varphi-$divergences

1.3.1 The minimum dual ϕ-divergence estimator

Liese and Vajda [2006] propose the following "supremal" representation of $\varphi-$divergences. Let $\mathcal{P}$ be a class of mutually absolutely continuous distributions such that for any triplet P, P_T and Q, $\varphi'(dP/dQ)$ is P_T-integrable. Theorem 17 in Liese and Vajda [2006] states that:

$$D_\varphi(P_T, P) = \sup_{Q \in \mathcal{P}} \int \varphi'\left(\frac{dQ}{dP}\right) dP_T + \int \varphi\left(\frac{dQ}{dP}\right) dP - \int \varphi'\left(\frac{dQ}{dP}\right) dQ \tag{1.3.1}$$

and the supremum is attained when $Q = P_T$.
Broniatowski and Keziou [2006] have also developed a similar and a more general representation of $D_\varphi(P, P_T)$. Let $\mathcal{F}$ be any class of $\mathcal{B}-$measurable real valued functions. Let $\mathcal{M}_\mathcal{F}$ be the subspace of the space of probability measures $\mathcal{M}$ defined by $\mathcal{M}_F = \{P \in \mathcal{M} \text{ s.t. } \int |f| dP < \infty, \forall f \in \mathcal{F}\}$. Assume that φ is differentiable and strictly convex. Then, for all $P \in \mathcal{M}_\mathcal{F}$ such that $D_\varphi(P, P_T)$ is finite and $\varphi'(dP/dP_T)$ belongs to $\mathcal{F}$, the $\varphi-$divergence admits the dual representation (see Theorem 4.4 in Broniatowski and Keziou [2006]):

$$D_\varphi(P, P_T) = \sup_{f \in \mathcal{F}} \int f dP - \int \varphi^*(f) dP_T, \tag{1.3.2}$$

where $\varphi^*(x) = \sup_{t \in \mathbb{R}} tx - \varphi(t)$ is the Fenchel-Legendre convex conjugate of φ. Moreover, the supremum is attained at $f = \varphi'(dP/dP_T)$.
When substituting $\mathcal{F}$ by the class of functions $\{\varphi'(dP/dQ)\}$, and using the property $\varphi^*(\varphi'(t)) = t\varphi'(t) - \varphi(t)$, we obtain the same representation given above in (1.3.1). Both formulations (1.3.1) and (1.3.2) are interesting in their own and in their proofs. The second formula gives us the opportunity to reproduce many supremal forms for the $\varphi-$divergence. In a parametric setup where $dP_\phi = p_\phi dx$ for $\phi \in \Phi \subset \mathbb{R}^d$ and the true distribution generating the data is a member of the model, i.e. $P_T = P_{\phi^T}$ for some $\phi^T \in \Phi$, Broniatowski and Keziou [2009b] propose to use the class of functions $\mathcal{F}_\phi = \{\varphi'(p_\phi/p_\alpha), \alpha \in \Phi\}$. Assume that φ and its convex dual are strictly convex. Suppose also the integrability condition:

$$\int \left|\varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x)\right| p_\phi(x) dx < \infty, \quad \forall \alpha, \phi \in \Phi. \tag{1.3.3}$$

Then, the dual representation of D_φ in the parametric setting is now written as:

$$D_\varphi(p_\phi, p_{\phi_T}) = \sup_{\alpha \in \Phi} \left\{ \int \varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x) p_\phi(x) dx - \int \varphi^\#\left(\frac{p_\phi}{p_\alpha}\right)(y) p_{\phi^T}(y) dy \right\}. \tag{1.3.4}$$

where $\varphi^\#(t) = t\varphi'(t) - \varphi(t)$. The idea behind this choice is that the supremum is attained when $\alpha = \phi^T$. Since p_{ϕ^T} is unknown, one thinks about replacing $p_{\phi^T} dy$ by the empirical

distribution. This seems very natural and does not cause any problem of absolute continuity as in formula (1.1.1). Moreover, no smoothing is needed. We now get the following approximation:

$$\hat{D}_\varphi(p_\phi, p_{\phi_T}) = \sup_{\alpha\in\Phi}\left\{\int \varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x)p_\phi(x)dx - \frac{1}{n}\sum_{i=1}^{n}\varphi^{\#}\left(\frac{p_\phi}{p_\alpha}\right)(y_i)\right\}. \tag{1.3.5}$$

Both Broniatowski and Keziou [2009b] and Liese and Vajda [2006] propose to estimate the set of parameters ϕ^T by:

$$\hat{\phi}_n = \arg\inf_{\phi\in\Phi}\sup_{\alpha\in\Phi}\hat{D}_\varphi(p_\phi, p_{\phi_T}). \tag{1.3.6}$$

This was called by Broniatowski and Keziou [2009b] the minimum dual $\varphi-$divergence estimator (MDφDE). The authors have studied the asymptotic properties and provided sufficient conditions for the consistency of this estimator. They have also built some statistical tests based on it. Toma and Broniatowski [2011] and Broniatowski and Vajda [2012] have studied the robustness of such an estimator from an influence function (IF) point of view. The IF is unfortunately unbounded in general and does not even depend on φ for the class of Cressie-Read functions φ_γ presented in the introduction. This fact is still not sufficient to conclude the non robustness of the MDφDE. It was pointed out by many authors in the context of $\varphi-$divergences that one may have an unbounded influence function, still the resulting estimators enjoy good robustness against outliers, see Beran [1977] for the Hellinger divergence in continuous models and Lindsay [1994] for a general class of $\varphi-$divergences in discrete models. In the former paper, Beran has studied the robustness by considering the *Hellinger* continuity of the approximate distribution for the estimator when the model varies in a small Hellinger neighborhood of the true distribution. In the later paper, Lindsay has studied the robustness through Pearson's residuals by introducing a new criterion called as the residual adjustment function (RAF). Robustness properties were studied through the RAF and by simulations. In the context of S-divergences, Ghosh et al. [2013] has shown that the robustness of the resulting estimator depends on two parameters although the IF only depends on one of them.
So far, and to the best of our knowledge, there is not even a simulation study of the robustness of the MDφDE although it is an estimator which, similarly to the power density estimator of Basu et al. [1998], does not require any smoothing or escort parameters. Besides, the asymptotic properties are proved with merely classical conditions on the model. The only simulation study is done by Frýdlovà et al. [2012] and focuses only on the Gaussian model. In their results, the MDφDE yields similar results to the maximum likelihood estimator when no contamination is present, while they get some cases where the MDφDE is robust under contamination, although they *should not* as we will see later in paragraph 1.4.2.

1.3.2 The Dual $\varphi-$divergence estimator

General facts and comments

The dual $\varphi-$divergence estimator (DφDE) was defined in Broniatowski and Keziou [2009b] (see also Keziou [2003]) as the argument of the supremum in (1.3.5). It is defined by:

$$\hat{\alpha}_n(\phi) = \arg\sup_{\alpha\in\Phi}\left\{\int \varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x)p_\phi(x)dx - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{p_\phi}{p_\alpha}\varphi'\left(\frac{p_\phi}{p_\alpha}\right) - \varphi'\left(\frac{p_\phi}{p_\alpha}\right)\right](y_i)\right\} \tag{1.3.7}$$

for a given "escort" parameter ϕ. This M-estimator is far more simple than the classical MDφDE defined by (1.3.6) since it needs only one optimization over α for a given choice of the escort parameter ϕ. Besides, Toma and Broniatowski [2011] proved that this estimator is robust in some scale and location models from an IF point of view, *provided a suitable choice of the escort parameter*. Toma and Broniatowski [2011], Toma and Leoni-Aubin [2010] and Keziou [2003] built robust tests using this estimator.

Relation with the density power divergences

The minimum density power divergence (MDPD) was first introduced by Basu et al. [1998]. It is defined by:

$$\begin{aligned}\hat{\phi}_n &= \arg\inf_{\phi\in\Phi} \int p_\phi^{1+a}(z)dz - \frac{a+1}{a}\frac{1}{n}\sum_i^n p_\phi^a(y_i) \\ &= \arg\inf_{\phi\in\Phi} \mathbb{E}_{P_\phi}\left[p_\phi^a\right] - \frac{a+1}{a}\mathbb{E}_{P_n}\left[p_\phi^a\right]. \end{aligned} \tag{1.3.8}$$

Let's look at the DφDE for power divergences with $\gamma = -a < 0$. It is given by:

$$\begin{aligned}\hat{\alpha}_n &= \arg\sup_{\alpha\in\Phi} \frac{1}{\gamma-1}\int \frac{p_\theta^\gamma}{p_\alpha^{\gamma-1}}(x)dx - \frac{1}{\gamma}\frac{1}{n}\sum_{i=1}^n\left[\frac{p_\theta}{p_\alpha}\right]^\gamma(y_i) \\ &= \arg\sup_{\alpha\in\Phi} -\frac{1}{a+1}\int \frac{p_\alpha^{a+1}}{p_\theta^a}(x)dx + \frac{1}{a}\frac{1}{n}\sum_{i=1}^n\left[\frac{p_\alpha}{p_\theta}\right]^a(y_i) \\ &= \arg\inf_{\alpha\in\Phi} \int \frac{p_\alpha^{a+1}}{p_\theta^a}(x)dx - \frac{1+a}{a}\frac{1}{n}\sum_{i=1}^n\left[\frac{p_\alpha}{p_\theta}\right]^a(y_i) \\ &= \arg\inf_{\alpha\in\Phi} \mathbb{E}_{P_\alpha}\left[\left(\frac{p_\alpha}{p_\theta}\right)^a\right] - \frac{a+1}{a}\mathbb{E}_{P_n}\left[\left(\frac{p_\alpha}{p_\theta}\right)^a\right]. \end{aligned} \tag{1.3.9}$$

By comparing (1.3.8) and (1.3.9), we can deduce that the DφDE seems to be a penalized form of the MDPD. This penalization by a density p_θ creates a big trouble from a robustness point of view. The robustness of the DφDE is now not only controlled by the divergence power $a = -\gamma$ but also through p_θ. We have seen in the previous paragraph that the robustness of the DφDE in a two-component Gaussian mixture varies according to the position of θ with respect to θ^T the true vector of parameters. The difficulty of the choice of this escort parameter constitutes the only drawback of the DφDE in comparison to the MDPD. In Broniatowski and Vajda [2012], the authors establish an interesting link between $\varphi-$divergences and the density power divergence (1.3.8), see their Theorem 4.1.2.

1.4 Limitations of the MDφDE and DφDE

1.4.1 The influence of the escort parameter on the robustness of the DφDE

The IF of the DφDE is given by (see Toma and Broniatowski [2011]):

$$\mathrm{IF}(y|\phi) = \left[\int J_f(x)p_{\phi^T}(x)dx\right]^{-1}\left[\int\left(\frac{p_\phi}{p_{\phi^T}}\right)^\gamma(x)\nabla_\phi p_{\phi^T}(x)dx - \left(\frac{p_\phi}{p_{\phi^T}}\right)^\gamma(y)\frac{\nabla_\phi p_{\phi^T}(y)}{p_{\phi^T}(y)}\right],$$

where:

$$f(\alpha, \phi, y) = \int \frac{p_\phi^\gamma}{p_\alpha^{\gamma-1}} p_\phi dx - \left[\frac{p_\phi}{p_\alpha}(y)\right]^\gamma.$$

Previous papers which discussed the choice of the escort parameter have either let the choice arbitrary in the region where the IF is bounded (Toma and Broniatowski [2011]), or proposed to use robust estimates for the escort parameters (Cherfi [2011] and Frýdlovà et al. [2012]). The first idea is very complicated since we have no idea about the true value of the parameters and a bad choice of the escort parameter even inside the region where the IF is bounded does not ensure a good result. In Frýdlovà et al. [2012] and Cherfi [2011], experimental results show that the DφDE in a Gaussian model is very close to the escort parameter and coincides with the escort parameter when the later is equal to the MLE. The last fact can be easily verified following the proof of Theorem 6 in Broniatowski [2014]. Indeed, one may show that the MLE is a zero of the estimating equation of the DφDE and has a definite negative Hessian matrix of the corresponding objective function. On the other hand, the use of a robust escort parameter is not always a good idea as we will show in the following two examples.

Example 1.4.1. Consider a two-component Gaussian mixture model. We will give some conditions on the escort parameter in order to make the IF bounded. The first term in the influence function is a matrix which is independent of y and is constant. Supposing that it is invertible, we investigate both the existence of the integral, which is also a constant, and the boundedness of the remaining term which depends on y. The integral exists since the the fraction is of order e^{ax} whereas the derivative is of order e^{-x^2}. Boundedness of the IF is therefore equivalent to the boundedness of the remaining term. One can show by simple limit calculus that the escort parameter needs to verify either of the following conditions according to the value of γ:

$$\mu_1 > \mu_1^T, \quad \mu_2 < \mu_2^T \qquad \text{if} \quad \gamma > 0; \tag{1.4.1}$$

$$\mu_1 < \mu_1^T, \quad \mu_2 > \mu_2^T \qquad \text{if} \quad \gamma < 0, \tag{1.4.2}$$

in order for the IF to be bounded. Simulation results show that the use of a robust escort parameter verifying the set of conditions (1.4.1, 1.4.2) leads to a more robust parameter than the escort. However, the use of a *robust* escort parameter which does *not* fulfill the set of conditions (1.4.1, 1.4.2) has a negative impact on the resulting estimator. In our simulations in Section 1.7, we have analyzed the mixture whose true set of parameters is $(\lambda^T = 0.35, \mu_1^T = -2, \mu_2^T = 1.5)$ where the dataset was contaminated by 10% of outliers, see paragraph 1.7.2 for more details. We used our new MDφDE, defined in Section 1.5, as an escort parameter $\hat{\phi}_1$ which is robust. The divergence criterion is the Hellinger divergence which corresponds to $\gamma = 0.5$. Thus, we are in the context of condition (1.4.1). The new MDφDE verifies this condition and the resulting DφDE has a better error, see table 1.4.1 here below. In the same table, we give another escort parameter $\hat{\phi}_2$ which is as good as the previous one based on the total variation distance (see Section 1.7 for the definition), and even slightly better. If we calculate the DφDE using the escort parameter $\hat{\phi}_2$ which clearly does not verify condition (1.4.1), the error is nearly doubled.

Example 1.4.2. Let p_ϕ be a generalized Pareto distribution:

$$p_{\nu,\sigma}(y) = \frac{1}{\sigma}\left(1 + \nu\frac{y}{\sigma}\right)^{-1-\frac{1}{\nu}}, \quad \text{for } y \geq 0.$$

Estimator	Total variation
$\hat{\phi}_1 = (\hat{\lambda} = 0.349, \hat{\mu}_1 = -1.767, \hat{\mu}_2 = 1.377)$	0.087
$\hat{\phi}_2 = (\hat{\lambda} = 0.36, \hat{\mu}_1 = -2.2, \hat{\mu}_2 = 1.7)$	0.079
DφDE($\hat{\phi}_1$)	0.076
DφDE($\hat{\phi}_2$)	0.115

Table 1.1: The influence of a robust escort parameter on the DφDE in a mixture of two Gaussian components. The error is calculated between the true distribution and the estimated one, see Sec. 1.7

The shape and the scale are supposed to be unknown and equal to $\nu^T = 0.7, \sigma^T = 3$. It is necessary for the IF of the DφDE to be bounded[7] following the value of γ to locate the shape of the escort parameter with respect to the true value of the shape parameter. If $\gamma \in (0,1)$, it is necessary for the IF to be bounded that $\nu < \nu^T$. If $\gamma < 0$, then the IF can be bounded whenever $\nu > \nu^T$. Our simulation results in paragraph 1.7.3 show that for $\gamma = 0.5$, the DφDE calculated using a robust escort parameter (our kernel-based MDφDE) has deteriorated the performance significantly. The total variation distance corresponding to the escort parameter is 0.05 whereas the total variation distance corresponding to the DφDE is 0.12.

The past two examples[8] form an opposed result to the conjecture of both articles Frýdlovà et al. [2012] and Cherfi [2011] about the use of robust escort parameter. The use of a robust escort is a gamble and does not guarantee a better estimator than the escort itself. Thus, we are taking a great risk by using the DφDE. Notice, finally, that the DφDE is still more robust than the MLE and the classical MDφDE even if the IF is not bounded. There is still a remedy, but we did not consider in our simulations yet. We may use several escort parameters and calculate for each of them the corresponding DφDE. Then, use a procedure to combine the results of such estimators. Lavancier and Rochet [2016] provide a way to combine several estimators in order to obtain a better one by searching for the "best linear" combination between initial estimates.

1.4.2 Lack of robustness of the MDφDE

Unboundedness of the IF The influence function of the MDφDE is given by (see Toma and Broniatowski [2011] or Broniatowski and Vajda [2012]):

$$\mathrm{IF}(y) = \left[\int \frac{\nabla_\phi p_{\phi^T}(x). \left(\nabla_\phi p_{\phi^T}(x)\right)^t}{p_{\phi^T}(x)} dx \right]^{-1} \frac{\nabla_\phi p_{\phi^T}(y)}{p_{\phi^T}(y)}.$$

The matrix is constant, hence if we suppose that it is invertible, boundedness properties of the IF is determined by the fraction $\frac{\nabla_\phi p_{\phi^T}(y)}{p_{\phi^T}(y)}$. We will calculate this fraction in two examples; a mixture of Gaussian distributions and a mixture of Weibull distributions. The fraction is unbounded in both examples. Besides, it is immediate to see that the same conclusion holds in an exponential family model.

Example 1.4.3. Consider the mixture of two Gaussian components

$$p_{(\lambda,\mu_1,\mu_2)}(y) = \lambda \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_1)^2} + (1-\lambda) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu_1)^2}.$$

[7] The IF contains an inverse of a 2×2 matrix which cannot be simply calculated. Since it is a mere constant, we only discussed the other terms in the IF.

[8] See the remaining of the simulations for more examples.

We have

$$\frac{\nabla_\phi p_{\phi^T}(y)}{p_{\phi^T}(y)} = \begin{bmatrix} \frac{e^{-\frac{1}{2}(y-\mu_1^T)^2} - e^{-\frac{1}{2}(y-\mu_2^T)^2}}{\lambda^T e^{-\frac{1}{2}(y-\mu_1^T)^2} + (1-\lambda^T)e^{-\frac{1}{2}(y-\mu_1^T)^2}} \\ \frac{\lambda^T(y-\mu_1)e^{-\frac{1}{2}(y-\mu_1^T)^2}}{\lambda^T e^{-\frac{1}{2}(y-\mu_1^T)^2} + (1-\lambda^T)e^{-\frac{1}{2}(y-\mu_1^T)^2}} \\ \frac{(1-\lambda^T)(y-\mu_2)e^{-\frac{1}{2}(y-\mu_2^T)^2}}{\lambda^T e^{-\frac{1}{2}(y-\mu_1^T)^2} + (1-\lambda^T)e^{-\frac{1}{2}(y-\mu_1^T)^2}} \end{bmatrix} = \begin{bmatrix} \frac{1-e^{(\mu_2^T-\mu_1^T)y+\frac{1}{2}(\mu_1^T)^2-\frac{1}{2}(\mu_2^T)^2}}{\lambda^T+(1-\lambda^T)e^{(\mu_2^T-\mu_1^T)y+\frac{1}{2}(\mu_1^T)^2-\frac{1}{2}(\mu_2^T)^2}} \\ \frac{\lambda^T(y-\mu_1)}{\lambda^T+(1-\lambda^T)e^{(\mu_2^T-\mu_1^T)y+\frac{1}{2}(\mu_1^T)^2-\frac{1}{2}(\mu_2^T)^2}} \\ \frac{(1-\lambda)^T(y-\mu_2)}{\lambda^T e^{(\mu_1^T-\mu_2^T)y+\frac{1}{2}(\mu_2^T)^2-\frac{1}{2}(\mu_1^T)^2+1-\lambda^T}} \end{bmatrix}.$$

Let's suppose that $\mu_1^T < \mu_2^T$. The first component of the previous vector is bounded at both plus and minus infinity. The second component is bounded at $+\infty$, whereas it has a $-\infty$ limit at $-\infty$. The third component is bounded at $-\infty$ whereas it has a $+\infty$ limit at $+\infty$. This shows that the IF of the MDφDE is unbounded.

Example 1.4.4. Consider the mixture of two Weibull components:

$$p_{(\lambda,\nu_1,\nu_2)}(x) = 2\lambda\nu_1(2x)^{\nu_1-1}e^{-(2x)^{\nu_1}} + (1-\lambda)\frac{\nu_2}{2}\left(\frac{x}{2}\right)^{\nu_2-1}e^{-\left(\frac{x}{2}\right)^{\nu_2}}.$$

We calculate the fraction $\frac{\nabla_\nu p_{\nu^T}(y)}{p_{\nu^T}(y)}$.

$$\frac{\nabla_\nu p_{\nu^T}(x)}{p_{\nu^T}(x)} = \begin{bmatrix} \frac{2\nu_1^T(2x)^{\nu_1^T-1}e^{-(2x)^{\nu_1^T}} - \frac{\nu_2^T}{2}\left(\frac{x}{2}\right)^{\nu_2^T-1}e^{-\left(\frac{x}{2}\right)^{\nu_2^T}}}{2\lambda^T\nu_1^T(2x)^{\nu_1^T-1}e^{-(2x)^{\nu_1^T}} + (1-\lambda^T)\frac{\nu_2^T}{2}\left(\frac{x}{2}\right)^{\nu_2^T-1}e^{-\left(\frac{x}{2}\right)^{\nu_2^T}}} \\ \frac{2\lambda^T\left(1+\nu_1^T\log(2x)-\nu_1^T\log(2x)(2x)^{\nu_1^T}\right)(2x)^{\nu_1^T-1}e^{-(2x)^{\nu_1^T}}}{2\lambda^T\nu_1^T(2x)^{\nu_1^T-1}e^{-(2x)^{\nu_1^T}} + (1-\lambda^T)\frac{\nu_2^T}{2}\left(\frac{x}{2}\right)^{\nu_2^T-1}e^{-\left(\frac{x}{2}\right)^{\nu_2^T}}} \\ \frac{\frac{1-\lambda^T}{2}\left(1+\nu_2^T\log(2x)-\nu_2^T\log(2x)(2x)^{\nu_2^T}\right)(2x)^{\nu_2^T-1}e^{-(2x)^{\nu_2^T}}}{2\lambda^T\nu_1^T(2x)^{\nu_1^T-1}e^{-(2x)^{\nu_1^T}} + (1-\lambda^T)\frac{\nu_2^T}{2}\left(\frac{x}{2}\right)^{\nu_2^T-1}e^{-\left(\frac{x}{2}\right)^{\nu_2^T}}} \end{bmatrix}.$$

The second component is clearly unbounded neither near zero (it is of order $\log(2x)$) nor at infinity (it is of order $e^{(2x)^{\nu_2^T}-(2x)^{\nu_1^T}}$). Hence the IF of the MDφDE is unbounded for the mixture of Weibull distribution.

Equality with MLE in exponential families. An important aspect about the classical MDφDE is that it coincides with the maximum likelihood estimator in full exponential models whenever the corresponding *true* divergence D_φ is finite, see Broniatowski [2014]. This covers the standard Gaussian model for which Frýdlovà et al. [2012] provided clear robust properties of the MDφDE when outliers are generated by the standard Cauchy distribution. This contradicts with the theoretical result presented in Broniatowski [2014] which is an exact one and depends only on analytic arguments. We have done similar simulations and found out that numerical problems may play a role here. Generally, such problems come from numerical approximations such as numerical integration. In a Gaussian model, all integrals in (1.3.5) have close formulas and easy to calculate, see Frýdlovà et al. [2012] or Broniatowski and Vajda [2012]. However, when using the standard Cauchy distribution to generate outliers, we get points with very large values superior to 100. These points participate only in the sum term in the MDφDE (1.3.5). A Gaussian density with parameters not very far from the standard ones ($\mu = 0, \sigma = 1$) will produce a value equal to 0 in numerical computer programs. Thus, numerical problems of the form

0/0 would appear when calculating the sum term in (1.3.5) since the summand is of the form $g(p_\theta/p_\alpha)(y_i)$. If one uses simple practical solutions to avoid this, such as adding a very small value (e.g. 10^{-100}) to the denominator or the nominator, a thresholding effect is produced and the *true* fraction is badly calculated. As a result, such outliers would have practically no effect in the procedure as if they were not added, and one would obtain "forged robust estimates". The same thresholding effect does not happen in the MLE since the likelihood function does not contain any fractions. On the other hand, if one calculates the fraction using the properties of the exponential function, i.e. $p_\theta(y_i)/p_\alpha(y_i) = \exp[(y_i-\alpha)^2/2 - (y_i-\phi)^2/2]$, the MD$\varphi$DE defined by (1.3.6) gives the same result as the maximum likelihood estimator and never better even with Cauchy contamination.
We have performed further simulations on several models which do not belong to the exponential family and found out that the MDφDE have a very similar behavior to the MLE, see Section 1.7 below. Papers such as Barron and Sheu [1991] discussed how one can estimate a probability density using exponential families and proved interesting convergence rates. Such paper can explain partially our claim passing by the result in Broniatowski [2014].

Non robustness of the MDφDE under outliers or, more generally, under misspecification can be explained. When P_T is a member of the model, the approximated dual formula converges to the $\varphi-$divergence, and the argument of the infimum to the corresponding one, as the number of observations increases, see Proposition 3.1 in Broniatowski and Keziou [2009b]. This result, however, does not hold when P_T is not a member of the model, i.e. under contamination or misspecification. Indeed, consistency is a consequence of the following limit:

$$\hat{D}_n(P_\phi, P_T) \to \sup_{\alpha\in\Phi}\left\{\int \varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x)p_\phi(x)dx - \int \varphi^{\#}\left(\frac{p_\phi}{p_\alpha}\right)(y)dP_T(y)\right\},$$

together with the fact that the arginf of the left hand side converges to the arginf of the right hand side. However, the limiting quantity is the dual representation of the $\varphi-$divergence, and since the equality in (1.3.4) holds uniquely when $p_\alpha = dP_T/dy$ (otherwise there is inequality) then it is never attained as long as P_T is not a member of the model. Moreover, the limiting quantity is a lower bound of the divergence and minimizing the former does not guarantee the minimization of the later. Figure 1.2 represents this idea on a standard Gaussian model where the mean is unknown and the standard deviation is fixed at 1 and is known. The true distribution is then contaminated by a Gaussian distribution $\mathcal{N}(\mu = 10, \sigma = 2)$. Thus P_T has the density $0.9\mathcal{N}(\mu = 0, \sigma = 1) + 0.1\mathcal{N}(\mu = 10, \sigma = 2)$. The model p_ϕ is a Gaussian model $\mathcal{N}(\mu, 1)$. Taking the Hellinger divergence, $\varphi(t) = (\sqrt{t}-1)^2/2$, we plot the dual $\varphi-$divergence formula (1.3.4) in Fig(a) and its empirical version (1.3.5) in Fig(b) using a 100-sample drawn from P_T. We also plot the true values of the Hellinger divergence calculated using formula (1.1.1). The minimum of the dual representation is attained at approximately $\mu = 1$ whereas it is attained at approximately 0 for the true divergence. The curve of the dual representation is almost all the time below the curve of the true divergence. We also included in the figures the alternative dual formula introduced in the following paragraph which overcomes this problem.

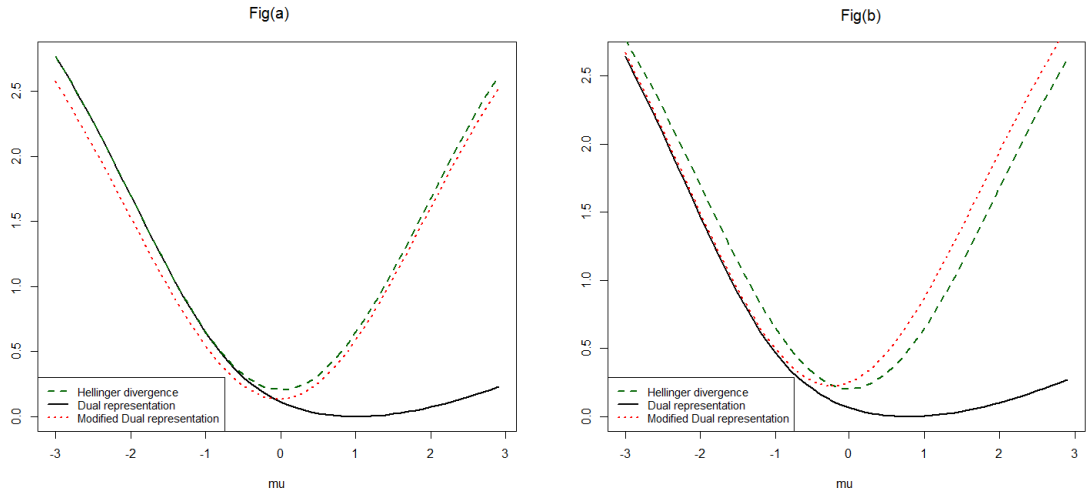

Figure 1.2: Underestimation caused by the classical dual representation compared to the new one. The true distribution is taken to be $0.9\mathcal{N}(\mu = 0, \sigma = 1) + 0.1\mathcal{N}(\mu = 10, \sigma = 2)$. Figure (a) shows the dual representation defined by (1.3.4) in comparison with the new reformulation defined by (1.5.1). Figure (b) shows the corresponding approximations when we replace the true distribution by its empirical version

1.5 A new robust estimator: kernel-based dual formula

1.5.1 New reformulation of the dual representation

As stated previously, if the model $(p_\alpha)_\alpha$ does not contain the true distribution p_T, the supremum in the dual formula is no longer attained and formula (1.3.4) is no longer an identity because the right hand side underestimates the divergence $D_\varphi(p_\phi, p_T)$.
An intuitive solution is to replace p_α by some adaptive (nonparametric) estimator of p_T which does not take into account the restriction of being in the model. In the resulting dual representation the supremum is attained whether we are under the model or not and we have equality between the dual representation and the $\varphi-$divergence. This way, the resulting criterion should inherit robustness properties against possible contamination as it approximates a $\varphi-$divergence.
One should be able to propose many solutions which correspond to this idea in order to reach a supremal attainment in the dual representation which may vary depending on the situation. For example, if we face a proportion of large-values outliers, one may add an extra component to p_α, i.e. replace p_α with the mixture $\lambda p_\alpha + (1-\lambda)q_\theta$. The extra component covers the outliers part in a smooth way. This suggestion is still very specific and treats only the case of contaminated data. Any nonparametric estimator of p_T can be used whose parameters may be determined automatically in the supremum calculus[9].
We propose here to use a kernel density estimator. In what follows $K_{n,w}$ denotes a kernel estimator of p_T defined using a symmetric or asymmetric kernel with or without bias-correction treatment.
For the definition of the new estimator, let $y_1, \cdots, y_n$ be an i.i.d. sample drawn from the probability law P_T. The number of observations n is fixed here. More formally, define the following class of functions $\mathcal{F}_{\phi,n} = \{\varphi'(p_\phi/K_{n,w}), w > 0\}$. The dual representation is now

[9]These parameters can be a window for a kernel estimator or parameters of a limited development in a suitable basis of functions, see Barron and Sheu [1991] for some examples of such approaches.

given by:

$$D_{\varphi}^{\mathcal{Y}_n}(p_\phi, p_T) = \sup_{w>0} \left\{ \int \varphi' \left(\frac{p_\phi}{K_{n,w}} \right)(x) p_\phi(x) dx - \int \varphi^{\#} \left(\frac{p_\phi}{K_{n,w}} \right)(y) p_T(y) dy \right\} \tag{1.5.1}$$

under a similar condition to (1.3.3) which is given by,

$$\int \left| \varphi' \left(\frac{p_\phi}{K_{n,w}} \right)(x) \right| p_\phi(x) dx < \infty, \quad \forall w > 0, \forall \phi \in \Phi. \tag{1.5.2}$$

We avoided to write $D_\varphi(p_\phi, p_T)$ in formula (eqn:NewExactDualForm), because the new dual formula may not ensure equality with the $\varphi-$divergence, but only a good approximation using a sample $\mathcal{Y}_n = \{Y_1, \cdots, Y_n\}$. A "good" choice of the window w_{opt} would yield[10]:

$$D_\varphi(p_\phi, p_T) \approx \int \varphi' \left(\frac{p_\phi}{K_{n,w_{\text{opt}}}} \right)(x) p_\phi(x) dx - \int \varphi^{\#} \left(\frac{p_\phi}{K_{n,w_{\text{opt}}}} \right)(y) p_T(y) dy.$$

Replace p_T by its empirical version. Our final approximation is given by:

$$\hat{D}_\varphi(p_\phi, p_T) = \int \varphi' \left(\frac{p_\phi}{K_{n,w_{\text{opt}}}} \right)(x) p_\phi(x) dx - \frac{1}{n} \sum_{i=1}^{n} \varphi^{\#} \left(\frac{p_\phi}{K_{n,w_{\text{opt}}}} \right)(y_i). \tag{1.5.3}$$

We avoid any indexation with respect to the sample or to n for the sake of clarity. Define now the new minimum dual $\varphi-$divergence estimator by:

$$\hat{\phi}_n = \arg\inf_{\phi \in \Phi} \int \varphi' \left(\frac{p_\phi}{K_{n,w_{\text{opt}}}} \right)(x) p_\phi(x) dx - \frac{1}{n} \sum_{i=1}^{n} \varphi^{\#} \left(\frac{p_\phi}{K_{n,w_{\text{opt}}}} \right)(y_i). \tag{1.5.4}$$

In comparison to the MDφDE defined by (1.3.6), we have removed the internal optimization procedure leaving only one simple optimization which keeps our procedure at the same level of complexity as other estimation procedures such as Beran's approach (Beran [1977]) and its generalization, and Basu et al. [1998].
An important question which arises now is: what should be the value of w_{opt} since its calculus demands knowing the true distribution? In the literature on kernel estimation, there exists many rules (automatic or not) to determine sub-optimum windows such as Silverman's (or Scott's) rule-of-thumb, cross-validation methods, etc; see for example Venables and Ripley [2013] Chap 5. Figure 1.2 shows in a Gaussian example contaminated by a Gaussian component $\mathcal{N}(10, 2)$ the use of Silverman's rule with a Gaussian kernel. The classical dual representation clearly underestimates the true divergence whereas the new reformulation stays close to it.
It is important to point out that the optimization problem in (1.5.4) is in general not convex. This is the case of the general class of divergence-(or disparity-)based estimators. Thus, we need to use a numerical optimization algorithm in order to calculate our kernel-based MDφDE, see Section 1.7 for more details.

[10] Recall that the supremum in (1.5.1) is attained at a window for which $K_{n,w}$ is as close as possible to p_T. Thus, a good choice of the window should result in a kernel estimator close to p_T and could be a good guess to the argument of the supremum in equation (1.5.1).

Remark 1.5.1. The new MDφDE keeps the MLE as a member of its class for the choice of $\varphi(t) = -\log(t) + t - 1$. Indeed, $\varphi'(t) = -1/t + 1$ and $t\varphi'(t) - \varphi(t) = \log(t)$. Thus,

$$\begin{aligned}\int \varphi'\left(\frac{p_\phi}{K_{n,w_{\text{opt}}}}\right)(x)p_\phi(x)dx &= 1,\\ \frac{1}{n}\sum_{i=1}^{n}\left[\frac{p_\phi}{K_{n,w_{\text{opt}}}}\varphi'\left(\frac{p_\phi}{K_{n,w_{\text{opt}}}}\right) - \varphi\left(\frac{p_\phi}{K_{n,w_{\text{opt}}}}\right)\right](y_i) &= \frac{1}{n}\sum_{i=1}^{n}\log\left(p_\phi\right) - \log\left(K_{n,w_{\text{opt}}}\right)(y_i).\end{aligned}$$

This entails that :

$$\begin{aligned}\hat{\phi}_n &= \arg\inf_{\phi\in\Phi} 1 - \frac{1}{n}\sum_{i=1}^{n}\log\left(p_\phi(y_i)\right) + \frac{1}{n}\sum_{i=1}^{n}\log\left(K_{n,w_{\text{opt}}}(y_i)\right)\\ &= \arg\sup_{\phi\in\Phi}\frac{1}{n}\sum_{i=1}^{n}\log\left(p_\phi(y_i)\right)\\ &= \text{MLE}.\end{aligned}$$

Remark 1.5.2. In the spirit of our approach, one can write a dual formula for Beran's approach in the case of the Hellinger divergence or more generally for any $\varphi-$divergence, see paragraph 1.2.1. Consider the class of functions $\mathcal{F}_{\phi,n} = \{\varphi'(p_\phi/K_{n,w}), w > 0\}$, then by (1.5.1) we can write:

$$\begin{aligned}D_\varphi\left(p_\phi, K_{n,w_0}\right) &= \sup_{w>0}\left\{\int\varphi'\left(\frac{p_\phi}{K_{n,w}}\right)(x)p_\phi(x)dx - \int\varphi^{\#}\left(\frac{p_\phi}{K_{n,w}}\right)(y)K_{n,w_0}(y)dy\right\}\\ &= \int\varphi'\left(\frac{p_\phi}{K_{n,w_0}}\right)(x)p_\phi(x)dx - \int\varphi^{\#}\left(\frac{p_\phi}{K_{n,w_0}}\right)(y)K_{n,w_0}(y)dy,\end{aligned}$$

where w_0 is a window calculated using an automatic rule as mentioned here above for w_{opt}. The only difference with the kernel-based dual formula (1.5.3) is that we are integrating function $\varphi^{\#}\left(\frac{p_\phi}{K_{n,w_0}}\right)$ with respect to the empirical distribution instead of a smoothed version of it.

1.6 Asymptotic properties and robustness of the new kernel-based MDφDE

We present in this section some of the asymptotic properties of the new MDφDE defined by (1.5.4). We use Theorem 5.7 from van der Vaart [1998] which we restate here. Consistency of the kernel-based MDφDE means that $\hat{\phi}_n$ defined by (1.5.4) converges in probability to ϕ^T the true vector of parameters when we are under the model, i.e. $P_T = P_{\phi^T}$. If we are not under the model, consistency holds towards the projection of P_T on the model in the sens of the divergence. In other terms, the projection P_{ϕ^T} is the member of the model P_ϕ whose parameters are defined by $\phi^T = \arg\inf_{\phi\in\Phi} D_\varphi(P_\phi, P_T)$.
Similarly to Basu and Lindsay [1994], there are some cases (which are rare) such as the location Gaussian model in which consistency of the kernel-based MDφDE does not require any condition on the kernel window. Thus, one may find simpler versions of the results we give below. We will be however interested in the general situation where the window needs to converge towards zero at a certain rate.
In a second part of this section, we calculate the limiting law of the new estimator under strong but standard assumptions. We, finally, calculate the influence function of the

kernel-based MDφDE for a fixed window, and show how the use of a kernel estimate in place of the model p_α in the dual formula (1.3.4) interferes to make the IF bounded.
We use the same notations as in van der Vaart [1998] to denote integration. Thus, if f is a $P-$integrable function, we denote Pf the integral $\int f dP$. Moreover, $K_w * P$ denotes the operation of smoothing dP by the kernel K_w with bandwidth equal to w. This smoothing can be done by simple convolution as in the case of Rosenblatt-Parzen kernel estimator. Other kinds of smoothing are presented in Section 1.2.2. In this section only, the smoothing is supposed to be an additive operator in the sense that $K_w * (P \pm Q) = K_w * P \pm K_w * Q$.

1.6.1 Consistency

Theorem 5.7 from van der Vaart [1998] permits to treat the consistency of a general class of M-estimates. It is stated as follows:

Theorem 1.6.1. *Let M_n be random functions and let M be a fixed function of ϕ such that for every $\varepsilon > 0$*

$$\sup_{\phi\in\Phi} |M_n(\phi) - M(\phi)| \xrightarrow{\mathbb{P}} 0, \tag{1.6.1}$$

$$\inf_{\phi:\|\phi-\phi^T\|\geq\varepsilon} M(\phi) > M(\phi^T). \tag{1.6.2}$$

Then any sequence of estimators $\hat{\phi}_n$ with $M_n(\hat{\phi}_n) \leq M_n(\phi^T) - o_P(1)$ converges in probability to ϕ^T.

In our approach, function M_n corresponds to the criterion function $P_n H(P_n, \phi)$, where $H(P_n, \phi, y)$ is defined by:

$$H(P_n, \phi, y) = \int \varphi'\left(\frac{p_\phi}{K_w * P_n}\right)(x) p_\phi(x) dx - \varphi^{\#}\left(\frac{p_\phi(y)}{K_w * P_n(y)}\right).$$

Function M is simply defined by the *expected*[11] limit in probability of M_n, since the Law of Large Numbers cannot be used because the average term is not a sum of i.i.d. random variables. It is given by $P_T h(P_T, \phi)$ where $h(P_T, \phi, y)$ is defined as:

$$h(P_T, \phi, y) = \int \varphi'\left(\frac{p_\phi}{p_T}\right)(x) p_\phi(x) dx - \varphi^{\#}\left(\frac{p_\phi}{p_T}\right)(y).$$

In order to prove (1.6.1), write:

$$\sup_{\phi\in\Phi} |P_n H(P_n, \phi) - P_T h(P_T, \phi)| \leq \sup_{\phi\in\Phi} |P_n H(P_n, \phi) - P_n h(P_T, \phi)| + \sup_{\phi\in\Phi} |P_T h(P_T, \phi) - P_n h(P_T, \phi)|. \tag{1.6.3}$$

Now, the second supremum tends to 0 in probability by the Glivenko-Cantelli theorem as soon as $\{h(P_T, \phi), \phi \in \Phi\}$ is a Glivenko-Cantelli class of functions, see van der Vaart [1998] Chap. 19 Section 2 and the examples therein. The problem then resides in finding conditions under which the first term tends to 0 in probability. The remaining of the paragraph will be concerned with the search for such conditions. In the whole section concerning the consistency of our new estimator, the window parameter w is supposed to depend on n in order to be able to use Theorem 1.6.1 without any modification. Besides,

[11] In the literal sense and not mathematically.

the construction of the estimator from (1.5.1) shows the explicit link of the window with n.
We next provide a set of sufficient conditions in order for the new estimator to be consistent. We treat the general class of $\varphi-$divergences in a first theorem. The result imposes strong but standard assumptions on the model. After that, We present a result for a subclass of $\varphi-$divergences with simpler conditions on the model. Some exceptional cases may be studied separately in order to deduce simpler conditions.

Remark 1.6.1. Large values of γ in absolute value are not interesting in general and may lead to practical complications. Values of γ greater than 1 leads to integrability problems in condition (1.5.2) for the new MDφDE and in (1.3.3) for the classical one in standard examples such as the scale Gaussian model, and the supremum in (1.3.4) is not well defined. The special case of $\gamma = 2$ which corresponds to the Pearson's χ^2 is included by this remark. This not very surprising. The Pearson's χ^2 is a very sensitive criterion and measures the relative error committed. Thus small errors committed at values where the distribution has small values will have the same influence as the values where distribution attributes a greater density.

An essential assumption will be the consistency of the kernel estimator. We refer to Wied and Weibßach [2012], Zambom and Dias [2013] or Libengue Dobele-kpoka [2013] Chap. 1 for a brief survey on symmetric kernels. When using asymmetric kernels, unfortunately consistency is proved only on every compact subset of the support of the distribution function, see Bouezmarni and Scaillet [2005] or Libengue Dobele-kpoka [2013] Chap. 3 for a more general approach. Thus, our proof does not cover these kernels.
Assumption (1.6.2) in Theorem 1.6.1 means that function $\phi \mapsto P_T h(P_T, \phi)$ has a unique and well separated minimum. Uniqueness is achieved when we are under the model ($P_T = P_{\phi^T}$) since function $\phi \mapsto P_T h(P_T, \phi)$ is non other than the dual representation (with the supremum calculated) of the $\varphi-$divergence $D_\varphi(P_\phi, P_{\phi^T})$. Using the property that $D_\varphi(p_\phi, p_{\phi^T}) = 0$ iff $p_\phi = p_{\phi^T}$, uniqueness is immediately verified as soon as the model is identifiable. If we are not under the model (misspecification), the projection of P_T on the model P_ϕ may not be unique and assumption (1.6.2) is still needed.

General Result

We will derive in this paragraph a result which concerns the general class of $\varphi-$divergences. It was difficult to build such result without imposing strong assumptions on the model. Hereafter, simpler conditions will be proved for the particular class of Cressie-Read functions φ_γ for $\gamma \in (-1, 0)$. As mentioned here above, using inequality (1.6.3), the difficult term is first one. It is given by:

$$P_n H(P_n, \phi) - P_n h(P, \phi) = \int \left[\varphi' \left(\frac{p_\phi}{K_w * P_n} \right) - \varphi' \left(\frac{p_\phi}{p_T} \right) \right] (x) p_\phi(x) dx \\ - \frac{1}{n} \sum_{i=1}^{n} \varphi^{\#} \left(\frac{p_\phi}{K_w * P_n} \right) (y_i) - \varphi^{\#} \left(\frac{p_\phi}{p_T} \right) (y_i).$$

The key idea is to treat each term (the integral and the sum) separately and prove its uniform convergence in probability towards 0. Another important step is to apply the mean value theorem in order to transfer the difference from functions φ' and $\varphi^{\#}$ into a difference between the kernel estimator and the true distribution where consistency of the former is exploited. The proof of the following theorem is differed to Appendix 1.8.1.

Theorem 1.6.2. *Assume that:*

1. *function $t \mapsto \varphi(t)$ is twice differentiable;*
2. *the kernel is defined on a compact and the kernel density estimator is consistent, i.e. $\sup_x |K_w * P_n(x) - p_T(x)| \to 0$ in probability;*
3. *the model p_ϕ is defined on a compact set and bounded independently of ϕ, and p_T is also defined on the same compact and bounded;*
4. *for any $\varepsilon > 0$, $\inf_{\phi:\|\phi-\phi^T\|\geq\varepsilon} P_T h(P_T, \phi) > P_T h(P_T, \phi^T)$,*

then the minimum dual $\varphi-$divergence estimator defined by (1.5.4) is consistent whenever it exists.

Notice that assumption 3 is strong but stays standard. It was already used in the literature, see for example Beran [1977]. The need to impose such restrictive assumptions stems from the nature of the dual formula which contains the quotient $p_\phi/K_{n,w}$ on the one hand. On the other hand, our approach which consists in using a data-based estimator results in sums of strongly dependent terms which cannot be treated in a simple and a general way. In the next paragraph, we treat a subclass of $\varphi-$divergences where we can control the terms of inequality (1.6.3) without the need to assumption 3.

Case of power divergences with $\gamma \in (-1, 0)$

Here we have:

$$P_n H(P_n, \phi) - P_n h(P, \phi) = \frac{1}{\gamma - 1} \int \frac{(K_w * P_n)^{1-\gamma} - p_T^{1-\gamma}}{p_\phi^{-\gamma}}(x)dx - \frac{1}{n\gamma} \sum_{i=1}^{n} \frac{(K_w * P_n)^{-\gamma} - p_T^{-\gamma}}{p_\phi^{-\gamma}}(y_i). \tag{1.6.4}$$

The key idea for proving the consistency is to use the uniform continuity of functions $t \mapsto t^{-\gamma}$ and $t \mapsto t^{(-\gamma+1)/2}$. The proof of the following theorem is differed to Appendix 1.8.2.

Theorem 1.6.3. *For the class of power divergences defined through the class of Cressie-Read functions φ_γ with $\gamma \in (-1, 0)$, assume that:*

1. *the kernel estimator is consistent, i.e. $\sup_x |K_w * P_n(x) - p_T(x)| \to 0$ in probability;*
2. *$\left\{ \left(\frac{p_\phi}{p_T}\right)^\gamma, \phi \in \Phi \right\}$ is a Glivenko-Cantelli class of functions;*
3. *there exists n_0 such that $\forall n \geq n_0$, the probability that the quantity*

$$\mathcal{A}_n = \sup_\phi \int \frac{(K_w * P_n)^{\frac{-\gamma+1}{2}}(x) + p_T^{\frac{-\gamma+1}{2}}(x)}{p_\phi^{-\gamma}(x)} dx$$

is upper bounded independently of n is greater than $1 - \eta_n$ for some $\eta_n \to 0$;

4. *there exists n_0 such that $\forall n \geq n_0$, the probability that the quantity*

$$\mathcal{B}_n = \sup_\phi \frac{1}{n} \sum_{i=1}^{n} p_\phi^\gamma(y_i)$$

is upper bounded independently of n is greater than $1 - \eta_n$ for some $\eta_n \to 0$;

5. *for any* $\varepsilon > 0$, $\inf_{\phi:\|\phi-\phi^T\|\geq\varepsilon} P_T h(P_T, \phi) > P_T h(P_T, \phi^T)$,

then the minimum dual $\varphi-$*divergence estimator defined by (1.5.4) is consistent whenever it exists.*

This result is clearly more general than the result of Theorem 1.6.2. We treat models which may be defined on the whole set $\mathbb{R}^d$. The assumptions are still accessible as we will demonstrate in the following example.

Example 1.6.1. We take a simple example of a Gaussian model with unknown mean $\phi = \mu$ which is supposed to be in a close interval $[\mu_{\min}, \mu_{\max}]$. We consider power divergences for which $\gamma \in (-1, 0)$. We use Theorem 1.6.3 to prove consistency. The Gaussian kernel is used. Assumption 1 is easily checked by considering the list of conditions in Theorem A in Silverman [1978]. Assumption 2 holds since

$$\frac{p_\phi^\gamma}{p_{\phi^T}^\gamma} p_{\phi^T}(x) = e^{-\frac{1}{2}x^2 - \mu y + \frac{1}{2}\mu^2}.$$

The verification of assumption 3 is technical, so that we let it to the end. For assumption 4, in order to study $\mathcal{B}_n$, it suffices to consider the quantity $\sup_\phi \int p_\phi^\gamma p_{\phi^T}$. Indeed, the Glivenko-Cantelli theorem states that both quantities $\sup_\phi \frac{1}{n}\sum_{i=1}^n p_\phi^\gamma(y_i)$ and $\sup_\phi \int p_\phi^\gamma p_{\phi^T}$ are uniformly close for sufficiently large n independently of ϕ, hence boundedness of either of them implies boundedness of the other. We have:

$$\int p_\phi^\gamma p_{\phi^T} = c_2 e^{-\frac{-\gamma}{1+\gamma}\frac{\mu^2}{2}}.$$

for some constant c_2. Here again, since μ is supposed to be in a closed interval, the previous quantity is bounded. This entails that $\mathcal{B}_n$ is bounded and assumption 4 is fulfilled.
We move now to assumption 5. By the dual representation of the divergence, we have $P_T h(P_T, \phi) = D_\varphi(p_\phi, p_{\phi^T})$. This implies that :

$$P_T h(P_T, \phi) = \frac{1}{\gamma(\gamma-1)} e^{\frac{\gamma^2-\gamma}{2}\mu^2} - \frac{1}{\gamma(\gamma-1)}.$$

This function clearly verifies assumption 5 since it has a minimum at $\mu = 0$ and this minimum is well separated.
We go back to assumption 3. The second term is given by:

$$\int \frac{p_{\phi^T}^{\frac{-\gamma+1}{2}}(x)}{p_\phi^{-\gamma}(x)} = c_1 e^{\frac{\gamma^2-\gamma}{2(1+\gamma)}\mu^2}$$

for some constant c_1. It is thus bounded since μ is supposed to be in a closed interval.

For the first term, let $\eta > 0$. By Jensen's inequality, we may write:

$$\begin{aligned}
\int \frac{(K_w * P_n)^{\frac{1-\gamma}{2}}(y)}{p_\phi^{-\gamma}(y)} dy &= \frac{(2\pi)^{\frac{1-\eta}{2}}}{\eta^{3/2}} \int \left(\frac{K_w * P_n}{p_\phi^{\frac{2\eta-\gamma}{1-\gamma}}} \right)^{\frac{1-\gamma}{2}} (y) \frac{1}{\sqrt{2\pi/\eta}} e^{-\eta \frac{(y-\mu)^2}{2}} dy \\
&\leq \frac{(2\pi)^{\frac{1-\eta}{2}}}{\eta^{3/2}} \left(\int \frac{K_w * P_n}{p_\phi^{\frac{2\eta-\gamma}{1-\gamma}}}(y) \frac{1}{\sqrt{2\pi/\eta}} e^{-\eta \frac{(y-\mu)^2}{2}} dy \right)^{\frac{1-\gamma}{2}} \\
&\leq (2\pi)^{\eta/2-\gamma-1/2} e^{\left(\frac{\eta-\gamma}{1-\gamma}-\frac{\eta}{2}\right)\mu^2} \times \left(\frac{1}{nw} \right. \\
&\left. \sum_{i=1}^{n} e^{-\frac{y_i^2}{2w^2}} \int e^{-\frac{1-\gamma-2(\eta-\gamma)w^2+\eta(1-\gamma)w^2}{2w^2(1-\gamma)} y^2 + \frac{(1-\gamma)y_i - 2w^2(\eta-\gamma)\mu + w^2(1-\gamma)\eta\mu}{w^2(1-\gamma)} y} dy \right)^{\frac{1-\gamma}{2}}.
\end{aligned}$$

We calculate each integral separately:

$$\int e^{-\frac{1-\gamma-2(\eta-\gamma)w^2+\eta(1-\gamma)w^2}{2w^2(1-\gamma)} y^2 + \frac{(1-\gamma)y_i - 2w^2(\eta-\gamma)\mu + w^2(1-\gamma)\eta\mu}{w^2(1-\gamma)} y} dy =$$

$$\sqrt{2\pi} w \sqrt{\frac{1-\gamma}{1-\gamma+(-\eta+2\gamma-\eta\gamma)w^2}} \exp\left[\frac{\left((1-\gamma)y_i - 2w^2(\eta-\gamma)\mu + w^2(1-\gamma)\eta\mu\right)^2}{2w^2(1-\gamma)\left(1-\gamma-2(\eta-\gamma)w^2+\eta(1-\gamma)w^2\right)} \right].$$

We now proceed to estimate the sum over i. First, the only important term in the precedent integral is the one with factor y_i^2. Therefore, we denote in the precedent integral c_2, c_1, c_0 respectively the coefficients of terms y_i^2, y_i and the constant term. We denote also c the constant before the exponential (without the w). We only give the form of c_2:

$$c_2 = \frac{1-\gamma}{2w^2\left(1-\gamma-2(\eta-\gamma)w^2+\eta(1-\gamma)w^2\right)}.$$

We now have:

$$\begin{aligned}
&\frac{1}{nw} \sum_{i=1}^{n} e^{-\frac{y_i^2}{2w^2}} \int e^{-\frac{1-\gamma-2(\eta-\gamma)w^2+\eta(1-\gamma)w^2}{2w^2(1-\gamma)} y^2 + \frac{(1-\gamma)y_i - 2w^2(\eta-\gamma)\mu + w^2(1-\gamma)\eta\mu}{w^2(1-\gamma)} y} dy \\
&\qquad = c\frac{1}{n} \sum_{i=1}^{n} \exp\left[\left(c_2 - \frac{1}{2w^2} \right) y_i^2 + c_1 y_i + c_0 \right] \\
&\qquad = c\frac{1}{n} \sum_{i=1}^{n} \exp\left[\frac{\eta - 2\gamma + \eta\gamma}{2\left(1-\gamma-2(\eta-\gamma)w^2+\eta(1-\gamma)w^2\right)} y_i^2 + c_1 y_i + c_0 \right].
\end{aligned}$$

The final step is to use a version of the law of large numbers for independent random variables such as the two series theorem of Kolomogrov (see [Feller, 1971] Chap VII, Theorem 3 page 238) since the terms of the sum do not have the same probability law, but guided by the standard Gaussian law. The general term of the sum is given by:

$$Z_i = \exp\left[\frac{\eta - 2\gamma + \eta\gamma}{2\left(1-\gamma-2(\eta-\gamma)w^2+\eta(1-\gamma)w^2\right)} y_i^2 + c_1 y_i + c_0 \right].$$

One can verify that the expectation of Z_i exists as soon as the following condition is fulfilled:

$$0 \leq \eta < 1 \qquad \text{and} \qquad \gamma > -1.$$

Indeed,

$$\mathbb{E}[Z_i] = \frac{1}{\sqrt{2\pi}} \int \exp\left[\frac{(\eta-1)(\gamma+1)+(\eta-2\gamma+\eta\gamma)w^2}{2\left(1-\gamma-2(\eta-\gamma)w^2+\eta(1-\gamma)w^2\right)}y^2 + (c_1+\mu)y + c_0 - \mu^2/2\right].$$

The dominating term in the integral is the one with y^2. It suffices then that the coefficient of y^2 to be negative so that the integral exists. We have

$$\frac{(\eta-1)(\gamma+1)+(\eta-2\gamma+\eta\gamma)w^2}{2\left(1-\gamma-2(\eta-\gamma)w^2+\eta(1-\gamma)w^2\right)} < 0$$

if the denominator is positive and the nominator is negative. The denominator is equal to $1-\gamma+(-\eta+(2-\eta)\gamma)w^2$. Suppose that $\eta \in (0,1)$, then the denominator is positive as soon as :

$$w^2 < \frac{1-\gamma}{\eta-(2-\eta)\gamma}. \tag{1.6.5}$$

On the other hand, the nominator is equal to $(\eta-1)(\gamma+1)+(\eta-2\gamma+\eta\gamma)w^2$. If $\eta \in (0,1)$, then the nominator is negative as soon as:

$$w^2 < \frac{(1-\eta)(1+\gamma)}{\eta-(2-\eta)\gamma}. \tag{1.6.6}$$

Combining this with (1.6.5) and since $(1-\eta)(1+\gamma) < 1-\gamma$, then if $\eta \in (0,1)$, the coefficient of y^2 is negative as soon as (1.6.6) is fulfilled. Recall that, both η and γ are fixed values which do not depend on n (the sample size). Moreover, the window w need to go to zero as n goes to infinity to ensure the consistency of the kernel density estimator. Thus condition (1.6.6) is fulfilled for any $\gamma \in (-1,0)$ as soon as $\eta \in (0,1)$.
The variance can be calculated similarly and proved to be finite under some condition to be identified. It suffices to calculate the second order moment. We have:

$$\mathbb{E}[Z_i^2] = \frac{1}{\sqrt{2\pi}} \int \exp\left[\frac{2\eta-3\gamma+2\eta\gamma-1+(\eta-2\gamma+\eta\gamma)w^2}{2\left(1-\gamma-2(\eta-\gamma)w^2+\eta(1-\gamma)w^2\right)}y^2 + (c_1+\mu)y + c_0 - \mu^2/2\right].$$

The coefficient of the dominating term is negative as soon as the denominator is positive (when $\eta \in (0,1)$) and the nominator is negative. This is translated into the following condition:

$$w^2 < \frac{-2\eta-2\eta\gamma+3\gamma+1}{\eta-2\gamma+\eta\gamma}.$$

The right hand side is positive only if:

$$\gamma > \frac{2\eta-1}{3-2\eta}.$$

Since $\eta \in (0,1)$ and the function $\eta \mapsto \frac{2\eta-1}{3-2\eta}$ is increasing, then possible values for γ where the variance is finite is $(-1/3, 0)$. It results that for $\gamma \in [-\frac{1}{3}, 0)$, the Kolomogrov's two series theorem applies and the average $\frac{1}{n}\sum Z_i$ now converges in probability. Besides, the remaining factor c also converges as n goes to infinity (and w goes to zero) to a constant (equal to 1). Thus, boundedness of $\int \frac{(K_w * P_n)^{\frac{1-\gamma}{2}}(y)}{p_\phi^{-\gamma}(y)} dy$ is ensured. The argument becomes uniform on μ since it is supposed to be inside a closed interval $[\mu_{\min}, \mu_{\max}]$. All assumptions of Theorem 1.6.3 are now verified, and the kernel-based MDφDE defined by (1.5.4) is consistent in the Gaussian model.

1.6.2 Asymptotic normality

In the literature on M-estimators, we study the asymptotic normality starting from the estimating equation, see van der Vaart [1998] Chap. 5 Section 5.3. The idea, then, consists in using a Taylor expansion. Keeping the same notation as in the previous paragraph, the estimating equation has the form:

$$\nabla P_n H(P_n, \hat{\phi}) = 0.$$

We apply a Taylor expansion on $\nabla P_n H(P_n, \phi)$ between $\hat{\phi}$ and ϕ^T.

$$\nabla P_n H(P_n, \hat{\phi}) = \nabla P_n H(P_n, \phi^T) + J_{P_n H(P_n, \phi^T)}(\hat{\phi} - \phi^T) + o_P\left(n^{-1/2}\right).$$

The left hand side is zero by definition of $\hat{\phi}$. The $o_P\left(n^{-1/2}\right)$ comes from a suitable control on the third derivatives of the objective function. If the matrix of second order derivatives $J_{P_n H(P_n, \phi^T)}$ converges in probability to an invertible matrix J, then, we may write:

$$\sqrt{n}(\hat{\phi} - \phi^T) = J^{-1}\sqrt{n}\nabla P_n H(P_n, \phi^T) + o_P(1). \tag{1.6.7}$$

The main problem resides in showing that $\sqrt{n}\nabla P_n H(P_n, \phi^T)$ has a multivariate Gaussian limit law. In the case of our kernel-based MDφDE, the vector $\sqrt{n}\nabla P_n H(P_n, \phi^T)$ is not a sum of i.i.d. terms (the case of M-estimates). It contains the difficulty of the case of divergences approximated by replacing the true distribution by a kernel density estimator (Beran [1977] or Park and Basu [2004]). Besides, there is a sum of strongly dependent terms making the treatment of this vector very complicated in a general setup. Strong, but standard, conditions are needed in order to study the asymptotic distribution. We only study the case of univarite densities. The general case of multivariate densities is more complicated, because there should be a correction term similarly to Tamura and Boos [1986]. We leave this part to a future work.
The following result covers all power divergences, i.e. $\varphi-$divergences with $\varphi = \varphi_\gamma$ with $\gamma \in \mathbb{R} \setminus \{0, 1\}$. The proof is differed to Appendix 1.8.3.

Theorem 1.6.4. *For the class of power divergences with $\varphi = \varphi_\gamma$ for $\gamma \neq \{0, 1\}$, assume that:*

1. *the kernel-based MDφDE $\hat{\phi}$ is consistent;*
2. *the kernel K is symmetric and has a compact support where it is of class $\mathcal{C}^1$. Moreover $z^2 K(z)$ is integrable;*
3. *the density p_{ϕ^T} is defined on a compact, is positive, bounded and twice derivable such that p''_{ϕ^T} is bounded. Moreover, there exists a neighborhood of ϕ^T such that the partial derivatives up to third order with respect to ϕ are bounded;*
4. *$\frac{n^{1/2}w}{-\log w} \to \infty$ and $n^{1/2}w^2 \to 0$,*

then,

$$\sqrt{n}(\hat{\phi} - \phi^T) \underset{\mathcal{L}}{\to} \mathcal{N}\left(0, (2\gamma^2 + 1)J^{-1}S\left(J^{-1}\right)^t\right), \tag{1.6.8}$$

where $S = \int \nabla p_{\phi^T} \nabla p_{\phi^T}^t$.

Assumptions on the model in this theorem are restrictive, but stay standard. Similar conditions were considered in the study of the rates of convergence of the kernel density estimation, see Wied and Weibßach [2012]. They were also considered in the study of asymptotic properties of $\varphi-$divergences, see Beran [1977] Theorem 4. Furthermore, conditions on the bandwidth are verified for $w = n^{-1/4-\delta}$ for $\delta \in (0, 1/8)$, see Bordes and Vandekerkhove [2010] Remark 3.1. Notice that even with such strong assumptions, the proof is not simple and demands several techniques and results from the theory on kernel density estimation.

Remark 1.6.2. Under the assumptions of Theorem 1.6.4, consistency of the kernel-based MDφDE is ensured by Theorem 1.6.2 provided the differentiability of φ up to second order (assumption 1) and the uniqueness and well separability of the minimum (assumption 4). Notice that the consistency of the kernel estimator (assumption 2) is also fulfilled, see Wied and Weibßach [2012] Theorem 2.

Remark 1.6.3. The condition on the existence and boundedness of p''_{ϕ^T} can be relaxed to being Lipschitz function. This demands however more assumptions on the bandwidth, see lemma 3.1 in Bordes and Vandekerkhove [2010].

1.6.3 Influence Function for a given window

In practice, the choice of the window is based on methods such as cross-validation, Gaussian approximations or even based on personal experience. Thus, it is interesting to study the robustness properties supposing that the window is generated by an external tool. Although in practice, we estimate the window on the basis of an observed dataset, this creates a complication in the definition and the calculation of the influence function. For this reason, we suppose that the window is fixed and independent of n.
The influence function (IF), although being limited to the existence of a noise-component, is easy to calculate in general[12] and gives an aspect of the robustness of an estimator whenever the IF is bounded. We derive in this paragraph the influence function of the new MDφDE for the class of power divergences. The general case of function φ seems to give an incomprehensible formula, and is not as interesting as the case of power divergences.

We recall the definition of the IF. Let C be a functional which gives for a probability distribution P the estimator corresponding to the argument of the infimum of $PH(P, \phi)$ defined earlier, i.e.

$$C(P) = \arg\inf_{\phi\in\Phi} \int \varphi'\left(\frac{p_\phi}{K_w * P}\right)(x)p_\phi(x)dx - \int \varphi^{\#}\left(\frac{p_\phi(y)}{K_w * P(y)}\right)dP(x).$$

Hence, $C(P_n)$ is non other than the estimator given by (1.5.4) for a given w. Fisher consistency is translated by $C(P_{\phi^T}) = \phi^T$. This is unfortunately not verified in general when the window is supposed to be calculated by an external tool, because the dual formula is a priori a lower bound of $D_\varphi(P_\phi, P_{\phi^T})$, and we cannot be sure that it would verify the same identifiability property, i.e. $D(Q, P) = 0$ iff $P = Q$ whenever φ is strictly convex. Example 1.6.1 shows, however, a case where Fisher consistency is attained for any value of the window w.
The influence function measures the impact of a small perturbation in the distribution P

[12] This is regardless of the theoretical justifications of its existence.

on the resulting estimator. It is hence defined by:

$$\mathrm{IF}(P,Q) = \lim_{\varepsilon\to 0} \frac{C\left((1-\varepsilon)P+\varepsilon Q\right) - C(P)}{\varepsilon}.$$

We generally detect the influence of an outlier x_0 by observing what happens when we replace P by $(1-\varepsilon)P + \varepsilon\delta_{x_0}$.
In the literature on M-estimates, one may derive the IF from the estimating equation. For power divergences, the estimating equation corresponding to P is given by:

$$\frac{\gamma}{\gamma-1}\int \frac{p_{C(P)}^{\gamma-1}\nabla p_{C(P)}}{(K_w * P)^{\gamma-1}}(x)dx = \int \frac{p_{C(P)}^{\gamma-1}\nabla p_{C(P)}}{(K_w * P)^{\gamma}}(x)dP(x), \tag{1.6.9}$$

where the gradient is calculated with respect to ϕ. The influence function is obtained by "derivation"[13] of the two sides with respect to ε after having replaced P by $(1-\varepsilon)P+\varepsilon Q$. Denote J_{p_ϕ} the matrix of second derivatives of p_ϕ with respect to ϕ. The following result gives the formula of the IF for power divergences when the noise is generated by an arbitrary distribution Q or when an outlier is present. The proof is differed to Appendix 1.8.4.

Theorem 1.6.5. *The influence function of the kernel-based MDφDE defined by (1.5.4) for a given window is given by:*

$$IF(P_T,Q) = \gamma A^{-1}\int \frac{p_{C(P_T)}^{\gamma-1}\left[K_w * Q\right]\nabla p_{C(P_T)}}{(K_w * P_T)^{\gamma}}\left(1-\frac{p_T}{K * P_T}\right)(x)dx + A^{-1}\int \frac{p_{C(P_T)}^{\gamma-1}\nabla p_{C(P_T)}}{(K_w * P_T)^{\gamma}}(x)dQ(x). \tag{1.6.10}$$

If C is Fisher consistent, i.e. $C(P_T) = \phi^T$, then the influence function is given by:

$$IF(P_T,Q) = \gamma A^{-1}\int \frac{p_{\phi^T}^{\gamma-1}\left[K_w * Q\right]\nabla p_{\phi^T}}{(K_w * P_T)^{\gamma}}\left(1-\frac{p_T}{K * P_T}\right)(x)dx + A^{-1}\int \frac{p_{\phi^T}^{\gamma-1}\nabla p_{\phi^T}}{(K_w * P_T)^{\gamma}}(x)dQ(x). \tag{1.6.11}$$

Finally, if $Q = \delta_{x_0}$, then the IF is given by:

$$IF(P_T,x_0) = \gamma A^{-1}\int \frac{p_{C(P_T)}^{\gamma-1}\left[K_w * \delta_{x_0}\right]\nabla p_{C(P_T)}}{(K_w * P_T)^{\gamma}}\left(1-\frac{p_T}{K_w * P_T}\right)(x)dx + A^{-1}\frac{p_{C(P_T)}^{\gamma-1}\nabla p_{C(P_T)}}{(K_w * P_T)^{\gamma}}(x_0), \tag{1.6.12}$$

where

$$A = \int\left(\frac{\gamma}{\gamma-1} - \frac{p_T(x)}{K_w * P_T}\right)\frac{\left[(\gamma-1)\nabla p_{C(P_T)}\left(\nabla p_{C(P_T)}\right)^t + p_{C(P_T)}J_{p_{C(P_T)}}\right]p_{C(P_T)}^{\gamma-2}}{(K_w * P_T)^{\gamma-1}}. \tag{1.6.13}$$

[13] The arginf function is a troublesome function when it comes to continuity and derivatives.

Remark 1.6.4. The form of the IF is somewhat similar to the IF of the classical MDφDE defined by (1.3.6). Toma and Broniatowski [2011] show that the IF of the classical MDφDE is given by:

$$\mathrm{IF}(P_T, x) = J^{-1}\frac{\nabla p_{\phi^T}}{p_{\phi^T}}, \tag{1.6.14}$$

where J is the information matrix given by $\int \frac{\nabla p_{\phi^T}(\nabla p_{\phi^T})^t}{p_{\phi^T}}$. Going back to the IF of the new MDφDE given by (1.6.12) and making w goes to zero would replace $K_w * P_T$ by p_{ϕ^T}. The first term thus disappears, and the IF would give $A^{-1}\frac{\nabla p_{\phi^T}}{p_{\phi^T}}$, where $A = J + \frac{1}{\gamma-1}J_{p_{\phi^T}}$ and $J_{p_{\phi^T}}$ is the matrix of second derivatives of p_ϕ with respect to ϕ.
On the other hand, a general comparison for $w > 0$ is also interesting. Indeed, the second term in (1.6.12) is a function of x and seems to be the term guiding the boundedness of the IF. We can rewrite it as follows:

$$A^{-1}\frac{p_{C(P_T)}^{\gamma-1}\nabla p_{C(P_T)}}{(K_w * P_T)^\gamma}(x_0) = A^{-1}\frac{p_{C(P_T)}^{\gamma}}{(K_w * P_T)^\gamma}\frac{\nabla p_{C(P_T)}}{p_{C(P_T)}}(x_0).$$

A direct comparison with (1.6.14) shows that our approach has resulted in the term $\frac{p_{C(P_T)}^{\gamma}}{(K_w * P_T)^\gamma}$ which could oblige the IF to be bounded in some cases. This is the ratio between the true density and the smoothed one. When $\gamma > 0$, it is surprising that the IF becomes *more* bounded as the ratio between the true distribution and the smoothed one decreases, which means that the smoothing is producing over estimation at the tail of the distribution.

Example 1.6.2. We resume the univariate Gaussian example. It can be proved that in this model, the kernel-based MDφDE is Fisher consistent. Indeed, function $P_T H(P_T, \mu)$ is given by:

$$P_T H(P_T, \mu) = \frac{1}{\gamma - 1}\sqrt{\frac{1+w^2}{1+\gamma w^2}}e^{-\frac{\gamma(1-\gamma)}{2(1+\gamma w^2)}\mu^2} - \frac{1}{\gamma}\sqrt{\frac{1+w^2}{(\gamma+1)w^2+1}}e^{-\frac{\gamma(w^2+1-\gamma)}{2(1+(\gamma+1)w^2)}\mu^2} - \frac{1}{\gamma(\gamma-1)}.$$

This formula holds for $\gamma \in [-1, \infty) \setminus \{0, 1\}$ whatever the value of w. It also holds for $\gamma < -1$ whenever $w^2 \leq -\frac{1}{\gamma}$. This function has a minimum at $\mu = 0$ whatever the value of w when $\gamma > 0$ (but different from 1), and has a local[14] minimum at zero when $\gamma < 0$ whatever the value of w, see figure (1.3).

Let's calculate the IF given by (1.6.12). We leave the calculus of the matrix A to the end. The second term is given by:

$$\frac{p_{\phi^T}^{\gamma-1}\nabla p_{\phi^T}}{(K_w * P)^\gamma}(x_0) = (1+w^2)^{\gamma/2}x_0 e^{-\frac{\gamma w^2}{2(1+w^2)}x_0^2}.$$

[14]The minimum of $P_T H(P_T, \mu)$ is $-\infty$ and is attained at $\pm\infty$, but when we restrain the set of possible parameters to a compact subset around zero, the estimation procedure becomes possible. Recall also that consistency was only proved when $\mu \in [\mu_{\min}, \mu_{\max}]$, see example 1.6.1 in this section.

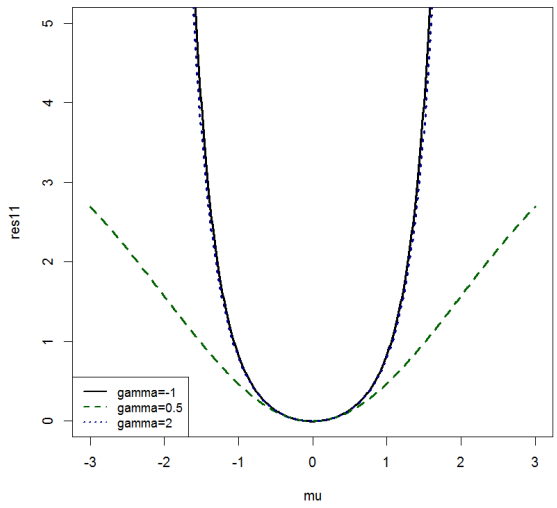

Figure 1.3: Function $P_T H(P_T, \mu)$ for different windows and divergences. They all have an infimum at zero

Hence, this quantity is bounded as soon as $\gamma > 0$. The second quantity is an integral which needs to exist and be finite. We have:

$$\frac{p_{\phi^T}^{\gamma-1}(x)K((x-x_0)/w)\nabla p_{\phi^T}(x)}{(K_w * P)^\gamma(x)}\left(1-\frac{p}{K_w * P}\right)(x) =$$
$$\frac{(1+w^2)^{\frac{\gamma+1}{2}}}{w}\exp\left[-\frac{\gamma w^4+1}{2w^2(1+w^2)}x^2+\frac{xx_0}{w^2}-\frac{x_0^2}{2w^2}\right]\times\left(\frac{1}{\sqrt{1+w^2}}e^{-\frac{x^2}{2(1+w^2)}}-e^{-\frac{x^2}{2}}\right).$$

It is clear now that if $\gamma > 0$, the integral exists. We should not forget that the integral term also depends on x_0. The dominating term is $e^{-x_0^2}$, so that the integral term is bounded as a function of x_0 as soon as the integral exists.
It remains to show that the term A exists and is invertible. Since $\nabla p_{\phi^T}(x) = xe^{-x^2/2}$, and $J_{p_{\phi^T}} = (1+x^2)e^{-x^2/2}$, then:

$$A = \sqrt{\frac{1+w^2}{2\pi}}\frac{\gamma}{\gamma-1}\left(\sqrt{\frac{2\pi}{a}}+\gamma\sqrt{\frac{2\pi}{a^3}}\right)-\frac{1+w^2}{\sqrt{2\pi}}\left(\sqrt{\frac{2\pi}{b}}+\gamma\sqrt{\frac{2\pi}{b^3}}\right),$$

where $a = \frac{\gamma w^2+1}{1+w^2}$ and $b = \frac{\gamma w^2+w^2+1}{1+w^2}$. It is clear that for $\gamma \in (0,1)$, the two terms constituting A have the same sign, hence A cannot be zero since it is the sum of two negative terms. However, if $\gamma > 1$, A may by zero for some cases. Indeed, A is 0 whenever $\gamma^2(1+\gamma+2\gamma w^2)^2(1+(\gamma+1)w^2)^3-(\gamma-1)(1+w^2)(1+\gamma+(\gamma+2)w^2)^2 = 0$. Notice that function $w \mapsto \gamma^2(1+\gamma+2\gamma w^2)^2(1+(\gamma+1)w^2)^3-(\gamma-1)(1+w^2)(1+\gamma+(\gamma+2)w^2)^2$ is equal to $2\gamma-1 > 0$ when $w = 0$, whereas it has a $-\infty$ limit at $+\infty$. Thus, it passes by zero for some $w > 0$ since it is a continuous function.
Previous arguments permit us to conclude for sure that for $\gamma \in (0,1)$, the influence function of the estimator defined by (1.5.4) is bounded in the Gaussian model independently of the bandwidth of the Gaussian kernel. Moreover, it is unbounded for $\gamma < 0$. Hence, one can hope to get a robust estimation when $\gamma \in (0,1)$, whereas further investigations are needed for the case of $\gamma < 0$.

1.7 Simulation study: comparison

Summary of the estimation methods and error criterion. We summarize the results of 100 experiments by giving the average of the estimates and the error committed, and the corresponding standard deviation. We consider two error criteria; the total variation distance (TVD) and the Chi square divergence between the true distribution and the estimated one. These criteria are defined as follows:

$$\sqrt{\chi^2(p_\phi, p_{\phi^T})} = \sqrt{\int \frac{\left(p_\phi(y) - p_{\phi^T}(y)\right)^2}{p_{\phi^T}(y)} dy}; \tag{1.7.1}$$

$$\mathrm{TVD}(p_\phi, p_{\phi^T}) = \sup_{A \in \mathcal{B}_n(\mathbb{R})} \left| dP_\phi(A) - dP_{\phi^T}(A) \right|. \tag{1.7.2}$$

We prefer to use the Chi square divergence, because it measures the relative error between two probability laws. Hence, the error committed on sets where the true distribution attributes small values is penalized in a similar way to sets where the true distribution attributes large values. We use also the TVD because it has the property of measuring the largest error committed when measuring a set A using the estimated distribution instead of the true one. The TVD can be directly calculated using the $L1$ distance. Indeed, the Scheffé lemma (see Meister [2009] page 129.) states that:

$$\sup_{A \in \mathcal{B}_n(\mathbb{R})} \left| dP_\phi(A) - dP_{\phi^T}(A) \right| = \frac{1}{2} \int_{\mathbb{R}} \left| p_\phi(y) - p_{\phi^T}(y) \right| dy.$$

We consider the Hellinger divergence for estimators based on $\varphi-$divergences. Our preference of the Hellinger divergence is that we hope to obtain robust estimators without loss of efficiency, see Jiménz and Shao [2001]. The parameter vector is estimated using six methods:

1. Maximum likelihood (MLE) which is calculated using EM for mixture models;
2. The classical MDφDE defined by (1.3.6);
3. Our kernel-based MDφDE defined by (1.5.4) with different choices for the kernel and its bandwidth;
4. The Basu-Lindsay approach with different choices for the kernel and its bandwidth;
5. The dual φ–divergence estimator (DφDE) defined by (1.3.7) with escort parameter the result of our kernel-based MDφDE with the best choice of the kernel and window among presented possibilities;
6. The minimum power density estimator (MDPD) of Basu et al. [1998] defined by (1.3.8) for $a \in \{0.1, 0.25, 0.5, 0.75, 1\}$.

We give for each experiment a summary of the results with comments, and precise the used kernels and the corresponding windows choices. We finally give an overall conclusion with some practical remarks.

Practical issues: Optimization was done using the Nelder-Mead algorithm. Integrations calculus were done using function `distrExIntegrate` of package `distrEx` which is a slight modification of the standard function `integrate`. It performs a Gauss-Legendre quadrature when function `integrate` returns an error. We have noticed that functions such as `integral` of package `pracma`[15], although has a good performance, is slow. Besides, function `int` of package `rmutil`, which uses either the Romberg method or algorithm 614 of the collected algorithms from ACM, seems to underestimate the value of the integral in slightly difficult circumstances such as heavy tailed distributions. For example, when we used it to calculate the classical MDφDE in the GPD case, it gave robust results because it underestimated the infinity part of the integral (forged thresholding effect). Finally, during some experiences on GPD observations and Weibull distributions based on the Basu-Lindsay approach, function `distrExIntegrate` failed to converge and function `integral` was used to attain a result.

Summary of the models and presentation of the results. Our simulation study covers the following models:

1. Gaussian model with unknown mean and variance;
2. Gaussian mixture with two components where the proportion and the two means are unknown;
3. Generalized Pareto distribution with unknown shape and scale;
4. Three Weibull mixtures with two components where the proportion and the two shapes are unknown.

Outliers were added in the original data in many ways which will be specified according to each case. We have either added noise outside the support of the dataset or by dispersing the noise over the whole dataset. We have also used different distributions to produce the noise.
For the first two models, we only used a Gaussian kernel with window chosen using either Silverman's rule (nrd0 in the statistical tool R) or Sheather and Jones' rule (SJ). For the heavy tailed models which are defined on the half real line, we needed to use non classical kernels such as asymmetric kernels (RIG: reciprocal inverse Gaussian and GA: gamma kernels) and the varying KDE of Mnatsakanov and Sarkisian [2012] denoted here as MT (Mellin transform) defined here above by (1.2.4), followed by the value of the bandwidth $\alpha \in \{5, 10, 15, 20\}$. In the GPD model and the first Weibull mixture, we present a simple comparison between symmetric kernels and other non classical methods and show the advantage of the later in such context. We therefore avoided using symmetric kernels for other Weibull mixtures. For the Basu-Lindsay approach, we did not implement asymmetric kernels, see discussion in paragraph 1.2.2. We only used the varying KDE.
Concerning the rule for deciding the window for the non classical kernels, we have tried out the cross-validation method (CV), but it resulted always in large (small for the varying KDE) and inconvenient windows especially when outliers are inserted. We were, therefore, obliged to use fixed windows in order to obtain good results. For each kernel and method, the window value or the rule used to calculate it is written next to it. More details can be found at each paragraph.

[15] Function `integral` includes a variety of adaptive numerical integration methods such as Kronrod-Gauss quadrature, Romberg's method, Gauss-Richardson quadrature, Clenshaw-Curtis (not adaptive) and (adaptive) Simpson's method.

1.7.1 Univariate Gaussian model

We consider the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ when both parameters μ and σ are unknown. We generate at each run a 100-sample of the standard Gaussian distribution $\mathcal{N}(0, 1)$. Outliers are added simply by replacing the 10 largest values in the sample by the value 10.
The maximum likelihood estimator of the parameters are simply the empirical mean and variance $\hat{\mu} = \frac{1}{100}\sum y_i, \hat{\sigma}^2 = \frac{1}{99}\sum (y_i - \hat{\mu})^2$. For methods which need kernels, we used a Gaussian kernel with two rules for the window; Silverman's rule and Sheather and Jones' one. We calculate the minimum density power divergence estimator (MDPD) for values of the tradeoff parameter $a \in \{0.1, 0.25, 0.5, 0.75, 1\}$. The D$\varphi$DE was calculated using the kernel-based MDφDE as an escort with Silverman's rule. Estimation results are summarized in table 1.2. Estimation error is calculated in table 1.3.
When we are under the model, all compared methods give the same result with very slight differences. As we add 10% outliers, the classical MDφDE and the MLE give the same result which is positively deviated from the true mean with a large variance. This is already expected by virtue of the result of [Broniatowski, 2014]. Other methods, ours included, give robust results except for MDPD with $a = 0.1$. Our estimator (for both windows choices) is at the same level of efficiency as the MLE under the model. Besides, the window choice seems irrelevant for methods based on kernels but for Beran's method where Silverman's rule is slightly better. The MDPD seems to give the best tradeoff between efficiency and robustness for $a = 0.5$ conquering other methods. The kernel-based MDφDE and the Basu-Lindsay approaches give slightly better efficiency which is traded with slightly lower robustness in comparison to the result of MDPD with $a = 0.5$.

Estimation method	No Outliers				10% Outliers			
	μ	sd(μ)	σ	sd(σ)	μ	sd(μ)	σ	sd(σ)
Hellinger								
Classical MDφDE	0.005	0.111	0.983	0.082	0.833	0.103	3.157	0.039
New MDφDE - Silverman	0.005	0.113	0.967	0.081	-0.187	0.114	0.810	0.069
New MDφDE - SJ	0.005	0.113	0.973	0.082	-0.191	0.114	0.800	0.068
Basu-Lindsay - Silverman	0.005	0.114	0.968	0.081	-0.191	0.114	0.805	0.068
Basu-Lindsay - SJ	0.005	0.113	0.970	0.081	-0.193	0.114	0.799	0.067
Beran - Silverman	0.005	0.113	1.024	0.087	-0.191	0.114	0.878	0.075
Beran - SJ	0.005	0.112	1.048	0.089	-0.192	0.114	0.853	0.073
MDPD 0.1	0.005	0.112	0.983	0.082	0.319	0.111	2.451	0.079
MDPD 0.25	0.006	0.112	0.983	0.083	-0.145	0.114	0.854	0.074
MDPD 0.5	0.008	0.117	0.979	0.087	-0.115	0.116	0.875	0.081
MDPD 0.75	0.010	0.123	0.975	0.093	-0.093	0.120	0.894	0.089
MDPD 1	0.012	0.129	0.971	0.098	-0.077	0.124	0.910	0.094
DφDE	0.005	0.112	0.982	0.082	-0.164	0.114	0.873	0.080
MLE	0.005	0.111	0.988	0.082	0.833	0.103	3.172	0.039

Table 1.2: The mean value and the standard deviation of the estimates in a 100-run experiment in the standard Gaussian model. The divergence criterion is the Hellinger divergence. The escort parameter of the DφDE is taken as the new MDφDE with Silverman's rule.

Estimation method	No Outliers				10% Outliers			
	χ^2	sd(χ^2)	TVD	sd(TVD)	χ^2	sd(χ^2)	TVD	sd(TVD)
Hellinger								
Classical MDφDE	0.104	0.052	0.054	0.026	8.503	0.113	0.516	0.002
New MDφDE - Silverman	0.106	0.052	0.056	0.028	0.230	0.063	0.136	0.041
New MDφDE - SJ	0.105	0.052	0.055	0.027	0.239	0.062	0.141	0.041
Basu-Lindsay - Silverman	0.105	0.052	0.055	0.028	0.235	0.062	0.139	0.040
Basu-Lindsay - SJ	0.105	0.052	0.055	0.027	0.240	0.062	0.142	0.040
Beran - Silverman	0.114	0.063	0.054	0.025	0.191	0.067	0.110	0.042
Beran - SJ	0.125	0.076	0.057	0.026	0.205	0.066	0.119	0.042
DφDE	0.104	0.052	0.054	0.026	0.183	0.068	0.105	0.042
MDPD 0.1	0.104	0.051	0.053	0.026	5.772	0.356	0.411	0.013
MDPD 0.25	0.105	0.052	0.054	0.026	0.185	0.066	0.107	0.042
MDPD 0.5	0.110	0.054	0.057	0.028	0.165	0.068	0.094	0.042
MDPD 0.75	0.116	0.060	0.060	0.032	0.152	0.070	0.086	0.043
MDPD 1	0.121	0.066	0.063	0.036	0.144	0.070	0.080	0.043
MLE	0.104	0.052	0.053	0.025	8.522	0.111	0.518	0.002

Table 1.3: The mean value of errors committed in a 100-run experiment with the standard deviation. The divergence criterion is the Hellinger divergence. The escort parameter of the DφDE is taken as the new MDφDE with Silverman's rule.

1.7.2 Mixture of two Gaussian components

We show in this paragraph several simulations from a two-component Gaussian mixture where the data is contaminated or not by a 10% of outliers. The true values of the mixture parameters are $\lambda = 0.35, \mu_1 = -2, \mu_2 = 1.5$. The variance of both components is supposed to be known and fixed at 1. Contamination was done for the first mixture by adding in the original sample to the 5 lowest values random observations from the uniform distribution $\mathcal{U}[-5, -2]$. We also added to the 5 largest values random observations from the uniform distribution $\mathcal{U}[2, 5]$. Estimation results are summarized in table 1.4. Estimation error is calculated in table 1.5. Maximum likelihood estimates are calculated using the EM algorithm. Table 1.7.2 contains a simulation with 1000 observations in each sample to illustrate that the comparison holds with higher number of observations.
Under the model, all compared methods give the same performance. When outliers are added, both classical MDφDE and MLE are not robust and give the same result. Other methods provide robust results. Error values are close for $\varphi-$divergence-based estimators and very close to results obtained by the MDPD which gives slightly better performances.

Table 1.4: The mean value and the standard deviation of the estimates in a 100-run experiment in the two-component Gaussian mixture

Estimation method	No Outliers						10% Outliers					
	λ	sd(λ)	μ_1	sd(μ_1)	μ_2	sd(μ_2)	λ	sd(λ)	μ_1	sd(μ_1)	μ_2	sd(μ_2)
Classical MDφDE	0.360	0.054	-1.989	0.204	1.493	0.136	0.342	0.064	-2.617	0.288	1.713	0.172
New MDφDE - Gauss:Silverman	0.360	0.054	-1.993	0.208	1.499	0.133	0.349	0.058	-1.767	0.226	1.377	0.135
New MDφDE - Gauss:1.2	0.359	0.054	-2.024	0.210	1.523	0.132	0.348	0.058	-1.811	0.218	1.411	0.132
Basu-Lindsay - Gauss:Silverman	0.361	0.055	-1.979	0.207	1.490	0.139	0.339	0.062	-1.927	0.305	1.377	0.158
Basu-Lindsay - Gauss:0.9	0.361	0.055	-1.976	0.215	1.489	0.143	0.334	0.066	-1.987	0.288	1.378	0.162
Beran - Gauss:Silverman	0.371	0.050	-1.985	0.203	1.546	0.132	0.369	0.053	-1.788	0.218	1.477	0.134
Beran - Gauss:0.9	0.381	0.046	-1.968	0.202	1.594	0.127	0.375	0.048	-1.785	0.218	1.502	0.130
DφDE	0.361	0.054	-1.988	0.203	1.492	0.136	0.355	0.056	-2.132	0.224	1.605	0.137
MDPD 0.1	0.360	0.054	-1.991	0.207	1.493	0.134	0.346	0.059	-2.052	0.243	1.452	0.144
MDPD 0.25	0.360	0.053	-1.994	0.213	1.492	0.133	0.351	0.057	-1.832	0.223	1.394	0.134
MDPD 0.5	0.360	0.053	-1.997	0.226	1.489	0.136	0.353	0.056	-1.819	0.218	1.404	0.132
MLE (EM)	0.360	0.054	-1.989	0.204	1.493	0.136	0.342	0.064	-2.617	0.288	1.713	0.172

Table 1.5: The mean value of errors committed in a 100-run experiment with the standard deviation in the two-component Gaussian mixture

Estimation method	No Outliers				10% Outliers			
	$\sqrt{\chi^2}$	sd($\sqrt{\chi^2}$)	TVD	sd(TVD)	$\sqrt{\chi^2}$	sd($\sqrt{\chi^2}$)	TVD	sd(TVD)
Classical MDφDE	0.113	0.044	0.064	0.025	0.335	0.102	0.150	0.034
New MDφDE - Gauss:Silverman	0.113	0.045	0.064	0.025	0.155	0.059	0.087	0.033
New MDφDE - Gauss:1.2	0.114	0.047	0.064	0.025	0.139	0.053	0.078	0.030
Basu-Lindsay - Gauss:Silverman	0.115	0.043	0.065	0.024	0.155	0.073	0.085	0.033
Basu-Lindsay - Gauss:0.9	0.118	0.043	0.067	0.024	0.147	0.059	0.083	0.034
Beran - Gauss:Silverman	0.113	0.046	0.064	0.025	0.132	0.050	0.073	0.027
Beran - Gauss:0.9	0.117	0.050	0.066	0.028	0.127	0.049	0.070	0.026
DφDE	0.112	0.044	0.064	0.025	0.142	0.061	0.076	0.031
MDPD 0.1	0.113	0.044	0.064	0.025	0.124	0.052	0.069	0.029
MDPD 0.25	0.114	0.045	0.064	0.025	0.140	0.054	0.079	0.030
MDPD 0.5	0.117	0.047	0.065	0.025	0.138	0.053	0.078	0.030
MLE	0.113	0.044	0.064	0.025	0.335	0.102	0.150	0.034

Estimation method	No Outliers				10% Outliers			
	$\sqrt{\chi^2}$	sd($\sqrt{\chi^2}$)	TVD	sd(TVD)	$\sqrt{\chi^2}$	sd($\sqrt{\chi^2}$)	TVD	sd(TVD)
Classical MDφDE	0.036	0.016	0.020	0.009	0.308	0.031	0.142	0.013
New MDφDE - Gauss:Silverman	0.036	0.015	0.020	0.009	0.146	0.024	0.082	0.014
New MDφDE - Gauss:1.2	0.042	0.017	0.023	0.009	0.095	0.022	0.051	0.012
Basu-Lindsay - Gauss:Silverman	0.037	0.016	0.021	0.009	0.132	0.026	0.074	0.014
Basu-Lindsay - Gauss:1.2	0.040	0.017	0.022	0.009	0.129	0.041	0.071	0.022
Basu-Lindsay - Gauss:1	0.039	0.017	0.022	0.009	0.076	0.022	0.045	0.014
Beran - Gauss:Silverman	0.038	0.016	0.021	0.009	0.116	0.024	0.062	0.013
Beran - Gauss:1.2	0.114	0.018	0.066	0.010	0.114	0.022	0.065	0.012
Beran - Gauss:1	0.078	0.018	0.045	0.010	0.087	0.022	0.044	0.011
DφDE	0.045	0.016	0.020	0.009	0.079	0.023	0.044	0.013
MDPD 0.1	0.036	0.015	0.020	0.009	0.046	0.016	0.027	0.010
MDPD 0.25	0.036	0.015	0.020	0.009	0.095	0.023	0.053	0.013
MDPD 0.5	0.037	0.015	0.021	0.009	0.092	0.022	0.050	0.012
MLE	0.036	0.016	0.020	0.009	0.308	0.031	0.142	0.013

Table 1.6: The mean value of errors committed in a 100-run experiment with the standard deviation in the two-component Gaussian mixture. Number of observations is 1000

1.7.3 Generalized Pareto distribution

We show in this paragraph several simulations from the generalized Pareto distribution (GPD) where the data is contaminated or not by a 10% of outliers. A GPD with a fixed location at zero, a scale parameter $\sigma > 0$ and a shape parameter $\nu > 0$ is defined by:

$$p_{\nu,\sigma}(y) = \frac{1}{\sigma}\left(1 + \nu\frac{y}{\sigma}\right)^{-1-\frac{1}{\nu}}, \quad \text{for } y \geq 0.$$

The true set of parameters is $\nu = 0.7, \sigma = 3$. Outliers are added by replacing 10 observations (chosen randomly) from each sample by observations from the distribution GPD($\nu = 1, \sigma = 10, \mu = 500$) where μ is the location parameter. Estimation results are summarized in table 1.7. Estimation error is calculated in table 1.8. The maximum likelihood estimator was calculated using the `gpd.fit` function of package `ismev`.

Under the model, all presented methods except for the Basu-Lindsay approach have close performance to the MLE and sometimes even better for given choices of the kernel or the tradeoff parameter. Our kernel-based MDφDE attained a similar performance to the MLE for *all* non classical kernels and the corresponding choices of the window, and attained an even better efficiency than the MLE. Beran's method attained this performance only with the varying KDE. The MDPD attained it only for small values of a (=0.1). The use of a symmetric kernel (here the Gaussian) did not give good results in kernel-based methods except for our kernel-based MDφDE with a Silverman's rule for the window[16]. This may be some indication of low sensitivity to the kernel used.
When outliers are added, the performance of kernel-based methods was slightly deteriorated whereas other methods (the MDPD included for all values of a) were greatly influenced, and the error is at least doubled; MDPD for all cases included. The use of asymmetric kernels seems to be the most convenient for a GPD model. Our kernel-based MDφDE seems to give the best result (in χ^2 and TVD) for all kernels and corresponding windows keeping a great gap in its favor in comparison with other methods.

Remark 1.7.1. The nature of the heavy tail of the GPD (slow decrease at infinity) made integration calculus difficult, and some integration functions failed to give fairly correct results. We, therefore, and in order to avoid integration on an infinite interval $[0, \infty)$, propose to use a quantile trick which is translated by the change of variable:

$$\int_0^\infty \varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x)p_\phi(x)dx = \int_0^1 \varphi'\left(\frac{p_\phi}{p_\alpha}\right)p_\phi(\mathbb{F}_\phi^{-1}(y))dy,$$

where $\mathbb{F}_\phi^{-1}(y) = \frac{\sigma}{\nu}((1-y)^{-\nu} - 1)$ is the quantile of the GPD probability law P_ϕ. This idea may appear ineffective since it does not change anything in the integral (the quantile function takes back values from $[0, 1)$ into $[0, \infty)$). In fact, integration methods perform in general better when integrating on a finite interval than when integrating on an infinite one.

[16] The Sheather and Jones' rule did not give satisfactory results.

Estimation method	No Outliers				10% Outliers			
	ν	sd(ν)	σ	sd(σ)	ν	sd(ν)	σ	sd(σ)
Hellinger								
Classical MDφDE	0.721	0.174	3.029	0.575	1.655	0.113	2.694	0.491
New MDφDE - Gauss Silverman	0.463	0.142	2.719	0.586	0.571	0.197	2.427	0.599
New MDφDE - Gauss SJ	0.343	0.108	2.858	0.597	0.368	0.141	2.798	0.569
New MDφDE - RIG CV	0.528	0.140	3.125	0.611	0.775	0.202	2.844	0.571
New MDφDE - RIG Nrd0	0.562	0.139	3.133	0.605	0.817	0.219	2.815	0.545
New MDφDE - RIG SJ	0.522	0.129	3.138	0.616	0.688	0.191	2.903	0.574
New MDφDE - GA CV	0.530	0.139	3.117	0.610	0.766	0.204	2.833	0.577
New MDφDE - GA Nrd0	0.564	0.139	3.112	0.601	0.814	0.211	2.787	0.544
New MDφDE - GA SJ	0.520	0.126	3.135	0.607	0.691	0.185	2.895	0.576
New MDφDE - MT 5	0.641	0.156	3.217	0.615	1.202	0.161	2.806	0.510
New MDφDE - MT 10	0.607	0.153	3.272	0.628	1.090	0.195	2.876	0.552
New MDφDE - MT 15	0.588	0.150	3.307	0.636	1.026	0.206	2.920	0.565
New MDφDE - MT 20	0.573	0.148	3.331	0.643	0.979	0.212	2.956	0.577
Basu-Lindsay - Gauss Silverman	0.128	0.125	6.022	1.522	0.122	0.109	7.151	2.025
Basu-Lindsay - Gauss SJ	0.078	0.066	4.603	1.057	0.097	0.087	4.843	1.316
Basu-Lindsay - MT 5	0.833	0.156	2.232	0.651	0.765	0.189	2.937	0.666
Basu-Lindsay - MT 10	0.853	0.197	2.297	0.659	0.777	0.193	2.880	0.704
Basu-Lindsay - MT 15	0.881	0.176	2.293	0.517	1.164	0.169	2.893	0.530
Basu-Lindsay - MT 20	0.907	0.180	2.337	0.603	0.936	0.206	2.694	0.580
Beran - Gauss Nrd0	0.216	0.108	5.165	1.218	0.197	0.125	6.084	1.546
Beran - Gauss SJ	0.231	0.108	3.988	0.919	0.229	0.134	4.135	0.939
Beran - RIG CV	0.516	0.134	3.890	0.832	0.833	0.218	3.944	0.745
Beran - RIG Nrd0	0.515	0.138	4.441	1.026	0.878	0.233	4.229	0.954
Beran - RIG SJ	0.507	0.136	3.813	0.787	0.732	0.200	3.641	1.113
Beran - GA CV	0.486	0.134	3.936	0.847	0.745	0.207	4.097	0.822
Beran - GA Nrd0	0.475	0.139	4.510	0.998	0.778	0.220	4.547	1.032
Beran - GA SJ	0.503	0.133	3.780	0.773	0.703	0.186	3.589	0.781
Beran - MT 5	0.711	0.150	3.384	0.640	1.339	0.140	2.979	0.551
Beran - MT 10	0.665	0.150	3.315	0.620	1.231	0.155	2.900	0.530
Beran - MT 15	0.637	0.154	3.310	0.640	1.164	0.169	2.893	0.530
Beran - MT 20	0.627	0.156	3.302	0.637	0.936	0.206	2.694	0.580
DφDE	0.720	0.179	3.026	0.580	1.45	0.290	2.749	0.524
MDPD 1	0.729	0.402	3.023	0.660	1.039	0.483	3.273	0.681
MDPD 0.75	0.716	0.331	3.025	0.631	1.021	0.416	3.242	0.645
MDPD 0.5	0.715	0.263	3.023	0.603	1.028	0.361	3.171	0.605
MDPD 0.25	0.722	0.200	3.019	0.581	1.292	0.240	2.955	0.532
MDPD 0.1	0.723	0.175	3.019	0.568	1.564	0.154	2.779	0.500
MLE	0.719	0.174	3.031	0.58	1.654	0.113	2.695	0.492

Table 1.7: The mean value and the standard deviation of the estimates in a 100-run experiment in the GPG model. The escort parameter of the DφDE is taken as the new MDφDE with Silverman's rule.

Estimation method	No Outliers				10% Outliers			
	χ^2	sd(χ^2)	TVD	sd(TVD)	χ^2	sd(χ^2)	TVD	sd(TVD)
Hellinger								
Classical MDφDE	0.099	0.077	0.044	0.026	1.027	0.195	0.142	0.014
New MDφDE - Silverman	0.159	0.056	0.087	0.034	0.171	0.070	0.097	0.044
New MDφDE - SJ	0.189	0.052	0.100	0.035	0.183	0.066	0.098	0.042
New MDφDE - RIG CV	0.109	0.045	0.058	0.027	0.114	0.065	0.053	0.029
New MDφDE - RIG Nrd0	0.100	0.044	0.054	0.027	0.142	0.130	0.056	0.029
New MDφDE - RIG SJ	0.110	0.044	0.059	0.027	0.104	0.056	0.054	0.030
New MDφDE - GA CV	0.108	0.045	0.058	0.027	0.114	0.063	0.054	0.029
New MDφDE - GA Nrd0	0.100	0.044	0.054	0.027	0.132	0.092	0.056	0.028
New MDφDE - GA SJ	0.109	0.044	0.058	0.027	0.104	0.056	0.054	0.030
New MDφDE - MT 5	0.093	0.053	0.049	0.028	0.472	0.307	0.089	0.024
New MDφDE - MT 10	0.095	0.050	0.051	0.028	0.336	0.243	0.078	0.026
New MDφDE - MT 15	0.097	0.048	0.053	0.028	0.268	0.193	0.072	0.027
New MDφDE - MT 20	0.099	0.047	0.054	0.029	0.226	0.154	0.068	0.028
Basu-Lindsay - Silverman	0.301	0.08	0.179	0.048	0.361	0.110	0.214	0.061
Basu-Lindsay - SJ	0.256	0.046	0.145	0.033	0.264	0.055	0.151	0.039
Basu-Lindsay - MT 5	0.155	0.082	0.090	0.047	0.100	0.077	0.051	0.036
Basu-Lindsay - MT 10	0.155	0.080	0.085	0.043	0.102	0.078	0.053	0.038
Basu-Lindsay - MT 15	0.140	0.107	0.071	0.050	0.421	0.278	0.086	0.025
Basu-Lindsay - MT 20	0.157	0.085	0.078	0.044	0.160	0.083	0.059	0.031
Beran - Gauss Nrd0	0.241	0.072	0.142	0.045	0.297	0.090	0.177	0.053
Beran - Gauss SJ	0.199	0.049	0.109	0.034	0.207	0.044	0.114	0.032
Beran - RIG CV	0.133	0.060	0.076	0.038	0.226	0.128	0.094	0.041
Beran - RIG Nrd0	0.164	0.085	0.097	0.051	0.306	0.235	0.114	0.054
Beran - RIG SJ	0.123	0.060	0.069	0.039	0.146	0.097	0.070	0.048
Beran - GA CV	0.136	0.060	0.078	0.038	0.195	0.100	0.094	0.044
Beran - GA Nrd0	0.169	0.078	0.101	0.048	0.267	0.186	0.121	0.057
Beran - GA SJ	0.120	0.058	0.068	0.037	0.130	0.078	0.065	0.040
Beran - MT 5	0.103	0.067	0.052	0.030	0.915	0.729	0.111	0.022
Beran - MT 10	0.093	0.057	0.049	0.029	0.581	0.615	0.095	0.023
Beran - MT 15	0.094	0.054	0.050	0.029	0.421	0.278	0.086	0.025
Beran - MT 20	0.095	0.055	0.051	0.029	0.371	0.298	0.081	0.026
DφDE	0.099	0.077	0.048	0.028	0.843	0.407	0.120	0.030
MDPD 1	0.211	0.310	0.068	0.038	0.477	0.665	0.089	0.047
MDPD 0.75	0.204	0.389	0.062	0.034	0.424	0.545	0.085	0.043
MDPD 0.5	0.141	0.160	0.056	0.030	0.419	0.515	0.082	0.039
MDPD 0.25	0.106	0.082	0.049	0.028	0.669	0.441	0.104	0.030
MDPD 0.1	0.099	0.083	0.047	0.027	0.955	0.326	0.133	0.019
MLE	0.099	0.077	0.048	0.026	1.025	0.195	0.142	0.014

Table 1.8: The mean value of errors committed in a 100-run experiment with the standard deviation for the GPD model. The escort parameter of the DφDE is taken as the new MDφDE with the gamma kernel.

1.7.4 Mixtures of Two Weibull Components

We present the results of estimating three different two-component Weibull mixtures. The model has the following density:

$$p_\phi(x) = 2\lambda\nu_1(2x)^{\nu_1-1}e^{-(2x)^{\nu_1}} + (1-\lambda)\frac{\nu_2}{2}\left(\frac{x}{2}\right)^{\nu_2-1}e^{-\left(\frac{x}{2}\right)^{\nu_2}}.$$

Scale parameters are supposed to be known and equal to 0.5 for the first component and 2 for the second component. The proportion is unknown and fixed at 0.35. Shape parameters are supposed unknown. Our examples cover a variety of cases of a Weibull mixture where the density function has either a finite limit at zero or goes to infinity for one of the components:

1. a mixture with close modes $\nu_1 = 1.2, \nu_2 = 2$;
2. a mixture with one mode and with limit equal to infinity at zero $\nu_1 = 0.5, \nu_2 = 3$;
3. a mixture with no modes and with limit equal to infinity at zero $\nu_1 = 0.5, \nu_2 = 1$.

We plot these mixtures in figure 1.4. Outliers were added in different ways to illustrate several scenarios. For the first mixture, outliers were added by replacing 10 observations of each sample chosen randomly by 10 observations drawn independently from a Weibull distribution with shape $\nu = 0.9$ and scale $\sigma = 3$. See tables (1.9) and (1.10). For the second mixture, we added to the 10 largest observations of each sample a random observation drawn from the uniform distribution $\mathcal{U}[2, 10]$. See tables 1.11 and 1.12. For the third one, outliers were added by replacing 10 observations, chosen randomly, of each sample by observations from the uniform distribution $\mathcal{U}[\max y_i, 75]$ after having verified that no observation in the overall data has exceeded the value 50. See tables 1.13 and 1.14.

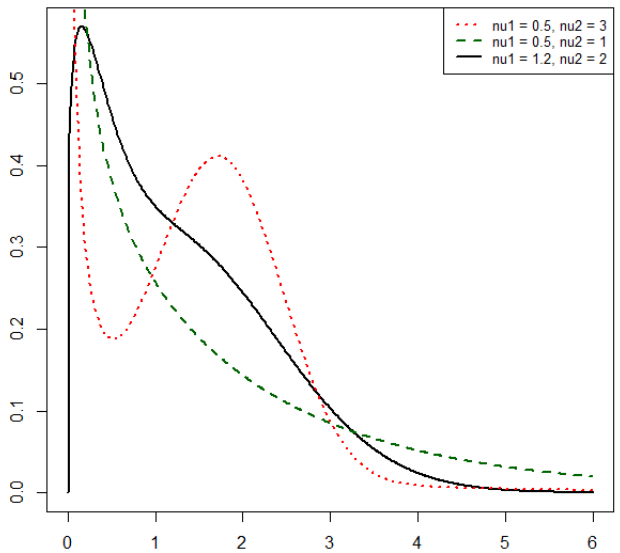

Figure 1.4: The three Weibull mixtures used in our experience.

The caclulus of the χ^2 divergence between the estimated model and the true distribution gave often infinity on all mixtures for all estimation methods even under the model. This is because a small bias in the estimation of the shape parameter results in a great relative error in both the tail behavior and near zero. We therefore, only provide the TVD as an error criterion.
The first Weibull mixture was the least complicated case. We were able to get satisfactory results for our kernel-based MDφDE using a Gaussian kernel. The two other mixtures were more challenging, and we needed to use asymmetric kernels to solve the problem of the bias near zero. It is worth noting that the Basu-Lindsay approach provided very bad estimates in the three mixtures which keeps it out of the competition. Note also that the use of a Gaussian kernel gave very pleasant results for the first mixture in spite of the boundary bias. We excluded it from mixtures which have infinity limit at zero because it did not work well because of the large bias at zero.

For the first mixture, under the model all presented methods provide close results (and sometimes better) to the MLE except for the Basu-Lindsay approach with all available choices and Beran's method with the varying KDE (MT) for windows 5 and 10 which fail. Under contamination, our method gives better results than all other methods and have very close (even slightly better) performance to the MDPD for tradeoff parameter higher than 0.25.
For the second mixture, the Basu-Lindsay approach failed again. Beran's method gave good result under the model only in one case; the RIG with window 0.01. The MDPD worked very well only for a tradeoff parameter lower than 0.5 and gave a good compromise between robustness and efficiency. It gave the best compromise in the presented methods. Our kernel-based MDφDE has close results to the MDPD with difference of 0.01 in the TVD. It is worth noting that our kernel-based MDφDE gave faire results for the two proposed kernels; the asymmetric kernel RIG for window 0.01 as before and the varying KDE MT for windows 10, 15 and 20. A fact which was not verified for other kernel-based methods showing again a less sensibility towards the kernel.
For the third mixture, the Basu-Lindsay approach did not give good results especially under the model. The only satisfactory results (which gave a good tradeoff between robustness and efficiency) were obtained by our kernel-based MDφDE for RIG kernel with window 0.01, Beran's method with the same kernel and window and the MDPD for $a = 0.5$. Our method and Beran's gave the same result with difference of 0.015 in favor of the MDPD. Better efficiency were obtained by other choices but on the cost of the robustness of the resulting estimator under contamination.

Estimation method	No Outliers						10% Outliers					
	λ	sd(λ)	ν_1	sd(ν_1)	ν_2	sd(ν_2)	λ	sd(λ)	ν_1	sd(ν_1)	ν_2	sd(ν_2)
Hellinger												
Classical MDφDE	0.355	0.066	1.245	0.228	2.054	0.237	0.410	0.257	1.045	0.255	1.718	0.849
New MDφDE - Gauss Silverman	0.384	0.067	1.221	0.244	2.138	0.291	0.348	0.076	1.121	0.265	1.822	0.319
New MDφDE - Gauss SJ	0.387	0.067	1.227	0.240	2.188	0.308	0.356	0.076	1.133	0.261	1.905	0.319
New MDφDE - RIG 0.01	0.371	0.066	1.297	0.231	2.215	0.321	0.355	0.100	1.213	0.229	1.955	0.344
New MDφDE - RIG 0.1	0.358	0.065	1.233	0.210	2.065	0.267	0.330	0.117	1.127	0.226	1.741	0.304
New MDφDE - RIG SJ	0.351	0.066	1.217	0.207	2.001	0.245	0.324	0.132	1.107	0.226	1.670	0.297
New MDφDE - MT 5	0.328	0.112	1.301	0.235	1.809	0.192	0.363	0.229	1.195	0.213	1.592	0.356
New MDφDE - MT 10	0.330	0.091	1.355	0.235	1.923	0.220	0.351	0.204	1.247	0.230	1.645	0.285
New MDφDE - MT 15	0.327	0.076	1.383	0.234	1.973	0.237	0.348	0.199	1.275	0.233	1.680	0.294
New MDφDE - MT 20	0.328	0.076	1.403	0.233	2.002	0.249	0.348	0.198	1.295	0.235	1.702	0.297
Basu-Lindsay - Gauss Silverman	0.752	0.064	2.199	0.248	38.66	8.66	0.822	0.083	1.927	0.276	32.37	13.52
Basu-Lindsay - Gauss SJ	0.723	0.059	2.205	0.257	16.18	10.75	0.759	0.065	1.958	0.263	19.52	10.56
Basu-Lindsay - MT 5	0.403	0.072	1.339	0.224	3.241	0.547	0.346	0.076	1.260	0.210	2.874	0.338
Basu-Lindsay - MT 10	0.390	0.069	1.409	0.234	3.281	0.465	0.337	0.067	1.319	0.217	2.813	0.233
Basu-Lindsay - MT 15	0.393	0.067	1.458	0.248	3.297	0.476	0.333	0.062	1.340	0.232	2.823	0.257
Basu-Lindsay - MT 20	0.399	0.066	1.472	0.221	3.282	0.458	0.335	0.068	1.362	0.225	2.819	0.300
Beran - Gauss Silverman	0.254	0.058	1.313	0.087	2.010	0.200	0.182	0.074	1.174	0.162	1.703	0.253
Beran - Gauss SJ	0.295	0.067	1.371	0.104	2.085	0.225	0.240	0.079	1.284	0.127	1.794	0.266
Beran - RIG 0.01	0.368	0.064	1.240	0.198	2.147	0.277	0.339	0.094	1.151	0.200	1.858	0.332
Beran - RIG 0.1	0.345	0.061	1.117	0.103	1.897	0.172	0.289	0.095	1.033	0.125	1.570	0.247
Beran - RIG SJ	0.320	0.060	1.069	0.074	1.725	0.138	0.260	0.123	0.997	0.088	1.416	0.203
Beran - MT 5	0.453	0.307	1.146	0.178	1.386	0.180	0.626	0.349	1.055	0.172	1.461	0.531
Beran - MT 10	0.354	0.201	1.238	0.201	1.553	0.133	0.419	0.304	1.134	0.202	1.450	0.425
Beran - MT 15	0.334	0.153	1.286	0.211	1.664	0.143	0.404	0.277	1.178	0.188	1.500	0.370
Beran - MT 20	0.334	0.136	1.317	0.218	1.738	0.156	0.383	0.256	1.207	0.198	1.542	0.348
DφDE	0.356	0.066	1.248	0.232	2.069	0.278	0.332	0.142	1.113	0.248	1.700	0.289
MDPD 1	0.358	0.087	1.238	0.252	2.127	0.521	0.343	0.113	1.167	0.239	2.005	0.517
MDPD 0.75	0.353	0.073	1.236	0.237	2.088	0.397	0.341	0.108	1.164	0.235	1.951	0.432
MDPD 0.5	0.354	0.068	1.238	0.230	2.071	0.345	0.336	0.105	1.159	0.237	1.860	0.344
MDPD 0.25	0.354	0.066	1.239	0.226	2.053	0.272	0.324	0.131	1.132	0.235	1.699	0.321
MDPD 0.1	0.355	0.066	1.242	0.227	2.048	0.238	0.394	0.241	1.091	0.215	1.780	0.792
MLE (EM)	0.355	0.066	1.245	0.228	2.054	0.237	0.321	0.187	0.913	0.313	1.575	0.325

Table 1.9: The mean value and the standard deviation of the estimates in a 100-run experiment on a two-component Weibull mixture ($\lambda = 0.35, \nu_1 = 1.2, \nu_2 = 2$). The escort parameter of the DφDE is taken as the new MDφDE with the SJ bandwidth choice.

Estimation method	No Outliers			10% Outliers		
	mean	median	sd	mean	median	sd
Hellinger						
Classical MDφDE	0.052	0.048	0.025	0.108	0.094	0.099
New MDφDE - Gauss Silverman	0.058	0.054	0.029	0.068	0.065	0.034
New MDφDE - Gauss SJ	0.058	0.053	0.029	0.064	0.061	0.031
New MDφDE - RIG 0.01	0.058	0.052	0.030	0.059	0.057	0.030
New MDφDE - RIG 0.1	0.051	0.049	0.026	0.066	0.062	0.032
New MDφDE - RIG SJ	0.050	0.050	0.026	0.071	0.066	0.032
New MDφDE - MT 5	0.057	0.055	0.025	0.081	0.074	0.032
New MDφDE - MT 10	0.054	0.053	0.026	0.075	0.071	0.032
New MDφDE - MT 15	0.054	0.054	0.026	0.073	0.069	0.032
New MDφDE - MT 20	0.055	0.054	0.027	0.073	0.069	0.031
Basu Lindsay - Gauss Silverman	0.298	0.289	0.042	0.247	0.253	0.050
Basu Lindsay - Gauss SJ	0.252	0.256	0.051	0.242	0.246	0.044
Basu Lindsay - MT 5	0.127	0.141	0.046	0.121	0.111	0.042
Basu Lindsay - MT 10	0.133	0.136	0.039	0.117	0.111	0.036
Basu Lindsay - MT 15	0.134	0.141	0.039	0.118	0.110	0.038
Basu Lindsay - MT 20	0.132	0.138	0.039	0.117	0.109	0.039
Beran - Gauss Silverman	0.068	0.062	0.028	0.082	0.081	0.031
Beran - Gauss SJ	0.060	0.054	0.028	0.067	0.065	0.029
Beran - RIG 0.01	0.052	0.048	0.026	0.060	0.058	0.029
Beran - RIG 0.1	0.042	0.039	0.020	0.067	0.061	0.030
Beran - RIG SJ	0.045	0.044	0.017	0.079	0.076	0.030
Beran - MT 5	0.099	0.097	0.016	0.125	0.125	0.022
Beran - MT 10	0.073	0.070	0.021	0.102	0.100	0.028
Beran - MT 15	0.064	0.060	0.022	0.092	0.089	0.030
Beran - MT 20	0.059	0.055	0.023	0.086	0.084	0.030
DφDE	0.053	0.049	0.027	0.068	0.065	0.031
MDPD 1	0.065	0.061	0.034	0.068	0.064	0.030
MDPD 0.75	0.059	0.056	0.029	0.063	0.060	0.029
MDPD 0.5	0.056	0.052	0.029	0.061	0.056	0.029
MDPD 0.25	0.052	0.048	0.027	0.068	0.067	0.031
MDPD 0.1	0.051	0.048	0.026	0.088	0.083	0.039
MLE	0.052	0.048	0.025	0.095	0.098	0.035

Table 1.10: The mean value with the standard deviation of the TVA committed in a 100-run experiment on a two-component Weibull mixture ($\lambda = 0.35, \nu_1 = 1.2, \nu_2 = 2$). The escort parameter of the DφDE is taken as the new MDφDE with the SJ bandwidth choice.

Estimation	No Outliers						10% Outliers					
method	λ	sd(λ)	ν_1	sd(ν_1)	ν_2	sd(ν_2)	λ	sd(λ)	ν_1	sd(ν_1)	ν_2	sd(ν_2)
Hellinger												
Classical MDφDE	0.344	0.059	0.497	0.079	3.063	0.476	0.376	0.053	0.339	0.030	2.892	0.484
New MDφDE RIG - 0.01	0.330	0.061	0.540	0.140	3.170	0.503	0.338	0.061	0.432	0.105	3.055	0.583
New MDφDE RIG - 0.1	0.371	0.063	0.468	0.138	3.045	0.452	0.392	0.072	0.372	0.085	2.927	0.464
New MDφDE RIG - SJ	0.395	0.072	0.442	0.134	3.013	0.443	0.424	0.086	0.354	0.082	2.916	0.459
New MDφDE MT - 5	0.311	0.062	0.520	0.065	2.875	0.451	0.316	0.063	0.376	0.036	2.699	0.471
New MDφDE MT - 10	0.302	0.062	0.548	0.077	2.903	0.433	0.306	0.062	0.384	0.039	2.727	0.448
New MDφDE MT - 15	0.295	0.063	0.564	0.084	2.927	0.434	0.298	0.063	0.388	0.042	2.745	0.450
New MDφDE MT - 20	0.289	0.063	0.575	0.091	2.943	0.437	0.291	0.063	0.392	0.044	2.758	0.454
Basu-Lindsay MT - 5	0.250	0.070	0.834	0.168	2.849	0.733	0.185	0.074	0.715	0.208	2.189	0.155
Basu-Lindsay MT - 10	0.240	0.065	0.797	0.157	2.789	0.550	0.197	0.087	0.707	0.201	2.324	0.132
Basu-Lindsay MT - 15	0.254	0.073	0.745	0.140	2.915	0.584	0.204	0.078	0.674	0.181	2.352	0.092
Beran RIG - 0.01	0.298	0.058	0.647	0.082	3.017	0.437	0.295	0.057	0.486	0.081	2.842	0.460
Beran RIG - 0.1	0.234	0.054	0.652	0.105	2.374	0.245	0.216	0.053	0.408	0.056	2.149	0.291
Beran RIG - SJ	0.194	0.056	0.653	0.134	1.936	0.246	0.142	0.065	0.402	0.144	1.601	0.325
Beran MT - 5	0.250	0.070	0.463	0.058	1.603	0.140	0.245	0.083	0.340	0.062	1.494	0.208
Beran MT - 10	0.278	0.066	0.501	0.069	2.005	0.181	0.275	0.079	0.354	0.033	1.868	0.260
Beran MT - 15	0.286	0.065	0.524	0.075	2.224	0.218	0.284	0.071	0.365	0.033	2.068	0.280
DφDE	0.343	0.059	0.5004	0.084	3.047	0.474	0.372	0.056	0.357	0.056	2.897	0.502
MDE 0.75	0.444	0.126	0.595	0.080	3.466	0.643	0.417	0.127	0.602	0.087	3.233	0.606
MDE 0.5	0.376	0.067	0.551	0.093	3.159	0.488	0.357	0.067	0.555	0.097	2.980	0.484
MDE 0.25	0.347	0.061	0.512	0.096	3.057	0.472	0.331	0.062	0.471	0.068	2.879	0.491
MDE 0.1	0.344	0.059	0.496	0.084	3.050	0.470	0.343	0.058	0.384	0.037	2.859	0.484
MLE (EM)	0.344	0.059	0.498	0.079	3.063	0.476	0.376	0.053	0.339	0.303	2.892	0.482

Table 1.11: The mean value and the standard deviation of the estimates in a 100-run experiment in a two-component Weibull mixture ($\lambda = 0.35, \nu_1 = 0.5, \nu_2 = 3$). The escort parameter of the DφDE is taken as the new MDφDE with Silverman's rule.

Estimation method	No Outliers			10% Outliers		
	mean	median	sd	mean	median	sd
Hellinger						
Classical MDφDE	0.060	0.055	0.024	0.096	0.094	0.025
New MDφDE RIG - 0.01	0.074	0.070	0.034	0.076	0.073	0.039
New MDφDE RIG - 0.1	0.079	0.064	0.053	0.099	0.086	0.062
New MDφDE RIG - SJ	0.091	0.075	0.068	0.120	0.099	0.078
New MDφDE MT - 5	0.062	0.061	0.027	0.081	0.073	0.031
New MDφDE MT - 10	0.066	0.064	0.028	0.076	0.070	0.030
New MDφDE MT - 15	0.069	0.068	0.028	0.076	0.071	0.030
New MDφDE MT - 20	0.072	0.073	0.029	0.076	0.071	0.030
Basu-Lindsay MT - 5	0.119	0.114	0.039	0.131	0.121	0.029
Basu-Lindsay MT - 10	0.109	0.106	0.033	0.119	0.100	0.038
Basu-Lindsay MT - 15	0.107	0.103	0.030	0.112	0.097	0.033
Beran RIG - 0.01	0.077	0.080	0.026	0.066	0.063	0.029
Beran RIG - 0.1	0.105	0.104	0.025	0.112	0.108	0.038
Beran RIG - SJ	0.157	0.032	0.032	0.193	0.180	0.053
Beran MT - 5	0.182	0.183	0.025	0.207	0.202	0.032
Beran MT - 10	0.127	0.127	0.028	0.153	0.146	0.037
Beran MT - 15	0.102	0.104	0.029	0.126	0.121	0.036
DφDE	0.060	0.057	0.024	0.091	0.088	0.027
MDP 0.75	0.103	0.083	0.067	0.097	0.083	0.065
MDP 0.5	0.068	0.067	0.029	0.069	0.067	0.028
MDP 0.25	0.062	0.058	0.026	0.064	0.062	0.029
MDP 0.1	0.061	0.059	0.024	0.076	0.072	0.027
MLE	0.060	0.056	0.024	0.096	0.094	0.024

Table 1.12: The mean value with the standard deviation of the TVA committed in a 100-run experiment on a two-component Weibull mixture ($\lambda = 0.35, \nu_1 = 0.5, \nu_2 = 3$). The escort parameter of the DφDE is taken as the new MDφDE with the SJ bandwidth choice.

Estimation method	No Outliers						10% Outliers					
	λ	sd(λ)	ν_1	sd(ν_1)	ν_2	sd(ν_2)	λ	sd(λ)	ν_1	sd(ν_1)	ν_2	sd(ν_2)
Hellinger												
Classical MDφDE	0.367	0.102	0.550	0.104	1.054	0.194	0.352	0.158	0.273	0.050	1.051	0.407
New MDφDE - 0.01	0.445	0.103	0.562	0.135	1.212	0.284	0.409	0.133	0.464	0.156	1.148	0.293
New MDφDE - 0.1	0.432	0.101	0.502	0.141	1.139	0.241	0.460	0.210	0.378	0.125	1.114	0.302
New MDφDE - SJ	0.431	0.101	0.485	0.141	1.127	0.244	0.487	0.216	0.356	0.108	1.110	0.309
New MDφDE MT - 5	0.350	0.158	0.619	0.134	1.006	0.211	0.436	0.313	0.375	0.121	1.245	1.177
New MDφDE MT - 10	0.338	0.148	0.643	0.135	1.019	0.167	0.474	0.322	0.409	0.140	1.150	0.516
New MDφDE MT - 15	0.335	0.148	0.658	0.135	1.029	0.161	0.456	0.321	0.411	0.146	1.292	1.689
Basu-Lindsay MT - 5	0.392	0.178	0.734	0.122	1.042	0.022	0.351	0.225	0.757	0.177	1.048	0.026
Basu-Lindsay MT - 10	0.340	0.149	0.742	0.103	1.037	0.024	0.260	0.175	0.712	0.147	1.039	0.024
Basu-Lindsay MT - 15	0.340	0.149	0.742	0.103	1.037	0.024	0.222	0.126	0.696	0.125	1.043	0.016
Beran - 0.01	0.370	0.098	0.685	0.091	1.125	0.188	0.381	0.211	0.572	0.183	1.058	0.215
Beran - 0.1	0.234	0.093	0.747	0.113	1.028	0.118	0.419	0.372	0.479	0.211	1.181	0.553
Beran RIG - SJ	0.211	0.185	0.745	0.130	1.034	0.230	0.259	0.331	0.367	0.181	1.105	0.542
Beran MT - 5	0.302	0.205	0.584	0.129	0.867	0.120	0.471	0.388	0.376	0.128	1.097	0.738
Beran MT - 10	0.327	0.175	0.610	0.132	0.929	0.121	0.490	0.347	0.394	0.131	1.155	0.803
Beran MT - 15	0.331	0.165	0.623	0.128	0.962	0.128	0.470	0.340	0.400	0.132	1.174	0.893
DφDE	0.371	0.111	0.544	0.100	1.064	0.240	0.473	0.293	0.382	0.175	1.431	1.818
MDPD 0.75	0.494	0.181	0.619	0.089	1.341	0.689	0.505	0.243	0.625	0.087	1.313	0.641
MDPD 0.5	0.413	0.134	0.577	0.101	1.143	0.349	0.412	0.255	0.582	0.101	1.059	0.358
MDPD 0.25	0.366	0.108	0.542	0.110	1.064	0.349	0.554	0.348	0.503	0.117	1.205	0.995
MDPD 0.1	0.368	0.109	0.539	0.106	1.059	0.237	0.451	0.322	0.370	0.111	1.280	1.407
MLE (EM)	0.372	0.108	0.549	0.100	1.055	0.192	0.417	0.194	0.291	0.073	1.114	0.468

Table 1.13: The mean value and the standard deviation of the estimates in a 100-run experiment in a two-component Weibull mixture ($\lambda = 0.35, \nu_1 = 0.5, \nu_2 = 1$). The escort parameter of the DφDE is taken as the new MDφDE with Silverman's rule.

Estimation method	No Outliers			10% Outliers		
	mean	median	sd	mean	median	sd
Hellinger						
Classical MDφDE	0.056	0.055	0.026	0.124	0.114	0.035
New MDφDE RIG - 0.01	0.079	0.073	0.039	0.090	0.082	0.044
New MDφDE RIG - 0.1	0.079	0.065	0.059	0.112	0.101	0.050
New MDφDE RIG - SJ	0.076	0.065	0.041	0.129	0.117	0.065
New MDφDE MT - 5	0.063	0.058	0.029	0.114	0.095	0.041
New MDφDE MT - 10	0.067	0.063	0.028	0.112	0.102	0.038
New MDφDE MT - 15	0.069	0.067	0.028	0.111	0.105	0.036
Basu-Lindsay MT - 5	0.095	0.067	0.078	0.118	0.087	0.088
Basu-Lindsay MT - 10	0.094	0.074	0.073	0.112	0.088	0.080
Basu-Lindsay MT - 15	0.093	0.072	0.067	0.103	0.088	0.063
Beran RIG 0.01	0.079	0.081	0.028	0.089	0.087	0.033
Beran RIG 0.1	0.087	0.085	0.023	0.103	0.102	0.025
Beran RIG - SJ	0.094	0.092	0.023	0.100	0.097	0.021
Beran MT - 5	0.061	0.060	0.022	0.127	0.134	0.044
Beran MT - 10	0.059	0.055	0.025	0.115	0.096	0.041
Beran MT - 15	0.060	0.056	0.025	0.112	0.097	0.039
DφDE	0.057	0.055	0.028	0.117	0.113	0.034
MDPD 0.75	0.102	0.091	0.050	0.093	0.088	0.039
MDPD 0.5	0.072	0.067	0.032	0.075	0.074	0.033
MDPD 0.25	0.061	0.056	0.028	0.092	0.090	0.039
MDPD 0.1	0.058	0.055	0.027	0.108	0.087	0.039
MLE	0.056	0.055	0.026	0.122	0.117	0.029

Table 1.14: The mean value with the standard deviation of errors committed in a 100-run experiment on a two-component Weibull mixture ($\lambda = 0.35, \nu_1 = 0.5, \nu_2 = 1$). The escort parameter of the DφDE is taken as the new MDφDE with the SJ bandwidth choice.

1.7.5 Concluding remarks and comments

We summarize the most important remarks based on our simulations presented above.

- Our kernel-based MDφDE gave very good results in all situations. It has the best performance especially in difficult situations, and the lowest sensitivity to the choice of the kernel and the window in the class of $\varphi-$divergence-based estimators. In comparison to the MDPD, results were close in the Gaussian and the Weibull mixture, and the MDPD had slightly better results. In the GPD model, our kernel-based MDφDE had clearly better results than the MDPD making it a good competitor.

- The execution time of the compared methods varies. Both the classical MDφDE and the Basu-Lindsay approach were the most time consuming. The MLE and the MDPD were the best in execution time, whereas both our new kernel-based MDφDE and Beran's approach were in the middle with close execution time.

- Both the MLE and the classical MDφDE have the best performance under the model even in *difficult* models with heavy tails where kernel-based approaches could not give a satisfactory result. In regular situations such as the Gaussian mixture model, all methods were equivalent under the model.

- When contamination is present, the compared estimators gave results as expected. Both the MLE and the classical MDφDE are not robust against contamination. The DφDE guided by our kernel-based MDφDE gave very good results under the model. However, when contamination is present, there was no improvement and sometimes a deterioration in the performance in comparison to the escort parameter. This is the case of the Weibull mixtures and the GPD model. The obtained results are still better than MLE and the classical MDφDE.

- The Basu-Lindsay approach worked very well in regular situations and even showed a slight improvement in efficiency in comparison to the Beran's method which is concordant to the result of Basu and Sarkar [1994]. It gave surprisingly good results in the GPD model under contamination when we used the varying KDE in comparison to the situation under the model. Unfortunately, it did not give satisfactory results in the Weibull mixtures. This method seems very sensitive to the kernel under difficult situations since the model is already influenced by the kernel creating a loss of information.

- The minimum density power divergence gave very good results in all situations except for the GPD. The best tradeoff parameter from our set of candidates was $a = 0.5$.

- The Beran's method gave very good tradeoff (and many times the best) between robustness and performance under the model in most of the situations, but not very well in the GPD model. The best choice of the kernel for the GPD and the Weibull mixture was the RIG with window 0.01. It was sensitive to the choice of the kernel and its window in many situations.

- The applicability of our kernel-based MDφDE to multivariate situations is bound by the use of integration in higher dimensions which is the case of other $\varphi-$divergence-based estimators and the case of the MDPD when applied to mixture models except for the L2 distance (a=1) which still has its limitations. A general solution is to use Monte-Carlo approximation for the integral.

- The results obtained using a fixed window for symmetric or asymmetric kernels give rise to an interesting question about the choice of the window. This question will be discussed in future work.

1.8 Appendix: Proofs

1.8.1 Proof of Theorem 1.6.2

Proof. Let $\varepsilon > 0$. We want to prove that $\lim_{n\to\infty} \mathbb{P}\left(\sup_{\phi\in\Phi} |P_n H(P_n,\phi) - P_n h(P_T,\phi)| < \varepsilon\right) = 1$. Since φ is twice differentiable (which also implies the differentiability of $\varphi^{\#}$), then by the mean value theorem, there exist two functions $\lambda_1, \lambda_2 : \mathbb{R} \to (0,1)$ such that:

$$\begin{aligned}
\varphi'\left(\frac{p_\phi}{K_{w,n}}\right)(x) - \varphi'\left(\frac{p_\phi}{p_T}\right)(x) &= \varphi''\left(\lambda_1(x)\frac{p_\phi}{K_{w,n}}(x) + (1-\lambda_1(x))\frac{p_\phi}{p_T}(x)\right)\left[\frac{p_\phi}{K_{w,n}}(x) - \frac{p_\phi}{p_T}(x)\right], \\
&= \varphi''\left(\lambda_1(x)\frac{p_\phi}{K_{w,n}}(x) + (1-\lambda_1(x))\frac{p_\phi}{p_T}(x)\right)\frac{p_\phi}{K_{w,n}p_T}(x)\left[p_T - K_{w,n}\right](x) \\
&= \mathcal{A}_n(x,\phi)\left[p_T - K_{w,n}\right](x) \\
\varphi^{\#}\left(\frac{p_\phi}{K_{w,n}}\right)(y_i) - \varphi^{\#}\left(\frac{p_\phi}{p_T}\right)(y_i) &= \left(\varphi^{\#}\right)'\left(\lambda_2(y_i)\frac{p_\phi}{K_{w,n}}(y_i) + (1-\lambda_2(y_i))\frac{p_\phi}{p_T}(y_i)\right)\left[\frac{p_\phi}{K_{w,n}} - \frac{p_\phi}{p_T}\right](y_i) \\
&= \left(\varphi^{\#}\right)'\left(\lambda_2(y_i)\frac{p_\phi}{K_{w,n}}(y_i) + (1-\lambda_2(y_i))\frac{p_\phi}{p_T}(y_i)\right)\frac{p_\phi}{K_{w,n}p_T}(y_i) \\
&\qquad \times \left[p_T - K_{w,n}\right](y_i) \\
&= \mathcal{B}_n(y_i,\phi)\left[p_T - K_{w,n}\right](y_i).
\end{aligned}$$

We denoted:

$$\begin{aligned}
\mathcal{A}_n(x,\phi) &= \varphi''\left(\lambda_1(x)\frac{p_\phi}{K_{w,n}}(x) + (1-\lambda_1(x))\frac{p_\phi}{p_T}(x)\right)\frac{p_\phi}{K_{w,n}p_T}(x) \\
\mathcal{B}_n(y_i,\phi) &= \left(\varphi^{\#}\right)'\left(\lambda_2(y_i)\frac{p_\phi}{K_{w,n}}(y_i) + (1-\lambda_2(y_i))\frac{p_\phi}{p_T}(y_i)\right)\frac{p_\phi}{K_{w,n}p_T}(y_i).
\end{aligned}$$

Let n be sufficiently large such that:

$$\sup_x |K_w * P_n(x) - p_T(x)| \leq \min\left(\varepsilon, \frac{\varepsilon}{\mathcal{A}_n}, \frac{\varepsilon}{\mathcal{B}_n}\right)$$

where $\mathcal{A}_n = \sup_\phi \int \mathcal{A}_n(x)dx$ and $\mathcal{B}_n = \sup_\phi \frac{1}{n}\sum \mathcal{B}_n(y_i)$ which exist by virtue of assumption 3 of the present theorem on the one hand and on the other hand the fact that functions $x \mapsto \lambda_1(x)$ and $x \mapsto \lambda_2(x)$ are bounded uniformly inside $(0,1)$. This event occurs with probability $1-\eta_n$ with $\eta_n \to 0$ by the strong consistency assumption (point 2). This implies that both events:

$$\begin{aligned}
\left|\int\left[\varphi'\left(\frac{p_\phi}{K_w * P_n}\right) - \varphi'\left(\frac{p_\phi}{p_T}\right)\right]p_\phi\right| &\leq \varepsilon, \\
\left|\frac{1}{n}\sum_{i=1}^{n}\varphi^{\#}\left(\frac{p_\phi}{K_w * P_n}\right)(y_i) - \varphi^{\#}\left(\frac{p_\phi}{p_T}\right)(y_i)\right| &\leq \varepsilon
\end{aligned}$$

happen with probability greater than $1-\eta_n$ independently of ϕ. Finally, we conclude that

$$\mathbb{P}\left(\sup_{\phi\in\Phi}|P_nH(P_n,\phi)-P_nh(P_T,\phi)|<2\varepsilon\right)\geq 1-\eta_n,$$

and hence $\sup_{\phi\in\Phi}|P_nH(P_n,\phi)-P_nh(P_T,\phi)|\to 0$ in probability. To end the proof, we use assumption 3 together with the result in Example 19.9 in van der Vaart [1998] Chap. 19 which imply that $\{\varphi^{\#}\left(\frac{p_\phi}{p_T}\right),\phi\in\Phi\}$ is a Glivenko-Cantelli class of functions. Hence, $\sup_{\phi\in\Phi}|P_Th(P_T,\phi)-P_nh(P_T,\phi)|\to 0$ in probability. Using inequality (1.6.3), we conclude that $\sup_{\phi\in\Phi}|P_nH(P_n,\phi)-P_Th(P_T,\phi)|\to 0$ in probability. Finally, the previous arguments prove the first point (1.6.1) in Theorem 1.6.1. The second point in Theorem 1.6.1 is the same as assumption 4 of the present theorem. By definition of the kernel-based MDφDE as a minimum of the criterion function $\phi\mapsto P_nH(P_n,\phi)$, Theorem 1 entails the consistency of our new estimator. □ □

1.8.2 Proof of Theorem 1.6.3

We follow the same idea of the proof of Theorem 1.6.2. In order to treat the second term in the right hand side of equation (1.6.4), we use the uniform continuity of function $t\mapsto t^{-\gamma}$. Indeed, if $|K_w*P_n(x)-p_T(x)|<\delta_2$, then:

$$|(K_w*P_n)^{-\gamma}(y_i)-p_T^{-\gamma}(y_i)|<\frac{\varepsilon}{\sup_\phi\frac{1}{n}\sum p_\phi^{\gamma}(y_i)}.$$

By the consistency of the kernel estimator, the previous inequality happens with probability $1-\eta_n$ with $\eta_n\to 0$. Thus, $\mathcal{B}_n$ of Theorem 1.6.2 is now replaced by the simpler quantity

$$\mathcal{B}_n=\sup_\phi\frac{1}{n}\sum_{i=1}^{n}p_\phi^{\gamma}(y_i). \tag{1.8.1}$$

On the other hand, in order to treat the first term in the right hand side of equation (1.6.4), we rewrite the integral as follows:

$$\int\frac{(K_w*P_n)^{-\gamma+1}(x)-p_T^{-\gamma+1}(x)}{p_\phi^{-\gamma}(x)}dx=$$
$$\int\frac{\left[(K_w*P_n)^{\frac{-\gamma+1}{2}}(x)-p_T^{\frac{-\gamma+1}{2}}(x)\right]\left[(K_w*P_n)^{\frac{-\gamma+1}{2}}(x)+p_T^{\frac{-\gamma+1}{2}}(x)\right]}{p_\phi^{-\gamma}(x)}dx.$$

Now, using the uniform continuity of function[17] $t\mapsto t^{\frac{-\gamma+1}{2}}$, we may deduce that if $|K_w*P_n(x)-p_T(x)|<\delta_1$, then:

$$|(K_w*P_n)^{\frac{-\gamma+1}{2}}(x)-p_T^{\frac{-\gamma+1}{2}}(x)|<\frac{\varepsilon}{\sup_\phi\int\frac{(K_w*P_n)^{\frac{-\gamma+1}{2}}(x)+p_T^{\frac{-\gamma+1}{2}}(x)}{p_\phi^{-\gamma}(x)}dx}. \tag{1.8.2}$$

Again, by the consistency of the kernel estimator, the previous inequality happens with probability $1-\eta_n$ with $\eta_n\to 0$. Thus $\mathcal{A}_n$ of Theorem 1.6.2 is now replaced by the quantity

$$\mathcal{A}_n=\sup_\phi\int\frac{(K_w*P_n)^{\frac{-\gamma+1}{2}}(x)+p_T^{\frac{-\gamma+1}{2}}(x)}{p_\phi^{-\gamma}(x)}dx.$$

[17] notice that $\frac{-\gamma+1}{2}\in(0,1)$ since $\gamma\in(-1,0)$.

Existness and finitness of both $\mathcal{A}_n$ and $\mathcal{B}_n$ in probability are ensured by assumptions 3 and 4. Now, using inequalities (1.8.2) and (1.8.1), both events

$$\begin{aligned}\left|\int \frac{(K_w * P_n)^{1-\gamma} - p_T^{1-\gamma}}{p_\phi^{-\gamma}}(x)dx\right| &< \varepsilon;\\ \left|\frac{1}{n}\sum_{i=1}^{n}\frac{(K_w * P_n)^{-\gamma} - p_T^{-\gamma}}{p_\phi^{-\gamma}}(y_i)\right| &< \varepsilon,\end{aligned}$$

happen with probability greater than $1-\eta_n$ independently of ϕ. Finally, we conclude that

$$\mathbb{P}\left(\sup_{\phi\in\Phi}|P_nH(P_n,\phi) - P_nh(P_T,\phi)| < 2\varepsilon\right) \geq 1-\eta_n,$$

and hence $\sup_{\phi\in\Phi}|P_nH(P_n,\phi) - P_nh(P_T,\phi)| \to 0$ in probability. To end the proof, we use assumption 2 together with the Glivenko-Cantelli theorem to deduce that $\sup_{\phi\in\Phi}|P_Th(P_T,\phi) - P_nh(P_T,\phi)| \to 0$ in probability. Using inequality (1.6.3), we conclude that $\sup_{\phi\in\Phi}|P_nH(P_n,\phi) - P_Th(P_T,\phi)| \to 0$ in probability. Finally, the previous arguments prove the first point (1.6.1) in Theorem 1.6.1. The second point in Theorem 1.6.1 is the same as assumption 4 of the present theorem. By definition of the kernel-based MDφDE as a minimum of the criterion function $\phi \mapsto P_nH(P_n,\phi)$, Theorem 1 entails the consistency of our new estimator. □

1.8.3 Proof of Theorem 1.6.4

In the whole proof, the index T will be omitted from ϕ^T for the sake of clarity. We start with calculating the gradient $\nabla P_nH(P_n,\phi)$.

$$\nabla P_nH(P_n,\phi) = \frac{\gamma}{\gamma-1}\int \nabla p_\phi \frac{p_\phi^{\gamma-1}}{K_{n,w}^{\gamma-1}}dx - \frac{1}{n}\sum_{i=1}^{n}\nabla p_\phi \frac{p_\phi^{\gamma-1}}{K_{n,w}^{\gamma}}(y_i). \tag{1.8.3}$$

We treat each term separately. The first term can be rewritten as:

$$\int \nabla p_\phi \frac{p_\phi^{\gamma-1}}{K_{n,w}^{\gamma-1}}dx = \int \nabla p_\phi p_\phi^{\gamma-1}\left[K_{n,w}^{1-\gamma} - p_\phi^{1-\gamma}\right]dx + \int \nabla p_\phi dx.$$

The second term in the right hand side is zero because p_ϕ is a density, provided changeability between integration and differentiation. For the first term, we write a second order Taylor expansion of function $t \mapsto t^{1-\gamma}$:

$$K_{n,w}^{1-\gamma} - p_\phi^{1-\gamma} = (1-\gamma)(K_{n,w} - p_\phi)p_\phi^{-\gamma} + \frac{-\gamma}{2}(K_n - p_\phi)^2 M_n(x)^{-\gamma-1},$$

where $M_n(x)$ is a point in between $K_{n,w}(x)$ and $p_\phi(x)$. We now have:

$$\begin{aligned}\int \nabla p_\phi p_\phi^{\gamma-1}\left[K_{n,w}^{1-\gamma} - p_\phi^{1-\gamma}\right]dx = (1-\gamma)\int \nabla p_\phi p_\phi^{-1}\left[K_{n,w} - p_\phi\right]dx + \\ \frac{-\gamma}{2}\int \nabla p_\phi p_\phi^{-1} M_n(x)^{-\gamma-1}\left[K_{n,w} - p_\phi\right]^2 dx.\end{aligned} \tag{1.8.4}$$

Using equations (3.11-3.13) from Beran [1977], we may write:

$$\sqrt{n}\int \nabla p_\phi p_\phi^{-1}\left[K_{n,w} - p_\phi\right]dx \underset{\mathcal{L}}{\to} \mathcal{N}(0,S), \tag{1.8.5}$$

where $S = \int \nabla p_\phi \nabla p_\phi^t dx$.
The second term will be handled in a similar way to equations (3.11-3.13) from Beran [1977]. Let $K_n(x) = K(x/w_n)/w_n$. Write

$$\int \nabla p_\phi p_\phi^{-1} M_n(x)^{-\gamma-1} \left[K_{n,w} - p_\phi\right]^2 dx = \int \nabla p_\phi p_\phi^{-1} M_n(x)^{-\gamma-1} \left[K_{n,w} - K_n * P_\phi\right]^2 dx +$$
$$2\int \nabla p_\phi p_\phi^{-1} M_n(x)^{-\gamma-1} \left[K_{n,w} - K_n * P_\phi\right]\left[K_n * P_\phi - p_\phi\right] dx + \int \nabla p_\phi p_\phi^{-1} M_n(x)^{-\gamma-1} \left[K_n * P_\phi - p_\phi\right]^2 dx, \tag{1.8.6}$$

and prove that each term has a limit equal to zero when multiplied by $\sqrt{n}$. There are two essential arguments. The first one uses equation (3.11) from Beran [1977] to write:

$$\sup_x |K_n * P_\phi - p_\phi| \leq \frac{w^2}{2} \sup_x \left|p_\phi''(x)\right| \int x^2 K(x) dx. \tag{1.8.7}$$

The second one is a result of Corollary 5 from Wied and Weibßach [2012]:

$$\lim_{n\to\infty} \sqrt{\frac{nw}{-2\log w}} \sup_x \frac{|K_{n,w} - K_n * P_\phi|}{\sqrt{p_\phi}} = \left(\int K^2(y) dy\right). \tag{1.8.8}$$

We treat the first term in equation (1.8.6) using equation (1.8.8).

$$\begin{aligned}\sqrt{n}\int |\nabla p_\phi| \, p_\phi^{-1} M_n(x)^{-\gamma-1} \left[K_{n,w} - K_n * P_\phi\right]^2 dx &\leq \left[\sqrt{\frac{nw}{-2\log w}} \sup_x \frac{|K_{n,w} - K_n * P_\phi|}{\sqrt{p_\phi}}\right]^2 \\ &\quad \times \frac{-2\log(w)}{n^{1/2} w} \int |\nabla p_\phi| \, M_n(x)^{-\gamma-1} dx \\ &= \mathcal{O}\left(\frac{-2\log(w)}{n^{1/2} w}\right).\end{aligned}$$

We treat the second term in equation (1.8.6) using equations (1.8.8) and (1.8.7).

$$\begin{aligned}\sqrt{n}\int \frac{|\nabla p_\phi|}{p_\phi M_n(x)^{\gamma+1}} \left[K_{n,w} - K_n * P_\phi\right]\left[K_n * P_\phi - p_\phi\right] dx &\leq \sqrt{\frac{nw}{-2\log w}} \sup_x \frac{|K_{n,w} - K_n * P_\phi|}{\sqrt{p_\phi}} \\ \times \sup_x \left|p_\phi''(x)\right| \int x^2 K(x) dx \sqrt{-2\log(w)} &\xrightarrow{\frac{w^{3/2}}{2}} \int |\nabla p_\phi| p_\phi^{-1/2} M_n(x)^{-\gamma-1} dx \\ &= \mathcal{O}\left(w^{3/2}\sqrt{-2\log(w)}\right).\end{aligned}$$

We treat the third term in equation (1.8.6) using equation (1.8.7).

$$\begin{aligned}\sqrt{n}\int |\nabla p_\phi| p_\phi^{-1} M_n(x)^{-\gamma-1} \left[K_n * P_\phi - p_\phi\right]^2 dx &\leq \sqrt{n}\frac{h^4}{2} \sup_x \left|p_\phi''(x)\right| \left[\int x^2 K(x) dx\right]^2 \\ &\quad \times \int \frac{|\nabla p_\phi|}{p_\phi M_n(x)^{\gamma+1}} dx \\ &= \mathcal{O}\left(n^{1/2} h^4\right).\end{aligned}$$

We conclude using assumption 3 that :

$$\sqrt{n}\int \nabla p_\phi p_\phi^{-1} M_n(x)^{-\gamma-1} \left[K_{n,w} - p_\phi\right]^2 dx \underset{\mathbb{P}}{\to} 0.$$

This entails together with (1.8.5) that the first term in $P_nH(P_n,\phi)$ multiplied by $\sqrt{n}$ is a centered multivariate Gaussian with covariance matrix S.
The sum term in $P_nH(P_n,\phi)$ can be treated similarly. Firstly, write:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla p_\phi\frac{p_\phi^{\gamma-1}}{K_{n,w}^{\gamma}}(y_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla p_\phi p_\phi^{\gamma-1}\left[K_{n,w}^{-\gamma}-p_\phi^{-\gamma}\right](y_i)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\nabla p_\phi}{p_\phi}(y_i).$$

Now the second term in the right hand side, is asymptotically Gaussian with mean zero and covariance matrix equal to S. For the first term, we apply the mean value theorem on function $z^{-\gamma}$. There exists a bounded function $M_n(y_i)$ in between $K_{n,w}(y_i)$ and $p_\phi(y_i)$ such that:

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla p_\phi p_\phi^{\gamma-1}\left[K_{n,w}^{-\gamma}-p_\phi^{-\gamma}\right](y_i)=&\frac{-\gamma}{\sqrt{n}}\sum_{i=1}^{n}\frac{\nabla p_\phi}{p_\phi^2}\left[K_{n,w}-p_\phi\right](y_i)\\&+\frac{\gamma(\gamma+1)}{\sqrt{n}}\sum_{i=1}^{n}\frac{\nabla p_\phi}{p_\phi^3}\left[K_{n,w}-p_\phi\right]^2(y_i)\end{aligned}$$

The treatment of the second term in the right hand side can be done similarly to the second term in equation (1.8.4) and thus converges to zero when multiplied by $\sqrt{n}$. The first term will be proved to have the same asymptotic behavior to the first term in equation (1.8.4). Write the difference between these terms. Let $\psi(x)=\frac{\nabla p_\phi}{p_\phi^2}$.

$$\begin{aligned}\sqrt{n}\left|\frac{-1}{\sqrt{n}}\sum_{i=1}^{n}\psi(x)\left[K_{n,w}-p_\phi\right](y_i)-\int\psi(x)\left[K_{n,w}-p_\phi\right]p_\phi(x)dx\right| &\leq\\ \sup_x|K_{n,w}(x)-p_\phi(x)|\sqrt{n}\left|\int\psi(x)\left[K_{n,w}-p_\phi\right](dP_n-dP_\phi)(x)\right| &=\\ \sqrt{n}\left|\int(\mathbb{F}_n-\mathbb{F}_\phi)(x)d\left(\psi(x)\left[K_{n,w}-p_\phi\right]\right)(x)\right| &\leq\\ \sqrt{n}\sup|\mathbb{F}_n-\mathbb{F}_\phi|\left[\sup|K'_{n,w}-p'_\phi|\int\psi(x)dx+\sup|K_{n,w}-p_\phi|\int\psi'(x)dx\right] &.\end{aligned}$$

Now, using rates of convergence of the empirical distribution function (see for example van der Vaart [1998] p. 268), the kernel density estimator (see for example Bordes and Vandekerkhove [2010] Lemma 3.1) and the derivative of the kernel density estimator (see Schuster [1969] Theorem 2.5), we prove easily that the right hand side of the inequality in the previous display tends to zero in probability. This proves our claim. Now it remains to use the asymptotic normality limit in equation (1.8.5) to deduce that:

$$\sqrt{n}\left(\frac{-1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\nabla p_\phi}{p_\phi^2}\left[K_{n,w}-p_\phi\right](y_i)\right)\xrightarrow[\mathcal{L}]{}\mathcal{N}(0,S).$$

Collecting the three pieces which generate the asymptotic normality in the whole calculus, we may conclude that:

$$\sqrt{n}\nabla P_nH(P_n,\phi)\xrightarrow[\mathcal{L}]{}\mathcal{N}\left(0,(2\gamma^2+1)\int\nabla p_\phi\nabla p_\phi^t\right).$$

The matrix of second order partial derivatives $J_{P_nH(P_n,.)}$ can be treated in an easier way than the vector $\nabla P_nH(P_n,\phi)$. It can be shown using similar techniques to those used here above that $J_{P_nH(P_n,.)}$ converges in probability at rate $o_P(n^{-1/2})$. We may conclude now that the asymptotic normality result (1.6.8) holds.

1.8.4 Proof of Theorem 1.6.5

For a clearer writing, we omit the index T from P_T in this proof. Deriving the left hand side of the estimating equation (1.6.9) gives:

$$\frac{\gamma}{\gamma-1}\int \frac{\left[(\gamma-1)\nabla p_{C(P)}\left(\nabla p_{C(P)}\right)^t + p_{C(P)}J_{p_{C(P)}}\right]p_{C(P)}^{\gamma-2}}{(K_w*P)^{\gamma-1}}\mathrm{IF}(P,Q) \\ -\gamma\int \frac{p_{C(P)}^{\gamma-1}\left[K_w*(Q-P)\right]\nabla p_{C(P)}}{(K_w*P)^{\gamma}}(x)dx.$$

Deriving the right hand side of the estimating equation (1.6.9) gives:

$$\int \frac{\left[(\gamma-1)\nabla p_{C(P)}\left(\nabla p_{C(P)}\right)^t p_{C(P)}^{\gamma-2} + p_{C(P)}^{\gamma-1}J_{p_{C(P)}}\right]}{(K_w*P)^{\gamma}}(x)dP(x)\mathrm{IF}(P,Q) \\ -\gamma\int \frac{p_{C(P)}^{\gamma-1}\left[K_w*(Q-P)\right]\nabla p_{C(P)}}{(K_w*P)^{\gamma+1}}(x)dP(x) + \int \frac{p_{C(P)}^{\gamma-1}\nabla p_{C(P)}}{(K_w*P)^{\gamma}}(x)(dQ-dP)(x).$$

We have now:

$$A\ \mathrm{IF}(P,Q) = \gamma\int \frac{p_{C(P)}^{\gamma-1}\left[K_w*(Q-P)\right]\nabla p_{C(P)}}{(K_w*P)^{\gamma}}(x)dx + \int \frac{p_{C(P)}^{\gamma-1}\nabla p_{C(P)}}{(K_w*P)^{\gamma}}(x)(dQ-dP)(x) \\ -\gamma\int \frac{p_{C(P)}^{\gamma-1}\left[K_w*(Q-P)\right]\nabla p_{C(P)}}{(K_w*P)^{\gamma+1}}(x)dP(x),$$

where A is defined by formula (1.6.13). Assuming that A is invertible and using the estimating equation (1.6.9), we can write:

$$\mathrm{IF}(P,Q) = \gamma A^{-1}\int \frac{p_{C(P)}^{\gamma-1}\left[K_w*Q\right]\nabla p_{C(P)}}{(K_w*P)^{\gamma}}\left(1-\frac{p}{K*P}\right)(x)dx + A^{-1}\int \frac{p_{C(P)}^{\gamma-1}\nabla p_{C(P)}}{(K_w*P)^{\gamma}}(x)dQ(x).$$

The remaining of the proof is a simple substitution of $C(P)$ by ϕ^T when $P = P_{\phi^T}$, and replacing Q by the dirac measure on a point x_0. □

Chapter 2

Iterative Proximal-Point Algorithm for the Calculus of Divergence-Based Estimators with Application to Mixture Models

In the previous chapter, we have presented and introduced several estimators; an estimator based on Beran's approach (1.2.1), an estimator based on the Basu-Lindsay approach (1.2.2), the MDφDE (1.3.6), the DφDE (1.3.7), our new kernel-based MDφDE (1.5.4) and the MDPD (1.3.8). All these estimators, the MLE included, are in general non convex (or non concave for the DφDE) optimization problems. The calculus of these estimators in general is then not guaranteed to give a good result for a finite-sample setup when we use any standard optimization algorithm. There exist several optimization algorithms such as Gradient descent algorithms (first and second order gradient descent and gradient-conjugate algorithms), the BFGS algorithm, the Nelder-Mead's algorithm, Brent's algorithm among others, see Lange [2013]. These algorithms guarantee the convergence of the iterative procedure to a global optimum whose objective function it is a strictly convex (or concave) function. If it is not the case, the algorithm converges to a local optimum. Each optimization method has its own advantages and drawbacks. There are also some algorithms which treat functions which can be written as the difference of two convex functions called convex-concave optimization algorithms, see Yuille and Rangarajan [2003]. These algorithms, for example, give in general better results than convex optimization algorithms for this kind of functions.

There is on the other hand, another type of optimization algorithms which attack a modified version of the objective function, say $D(\phi) + g(\phi, \phi^k)$, where D is the objective function and g is a perturbation function which depends on the current iteration k. A perturbation of the objective function has a goal of giving it a "better form". The iterative procedure then proceeds to optimize the modified function iteratively as the perturbation becomes less and less important as the number of the iteration increases. This kind of algorithms is called proximal-point algorithms. It was first proposed by Martinet [1970] who used a perturbation of the form $g(\phi, \phi^k) = \|\phi - \phi^k\|$. Generally, the proximal term has a regularization effect in the sense that a proximal point algorithm is more stable and frequently outperforms classical optimization algorithms, see Goldstein and Russak [1987]. Furthermore, and as mentioned in [Chrétien and Hero, 2008], proximal point algorithms permit to avoid saddle points.

The EM algorithm is a very interesting example of proximal point algorithms, see paragraph 2.1.2 for a detailed calculus or the papers of Chrétien and Hero [1998] and Tseng [2004]. Indeed, one may rewrite the conditional expectation of the complete log-likelihood as a sum of the log-likelihood function and a distance-like function over the conditional densities of the labels provided an observation. Thus, the EM algorithm has the log-likelihood as an objective function which is being perturbed by a distance-like function. Chrétien and Hero Chretien and Hero [1998] proved superlinear convergence of a proximal point algorithm derived by the EM algorithm. Notice that EM-type algorithms usually enjoy no more than linear convergence.

Taking into consideration the need for robust estimators, and the fact that the MLE is the least robust estimator among the class of divergence-type estimators, we generalize the EM algorithm (and the version in Tseng [2004]) by replacing the log-likelihood function by an estimator of a $\varphi-$divergence between the true distribution of the data and the model. We, thus, propose to calculate divergence-based estimators mentioned here above using a proximal-point algorithm based on the work of Tseng [2004] on the log-likelihood function. This proximal-point algorithm extends the EM algorithm. Our convergence proof of the iterative procedure requires some regularity of the estimated divergence with respect to the parameter vector which can be easily checked using Lebesgue theorems except for the dual formula (1.3.5). Indeed, the supremal form of the estimated divergence in the dual formula complicates the situation. Recent results in Rockafellar and Wets [1998] provide sufficient conditions to solve this problem. It may at time be very difficult to prove that the objective function is differentiable with respect to ϕ, therefore, our results cover the case when the objective function is not differentiable.
We also propose a two-step iterative algorithm to calculate divergence-based estimators for mixture models motivated by the EM algorithm; a step to calculate the proportion and a step to calculate the parameters of the components. Proofs for this simplified version become more technical. The goal of this simplification is to reduce the dimension over which we optimize since in lower dimensions, optimization procedures are more efficient[1]. Another contribution of this work concerns the assumptions ensuring the convergence of the algorithm. In the previous works on such type of proximal algorithms such as the papers of Tseng [2004] and Chrétien and Hero [1998], the proximal term is supposed to verify an identifiability property. In other words $g(\phi, \phi') = 0$ if and only if $\phi = \phi'$. We show that such property is difficult to verify and it is often not fulfilled in mixture models. We provide a way to relax such condition without imposing further assumptions.

2.1 Development of the proximal-point algorithm from the EM algorithm

2.1.1 General context and notations

Let (X, Y) be a couple of random variables with joint probability density function $f(x, y|\phi)$ parametrized by a vector of parameters $\phi \in \Phi \subset \mathbb{R}^d$. Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be n copies of (X, Y) independently and identically distributed. Finally, let $(x_1, y_1), \cdots, (x_n, y_n)$ be n realizations of the n copies of (X, Y). The x_i's are the unobserved data (labels) and

[1]This does not cover all optimization methods. For example, the Nelder-Mead algorithm is considered as "unreliable" in univariate optimization. The Brent method can be used as an alternative. Note that these two algorithms are suitable for non differentiable functions since they only use function values to reach an optimum.

the y_i's are the observations. The vector of parameters ϕ is unknown and need to be estimated.
The observed data y_i are supposed to be real vectors and the labels x_i belong to a space $\mathcal{X}$ not necessarily finite unless mentioned otherwise. Denote dx the measure on the label space $\mathcal{X}$ (for example the counting measure if $\mathcal{X}$ is discrete). The marginal density of the observed data is given by $p_\phi(y) = \int f(x, y|\phi)dx$.
For a parametrized function f with a parameter a, we write $f(x|a)$. We use the notation ϕ^k for sequences with the index above. Derivatives of a real valued function ψ defined on $\mathbb{R}$ are written as ψ', ψ'', etc. We use ∇f for the gradient of real function f defined on $\mathbb{R}^d$, ∂f to its subgradient and J_f to the matrix of second order partial derivatives. For a generic function H of two variables (ϕ, θ), $\nabla_1 H(\phi, \theta)$ denotes the gradient with respect to the first (vectorial) variable ϕ.

2.1.2 EM algorithm and Tseng's generalization

The EM algorithm is a well-known method for calculating the maximum likelihood estimator of a model where incomplete data is considered. For example, when working with mixture models in the context of clustering, the labels or classes of observations are unknown during the training phase. Several variants of the EM algorithm were proposed, see McLachlan and Krishnan [2007]. The EM algorithm estimates the unknown parameter vector by generating the sequence (see [Dempster et al., 1977]):

$$\begin{aligned} \phi^{k+1} &= \arg\max_{\Phi} Q(\phi, \phi^k) \\ &= \arg\max_{\Phi} \mathbb{E}\left[\log(f(\mathbf{X}, \mathbf{Y}|\phi)) \,\middle|\, \mathbf{Y} = \mathbf{y}, \phi^k\right], \end{aligned}$$

where $\mathbf{X} = (X_1, \cdots, X_n)$, $\mathbf{Y} = (Y_1, \cdots, Y_n)$ and $\mathbf{y} = (y_1, \cdots, y_n)$. By independence between the couples (X_i, Y_i)'s, the previous iteration may be rewritten as:

$$\begin{aligned} \phi^{k+1} &= \arg\max_{\Phi} \sum_{i=1}^{n} \mathbb{E}\left[\log(f(X_i, Y_i|\phi)) \,\middle|\, Y_i = y_i, \phi^k\right] \\ &= \arg\max_{\Phi} \sum_{i=1}^{n} \int_{\mathcal{X}} \log(f(x, y_i|\phi)) h_i(x|\phi^k)dx, \end{aligned} \tag{2.1.1}$$

where $h_i(x|\phi^k)$ is the conditional density of the labels (at step k) provided y_i. It is given by:

$$h_i(x|\phi^k) = \frac{f(x, y_i|\phi^k)}{p_{\phi^k}(y_i)}. \tag{2.1.2}$$

This justifies the recurrence equation given by [Tseng, 2004]. It is slightly different from the EM recurrence defined in [Dempster et al., 1977]. The conditional expectation of the logarithm of the complete likelihood provided the data and the parameter vector of the previous iteration is calculated, here, on the vector of observed data. The expectation is replaced by an integral against the corresponding conditional density of the labels.
It is well-known that the EM iterations can be rewritten as a difference between the log-likelihood and a *Kullback-Liebler* distance-like function. Indeed, using (2.1.2) in (2.1.1),

one can write:

$$\begin{aligned}
\phi^{k+1} &= \arg\max_{\Phi} \sum_{i=1}^{n} \int_{\mathcal{X}} \log\left(h_i(x|\phi) \times p_\phi(y_i)\right) h_i(x|\phi^k)dx \\
&= \arg\max_{\Phi} \sum_{i=1}^{n} \int_{\mathcal{X}} \log\left(p_\phi(y_i)\right) h_i(x|\phi^k)dx + \sum_{i=1}^{n} \int_{\mathcal{X}} \log\left(h_i(x|\phi)\right) h_i(x|\phi^k)dx \\
&= \arg\max_{\Phi} \sum_{i=1}^{n} \log\left(p_\phi(y_i)\right) + \sum_{i=1}^{n} \int_{\mathcal{X}} \log\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k)dx \\
&\qquad + \sum_{i=1}^{n} \int_{\mathcal{X}} \log\left(h_i(x|\phi^k)\right) h_i(x|\phi^k)dx.
\end{aligned}$$

The final line is justified by the fact that $h_i(x|\phi)$ is a density, therefore it integrates to 1. The additional term does not depend on ϕ and, hence, can be omitted. We now have the following iterative procedure:

$$\phi^{k+1} = \arg\max_{\Phi} \sum_{i=1}^{n} \log\left(p_\phi(y_i)\right) + \sum_{i=1}^{n} \int_{\mathcal{X}} \log\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k)dx. \tag{2.1.3}$$

As stated in [Tseng, 2004], the previous iteration has the form of a proximal point maximization of the log-likelihood, i.e. a perturbation of the log-likelihood by a (modified) Kullback distance-like function defined on the conditional densities of the labels. Tseng proposed to generalize the Kullback distance-like term into other types of divergences. Tseng's recurrence is now defined by:

$$\phi^{k+1} = \arg\sup_{\phi} J(\phi) - D_\psi(\phi, \phi^k), \tag{2.1.4}$$

where J is the log-likelihood function and D_ψ is a distance-like function defined on the conditional probabilities of the classes provided the observations and is given by:

$$D_\psi(\phi, \phi^k) = \sum_{i=1}^{n} \int_{\mathcal{X}} \psi\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k)dx, \tag{2.1.5}$$

for a real positive convex function ψ such that $\psi(1) = \psi'(1) = 0$. $D_\psi(\phi_1, \phi_2)$ is positive and equals zero if $\phi_1 = \phi_2$. Moreover, $D_\psi(\phi_1, \phi_2) = 0$ if and only if $\forall i, h_i(x|\phi_1) = h_i(x|\phi_2)$ $dx-$almost everywhere. Clearly, (2.1.4) and (2.1.3) are equivalent for $\psi(t) = -\log(t)+t-1$.

2.1.3 Generalization of Tseng's algorithm

We use the relation between maximizing the log-likelihood and minimizing the Kullback-Liebler divergence to generalize the previous algorithm. We therefore replace the log-likelihood function by a $\varphi-$divergence D_φ (in the sense of [Csiszár, 1963]) between the true density of the data p_{ϕ_T} and the model p_ϕ. Since the value of the divergence depends on the true density which is unknown, an estimator of the divergence needs to be considered. We may use any estimator among (1.2.1), (1.2.2), (1.3.5) or (1.5.3). Our new algorithm is defined by the following recurrence:

$$\phi^{k+1} = \arg\inf_{\phi} \hat{D}_\varphi(p_\phi, p_{\phi_T}) + \frac{1}{n} D_\psi(\phi, \phi^k) \tag{2.1.6}$$

where $D_\psi(\phi, \phi^k)$ is defined by (2.1.5). When $\varphi(t) = -\log(t) + t - 1$, it is easy to see that we get recurrence (2.1.4). Take for example the case of the approximation (1.3.5). Since $\varphi'(t) = \frac{-1}{t} + 1$, we have $\int \varphi'\left(\frac{p_\phi}{p_\alpha}\right) p_\phi dx = 0$. Hence,

$$\hat{D}_\varphi(p_\phi, p_{\phi_T}) = \sup_\alpha \frac{1}{n}\sum_{i=1}^n \log(p_\alpha(y_i)) - \frac{1}{n}\sum_{i=1}^n \log(p_\phi(y_i)).$$

Using the fact that the first term in $\hat{D}_\varphi(p_\phi, p_{\phi_T})$ does not depend on ϕ, so it does not count in the $\arg\inf$ defining ϕ^{k+1}, we may rewrite (2.1.6) as:

$$\begin{aligned}
\phi^{k+1} &= \arg\inf_\phi \left\{ \sup_\alpha \frac{1}{n}\sum_{i=1}^n \log(p_\alpha(y_i)) - \frac{1}{n}\sum_{i=1}^n \log(p_\phi(y_i)) + \frac{1}{n} D_\psi(\phi, \phi^k) \right\} \\
&= \arg\inf_\phi \left\{ -\frac{1}{n}\sum_{i=1}^n \log(p_\phi(y_i)) + \frac{1}{n} D_\psi(\phi, \phi^k) \right\} \\
&= \arg\sup_\phi \left\{ \frac{1}{n}\sum_{i=1}^n \log(p_\phi(y_i)) - \frac{1}{n} D_\psi(\phi, \phi^k) \right\} \\
&= \arg\sup_\phi J(\phi) - D_\psi(\phi, \phi^k).
\end{aligned}$$

For notational simplicity, from now on, we redefine D_ψ with a normalization by n, i.e.

$$D_\psi(\phi, \phi^k) = \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{X}} \psi\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k) dx. \tag{2.1.7}$$

Hence, our set of algorithms is redefined by:

$$\phi^{k+1} = \arg\inf_\phi \hat{D}_\varphi(p_\phi, p_{\phi_T}) + D_\psi(\phi, \phi^k). \tag{2.1.8}$$

We will see later that this iteration forces the estimated divergence to decrease and that under suitable conditions, it converges to a (local) minimum of $\hat{D}_\varphi(p_\phi, p_{\phi_T})$. It results that, algorithm (2.1.8) is a way to calculate the minimum $\varphi-$divergence estimator defined by (1.2.1), (1.2.2), (1.3.6) or (1.5.4).
Before proceeding to study the convergence properties of such algorithm, we will propose another algorithm for the case of mixture models. In the EM algorithm, the estimation of the parameters of a mixture model is done mainly by two steps, see paragraph 2.5.2. The first step estimates the proportions of the classes whereas the second step estimates the parameters defining the classes. Our idea is based on a directional optimization of the objective function in (2.1.8). Convergence properties of the two-step algorithm will also be studied, but the proofs are more technical.

2.2 Two-step Algorithm for mixtures

Let p_ϕ be a mixture model with s components:

$$p_\phi(y) = \sum_{i=1}^s \lambda_i f_i(y|\theta_i). \tag{2.2.1}$$

Here, $\phi = (\lambda, \theta)$ with $\lambda = (\lambda_1, \cdots, \lambda_s) \in [0,1]^s$ such that $\sum_j \lambda_j = 1$, and $\theta = (\theta_1, \cdots, \theta_s) \in \Theta \subset \mathbb{R}^{d-s}$ such that $\Phi \subset [0,1]^s \times \Theta$. In the EM algorithm, the corresponding optimization

to (2.1.8) can be solved by calculating an estimate of the λ's as the proportions of classes, and then proceed to optimize on the θ's (see for example [Titterington et al., 1985]). This simplifies the optimization in terms of complexity (optimization in lower spaces) and clarity (separate proportions from classes parameters). We want to build an algorithm with the same property and divide the optimization problem into two parts. One which estimates the proportions λ and another which estimates the parameters defining the form of each component θ. We propose the following algorithm:

$$\lambda^{k+1} = \underset{\lambda\in[0,1]^s, s.t.(\lambda,\theta^k)\in\Phi}{\arg\inf} \hat{D}_\varphi(p_{\lambda,\theta^k}, p_{\phi_T}) + D_\psi((\lambda,\theta^k),\phi^k); \tag{2.2.2}$$

$$\theta^{k+1} = \underset{\theta\in\Theta, s.t.(\lambda^{k+1},\theta)\in\Phi}{\arg\inf} \hat{D}_\varphi(p_{\lambda^{k+1},\theta}, p_{\phi_T}) + D_\psi((\lambda^{k+1},\theta),\phi^k). \tag{2.2.3}$$

This algorithm corresponds to a directional optimization for recurrence (2.1.8) by considering simply the unit vectors as directions. We can therefore prove analogously that the estimated divergence between the model and the true density decreases as we proceed with the recurrence.

We end the first part of this chapter by three remarks:

- Function ψ defining the distance-like proximal term D_ψ needs not to be convex as in Tseng [2004]. As we will see in the convergence proofs, the only properties needed are: ψ is a non negative function defined on $\mathbb{R}_+$ verifying $\psi(t) = 0$ iff $t = 1$, and $\psi'(t) = 0$ iff $t = 1$.
- The simplified version is not restricted to mixture models. Indeed, any parametric model, whose vector of parameters can be separated into two independent parts, can be estimated using the simplified version.
- As we will see in the proofs, results on the simplified version (2.2.2, 2.2.3) can be extended to a further simplified one. In other words, one may even consider an algorithm which attack a lower level of optimization. We may optimize on each class of the mixture model instead of the whole set of parameters. Since the analytic separation is not evident, one should expect some loss of quality as a cost of a less optimization time.

The remaining of the chapter is devoted entirely to the study of the convergence of the sequences generated by either of the two sets of algorithms (2.1.8) and (2.2.2, 2.2.3) presented above. A key feature which will be needed in the proofs is the regularity of the objective function $\hat{D}_\varphi(p_\phi, p_{\phi_T})$. Regularity of all divergence estimators mentioned at the begining of this chapter can be checked using Lebesgue theorems except for the dual formula (1.3.5). Indeed, continuity and differentiability are not simple since the dual formula is defined through a supremum. The following section is devoted to the study of the regularity of a function written as the supremum of a bivariable function.

2.3 Analytical properties of the dual formula of $\varphi-$divergences

The dual formula defining the estimator of the divergence between the true density and the model defined by (1.3.5) seems quite complicated. This is basically because of a functional integral and a supremum over it. Continuity and differentiation of the integral is resolved by Lebesgue theorems. We only need that the integrand as well as its partial

derivatives to be uniformly bounded with respect to the parameter. However, continuity or differentiability of the supremum is more subtle. Indeed, even if the optimized function is $\mathcal{C}^\infty$, it does not imply the continuity of its supremum. Take for example function $f(x,u) = -e^{xu}$. We have:

$$\sup_x f(x,u) = \begin{cases} -1 & \text{if} \quad u = 0; \\ 0 & \text{if} \quad u \neq 0. \end{cases}$$

On the basis of the theory presented in [Rockafellar and Wets, 1998] about parametric optimization, we present two ways for studying continuity and differentiability of $\hat{D}_\varphi(p_\phi, p_{\phi_T})$ defined through (1.3.5). The first one is the most important because it is easier and demands less mathematical notations. In the first approach, we provide sufficient conditions in order to prove continuity and differentiability almost everywhere of the dual estimator of the divergence. This approach will be used in the study of the convergence of our proximal-point algorithm, see Section 2.6. The second approach is presented for the sake of completness of the study. We give sufficient conditions which permit to prove the differentiability *everywhere*.
We recall first the definition of a subgradient of a real valued function f.

Definition 2.3.1 (Definition 8.3 in Rockafellar and Wets [1998])**.** *Consider a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ and a point ϕ^* with $f(\phi^*)$ finite. For a vector v in $\mathbb{R}^d$, one says that:*

(a) v is a regular subgradient of f at ϕ^*, written $v \in \hat{\partial} f(\phi^*)$, if:

$$f(\alpha) \geq f(\phi^*) + < v, \alpha - \phi^* > + o\left(|\alpha - \phi^*|\right);$$

(b) v is a (general) subgradient of f at ϕ^*, written $v \in \partial f(\phi^*)$, if there are sequences $\alpha^n \to \phi^*$ with $f(\alpha^n) \to f(\phi^*)$, and $v^n \in \hat{\partial} f(\alpha^n)$ with $v^n \to v$.

2.3.1 A result of differentiability almost everywhere : Lower-$\mathcal{C}^1$ functions

Definition 2.3.2 ([Rockafellar and Wets, 1998] Chap 10.)**.** *A function $D : \Phi \to \mathbb{R}$, where Φ is an open set in $\mathbb{R}^d$, is said to be lower-$\mathcal{C}^1$ on Φ, if on some neighborhood V of each ϕ there is a representation*

$$D(\phi) = \sup_{\alpha \in T} f(\alpha, \phi)$$

in which the functions $\alpha \mapsto f(\alpha, \phi)$ are of class $\mathcal{C}^1$ on V and the set T is a compact set such that $f(\alpha, \phi)$ and $\nabla_\phi f(\alpha, \phi)$ depend continuously not just on $\phi \in \Phi$ but jointly on $(\alpha, \phi) \in T \times V$.

In our case, the supremum form is globally defined. Moreover, $T = \Phi$. In case Φ is bounded, it suffices then to take $T = cl(\Phi)$ the closure of Φ since $\alpha \mapsto f(\alpha, \phi)$ is continuous. The condition on T to be compact is essential here, and can not be compromised, so that it is necessary to reduce in a way or in another the optimization on α into a compact or at least a bounded set. For example, one may prove that the values of $\alpha \mapsto f(\alpha, \phi)$ near infinity are lower than some value inside Φ independently of ϕ.

Theorem 2.3.1 (Theorem 10.31 in [Rockafellar and Wets, 1998]). *Any lower-$\mathcal{C}^1$ function D on an open set $\Phi \subset \mathbb{R}^d$ is both (strictly[2]) continuous and continuously differentiable where it is differentiable. Moreover, if Δ consists of the points where D is differentiable, then $\Phi \setminus \Delta$ is negligible[3].*

The stated result can be ensured by simple hypotheses on the model p_ϕ and the function φ. Unfortunately, since the estimated divergence $\hat{D}_\varphi(p_\phi, p_{\phi_T})$ will not be everywhere differentiable, we can no longer talk about the stationarity of $\hat{D}_\varphi(p_\phi, p_{\phi_T})$ at a limit point of the sequence ϕ^k generated for example by (2.1.8). We therefore, use the notion of subgradients. Indeed, when a function g is not differentiable, a necessary condition for x_0 to be a local minimum of g is that $0 \in \partial g(x_0)$ and it becomes sufficient whenever g is proper convex[4]. Moreover, as g becomes differentiable at x_0, then $\nabla g(x_0) \in \partial g(x_0)$ with equality if and only if g is $\mathcal{C}^1$. In other words, proving that $0 \in \partial \hat{D}_\varphi(p_{\hat{\phi}}, p_{\phi_T})$ means that $\hat{\phi}$ is a sort of a *generalized stationary point* of $\phi \mapsto \hat{D}_\varphi(p_\phi, p_{\phi_T})$.
We will be studying later on in paragraphs (2.6.3) and (2.7.1) examples where we verify with more details the previous conditions and see the resulting consequences on the sequence $(\phi^k)_k$.

2.3.2 A result of everywhere differentiability: Level-bounded functions

Definition 2.3.3 ([Rockafellar and Wets, 1998] Chap 1.). *A function $f : \mathbb{R}^d \times \mathbb{R}^d \to \bar{\mathbb{R}}$ with values $f(\alpha, \phi)$ is (upper) level-bounded in α locally uniformly in ϕ if for each ϕ_0 and $a \in \mathbb{R}$ there is a neighborhood V for ϕ_0 such that the set $\{(\alpha, \phi) | \phi \in V, f(\alpha, \phi) \geq a\}$ is bounded in $\mathbb{R}^d \times \mathbb{R}^d$ for every $a \in \mathbb{R}$.*

For a fixed ϕ, the level-boundedness property corresponds to having $f(\alpha, \phi) \to -\infty$ as $\|\alpha\| \to \infty$. In order to state the main result for this case, let ϕ_0 be a point at which we need to study continuity and differentiability of $\phi \mapsto \sup_\alpha f(\alpha, \phi)$. A first result gives sufficient conditions under which the supremum function is continuous. We state it as follows:

Theorem 2.3.2 ([Rockafellar and Wets, 1998] Theorem 1.17). *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ be an upper semicontinuous function. Suppose that $f(\alpha, \phi)$ is level-bounded in α locally uniformly in ϕ. For function $\phi \mapsto \sup_\alpha f(\alpha, \phi)$ to be continuous at ϕ_0, a sufficient condition is the existence of $\alpha_0 \in \arg\max_\alpha f(\alpha, \phi_0)$ such that $\phi \mapsto f(\alpha_0, \phi)$ is continuous at ϕ_0.*

Since in general, we do not know exactly where the supremum will be, one proves the continuity of $\phi \mapsto f(\alpha, \phi)$ for every α.
A Further result about continuity and differentiability of the supremum function can also be stated. Define, at first, the sets $Y(\phi_0)$ and $Y_\infty(\phi_0)$ as follows:

$$Y(\phi_0) = \bigcup_{\alpha \in \arg\sup_\beta f(\beta, \phi_0)} M(\alpha, \phi_0), \quad \text{for } M(\alpha, \phi_0) = \{a | (0, a) \in \partial f(\alpha, \phi_0)\}$$

$$Y_\infty(\phi_0) = \bigcup_{\alpha \in \arg\sup_\beta f(\beta, \phi_0)} M_\infty(\alpha, \phi_0), \quad \text{for } M_\infty(\alpha, \phi_0) = \{a | (0, a) \in \partial^\infty f(\alpha, \phi_0)\}$$

[2] A strictly continuous function f is a local Lipschitz continuous function, i.e. for each $x_0 \in \text{int}\Phi$, the following limit exists and is finite

$$\limsup_{x, x' \to x_0} \frac{|f(x') - f(x)|}{x' - x}$$

[3] A set is called negligible if for every $\varepsilon > 0$, there is a family of boxes $\{B_k\}_k$ with $d-$dimensional volumes ε_k such that $A \subset \cup_k B_k$ and $\sum_k \varepsilon_k < \varepsilon$.
[4] See [Rockafellar and Wets, 1998] theorem 10.1.

where $\partial^\infty f$ is the horizon subgradient, see Definition 8.3 (c) in Rockafellar and Wets [1998]. We avoided to mention the definition here in order to keep the text clearer. Furthermore, in the whole chapter, the horizon subgradient will always be equal to the set $\{0\}$.

Theorem 2.3.3 (Corollary 10.14 in [Rockafellar and Wets, 1998]). *For a proper upper semicontinuous function $f : \mathbb{R}^d \times \mathbb{R}^d \to \bar{\mathbb{R}}$ such that $f(\alpha, \phi)$ is level-bounded in α locally uniformly in ϕ, and for $\phi_0 \in$ dom $\sup_\alpha f(\alpha, \phi)$:*

(a) If $Y_\infty(\phi_0) = \{0\}$, then $\phi \mapsto \sup_\alpha f(\alpha, \phi)$ is strictly continuous at ϕ_0;

(b) if $Y(\phi_0) = \{a\}$ too, then[5] $\phi \mapsto \sup_\alpha f(\alpha, \phi)$ is $\mathcal{C}^1$ at ϕ_0 with $\nabla \sup_\alpha f(\alpha, \phi) = a$.

In our examples, f will be a continuous function and even $\mathcal{C}^1(\Phi \times \Phi)$. This implies that $\partial^\infty f(\alpha, \phi) = \{0\}$ and $\partial f(\alpha, \phi) = \{\nabla f(\alpha, \phi)\}$, see Exercise 8.8 in Rockafellar and Wets [1998]. Hence, $Y_\infty(\phi_0) = \{0\}$ whatever ϕ_0 in Φ. Moreover $M(\alpha, \phi_0) = \{\nabla_\phi f(\alpha, \phi_0)\}$ so that $Y(\phi_0) = \bigcup\{\nabla_\phi f(\alpha, \phi_0)\}$ and the union is on the set of suprema of $\alpha \mapsto f(\alpha, \phi_0)$. If $f(\alpha, \phi)$ is level-bounded in α locally uniformly in ϕ, then the supremum function becomes strictly continuous. Moreover, if the function f has the same gradient with respect to ϕ for all the suprema of $\alpha \mapsto f(\alpha, \phi)$, then $\sup_\alpha f(\alpha, \phi)$ becomes continuously differentiable. This is for example the case when function $\alpha \mapsto f(\alpha, \phi)$ has a unique global supremum for a fixed ϕ, which is for example the case of a strictly concave function (with respect to α for a fixed ϕ).

Example 2.3.1. Let $(p_\phi)_\phi$ be an exponential model defined by:

$$p_\phi(x) = \exp\left[T(x).\phi - C(\phi)\right].$$

Let $\varphi(t) = t\log(t) - t + 1$. The dual representation of the divergence (formula (1.3.5)) is then given by:

$$\hat{D}_\varphi(p_\phi, p_{\phi*}) = \sup_\alpha \left\{ \mathbb{E}_{p_\phi}[T(X)].(\phi - \alpha) + C(\alpha) - C(\phi) - \frac{1}{n}\sum_{i=1}^n e^{T(y_i).(\phi-\alpha)+C(\alpha)-C(\phi)} \right\} + 1.$$

In order to prove that the optimized function is level-bounded in α locally uniformly in ϕ, we take a bounded open neighborhood around ϕ, and we prove that the optimized function tends to $-\infty$ as $\|\alpha\|$ tends to infinity. For example, for the Gaussian case with the mean μ as the parameter of interest, we have:

$$\hat{D}_\varphi(p_\mu, p_{\mu*}) = \sup_{\beta\in\mathbb{R}} \frac{1}{2}\mu^2 - \mu\beta + \frac{1}{2}\beta^2 - \frac{1}{n}\sum_{i=1}^n e^{y_i(\mu-\beta)+\frac{1}{2}\beta^2-\frac{1}{2}\mu^2} + 1.$$

It is clear that e^{β^2} is the dominant term at ∞, and by putting μ in a bounded interval, the limit of the optimized function when β tends to infinity is easily calculated and equals $-\infty$.
For the exponential case $p_a(x) = ae^{-ax}$ with $a > 0$ the parameter of interest, we have:

$$\hat{D}_\varphi(p_a, p_{a^*}) = \sup_{b>0} \frac{b}{a} - \log(b) + \log(a) - \frac{1}{n}\sum_{i=1}^n e^{-y_i(a-b)-\log(b)+\log(a)}.$$

[5] In the statement of the corollary in [Rockafellar and Wets, 1998], the supremum function becomes strictly differentiable, but to avoid extra vocabularies, we replaced it with an equivalent property.

Here again, the dominant term is $e^{y_i b}$ at infinity. Since an observation of an exponential law is positive, the limit when b tends to infinity is hence easily calculated and equals $-\infty$. For part (a) of Theorem 2.3.3 to be verified, we still need to prove that $Y_\infty(\phi_0) = \{0\}$. However, this is verified because the optimized function, here, is continuously differentiable, so that it is strictly continuous. This implies that $Y_\infty(\phi_0) = \{0\}$. Hence, $\hat{D}_\varphi(p_\phi, p_{\phi^T})$ is strictly continuous.
To prove that it is also $\mathcal{C}^1$, we need to prove that $Y(\phi_0)$ contains but one element. First of all, since the optimized function is differentiable, $Y(\phi_0) = \bigcup\{\nabla_\phi f(\alpha, \phi_0)\}$. The union is over the set $\{\arg\max_\alpha f(\alpha, \phi_0)\}$. Let's calculate the jacobian matrix with respect to α and see when it might be definite negative, and hence function $\alpha \mapsto f(\alpha, \phi)$ would be strictly concave and would only have one maximum whenever it exists[6]. If for any ϕ, it is definite negative, this should be sufficient to prove the claim.

$$\begin{aligned}
\nabla_\alpha f(\alpha, \phi) &= -\mathbb{E}_{p_\phi}[T(X)] + \nabla C(\alpha) - \frac{1}{n}\sum_{i=1}^{n} (\nabla C(\alpha) - T(y_i)) e^{T(y_i).(\phi-\alpha)+C(\alpha)-C(\phi)} \\
J_f(\alpha, \phi) &= J_C(\alpha) - \frac{1}{n}\sum_{i=1}^{n} \left(J_C(\alpha) + (\nabla C(\alpha) - T(y_i)).(\nabla C(\alpha) - T(y_i))^t\right) e^{T(y_i).(\phi-\alpha)+C(\alpha)-C(\phi)}
\end{aligned}$$

For the Gaussian example, we have:

$$\begin{aligned}
\frac{\partial f}{\partial \beta}(\beta, \mu) &= -\mu + \beta - \frac{1}{n}\sum_{i=1}^{n} (\beta - y_i) e^{y_i(\mu-\beta)+\frac{1}{2}\beta^2-\frac{1}{2}\mu^2} \\
\frac{\partial^2 f}{\partial \beta^2}(\beta, \mu) &= 1 - \frac{1}{n}\sum_{i=1}^{n} \left(1 + (\beta - y_i)^2\right) e^{y_i(\mu-\beta)+\frac{1}{2}\beta^2-\frac{1}{2}\mu^2}
\end{aligned}$$

The gradient has at least one zero since it is continuous and has $+\infty$ limit at $-\infty$ and $-\infty$ limit at $+\infty$. The second derivative with respect to β is unfortunately not necessarily negative so that function $\beta \mapsto f(\beta, \mu)$ is not concave. An analytical study of function f seems very difficult. Let's simulate a 10-sample of the standard Gaussian probability law, and let a mathematical tool such as Mathematica do the painting. Table (2.1) shows the dataset used.

y_i	0.644	-3.144	-1.029	-0.367	0.353	-0.704	1.148	0.674	0.148	-0.721

Table 2.1: A 10-sample Gaussian dataset.

We make a 3D plot for $f(\beta, \mu)$ in two parts. The first part for $\mu > 0$ and the second is for $\mu < 0$ to get a clear view about what happens when μ changes, see figure (2.1). Although the second derivative with respect to β is not necessarily negative, function $\beta \mapsto f(\beta, \mu)$ has only one maximum point. We conclude that function $\sup_\beta f(\beta, \mu)$ is continuously derivable *for the dataset provided in table (2.1).*

[6]Notice that for both the Gaussian and exponential example, the derivative passes by zero as will be explained later on. Therefore, strict concavity would imply the existence of one maximum.

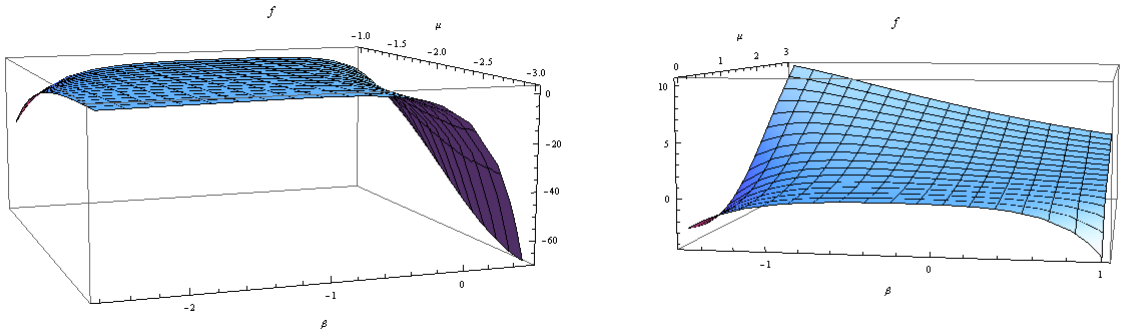

Figure 2.1: A 3D plot of function f in the Gaussian example shows that there is only one maximum for each value of μ.

For the exponential case, we have:

$$\begin{aligned}\frac{\partial f}{\partial b} &= \frac{1}{a} - \frac{1}{b} - \frac{1}{n}\sum_{i=1}^{n}\left(by_i - \frac{1}{b}\right)e^{-y_i(a-b)-\log(b)+\log(a)}\\ \frac{\partial^2 f}{\partial b^2} &= \frac{1}{b^2} - \frac{1}{n}\sum_{i=1}^{n}\left(y_i + \frac{1}{b^2} + \left(by_i - \frac{1}{b}\right)^2\right)e^{-y_i(a-b)-\log(b)+\log(a)}\end{aligned}$$

We similarly have the same previous problem. The second derivative is not necessarily negative, however, when simulating a dataset and plotting function f, we may conclude that it has only one maximum whenever a is fixed. Hence function $\sup_b f(a,b)$ becomes continuously derivable.
These two examples, although very simple, shows the difficulty in proving differentiability. Our proof, earlier, depends heavily on *graphical tools*, which may still appear not totally convincing. Another undesirable aspect is that even if we admit the *graphical* indications about the existence of a unique maximum, the final conclusion stays related to the dataset we are working on. An analytical proof for previous examples remain an open problem for further work.

Remark 2.3.1 (An implicit function point of view)**.** One may try in case $f(\alpha,\phi)$ is concave to calculate the gradient with respect to α. At the supremum, whenever it exists, we have $\nabla_\alpha f(\alpha,\phi) = 0$. The solution in α is *a priori* a function of ϕ, say $\alpha(\phi)$. The implicit function theorem provides a way to prove the existence of such function and gives sufficient conditions for continuity and differentiability. Notice that a *global* version of the implicit function theorem is needed here in order to define $\alpha(\phi)$ on the whole Φ and not locally. As soon as we have such a function, we may write $f(\alpha(\phi),\phi) = \hat{D}_\varphi(p_\phi, p_{\phi_T})$, and the divergence becomes differentiable using a simple chain rule. The problem with this solution is that the conditions to ensure a global function $\alpha(\phi)$ are not simple, see for example [Cristea, 2007] and the references therein.

2.4 Convergence properties

We adapt the ideas given in [Tseng, 2004] to develop a suitable proof for our proximal algorithm. We present some propositions which show how according to some possible situations one may prove convergence of the algorithms defined by recurrences (2.1.8) and (2.2.2, 2.2.3). Let $\phi^0 = (\lambda^0, \theta^0)$ be a given initialization for the parameters, and define the

following set

$$\Phi^0 = \{\phi \in \Phi : \hat{D}_\varphi(p_\phi, p_{\phi_T}) \leq \hat{D}_\varphi(\phi^0, \phi_T)\} \tag{2.4.1}$$

where $\hat{D}_\varphi(\phi, \phi_T)$ is any estimator of the $\varphi-$divergence among (1.2.1), (1.2.2), (1.3.5) or (1.5.3). We suppose that Φ^0 is a subset of $int(\Phi)$. The idea of defining such a set in this context is inherited from the paper of [Wu, 1983] which provided the first *correct proof* of convergence for the EM algorithm. Before going any further, we recall the following definition of a (generalized) stationary point.

Definition 2.4.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a real valued function. If f is differentiable at a point ϕ^* such that $\nabla f(\phi^*) = 0$, we then say that ϕ^* is a stationary point of f. If f is not differentiable at ϕ^* but the subgradient of f at ϕ^*, say $\partial f(\phi^*)$, exists such that $0 \in \partial f(\phi^*)$, then ϕ^* is called a generalized stationary point of f.*

Using continuity and differentiability assumptions on both $\hat{D}_\varphi$ and D_ψ, we will prove the following results:

- For both algorithms (2.1.8) and (2.2.2, 2.2.3), if Φ^0 is closed and $\{\phi^{k+1} - \phi^k\} \to 0$, then any limit point of $(\phi^k)_k$ is a stationary point of the objective function $\hat{D}_\varphi(\phi, \phi_T)$;
- For algorithm (2.1.8), if we only have Φ^0 is compact, then any limit point is a stationary point of the objective function;
- For algorithm (2.2.2, 2.2.3), if Φ^0 is compact and $\|\lambda^{k+1} - \lambda^k\| \to 0$, then any limit point is a stationary point of the objective function;
- For both algorithms (2.1.8) and (2.2.2, 2.2.3), if Φ^0 is compact and $D_\psi(\phi, \phi') > 0$ iff $\phi \neq \phi'$, then $\{\phi^{k+1} - \phi^k\} \to 0$ and any limit point is a stationary point of the objective function.
- In case the objective function $\phi \mapsto \hat{D}_\varphi(p_\phi|p_{\phi_T})$ is not continuously differentiable, we prove previous points for algorithm (2.1.8) with generalized stationary point instead of stationary point.

We will be using the following assumptions which will be checked in several examples later on.

A0. Functions $\phi \mapsto \hat{D}_\varphi(p_\phi|p_{\phi_T}), D_\psi$ are lower semicontinuous;

A1. Functions $\phi \mapsto \hat{D}_\varphi(p_\phi|p_{\phi_T}), D_\psi$ and $\nabla_1 D_\psi$ are defined and continuous on, respectively, $\Phi, \Phi \times \Phi$ and $\Phi \times \Phi$;

AC. $\nabla \hat{D}_\varphi(p_\phi|p_{\phi_T})$ is defined and continuous on Φ;

A2. Φ^0 is a compact subset of int(Φ);

A3. $D_\psi(\phi, \bar{\phi}) > 0$ for all $\bar{\phi} \neq \phi \in \Phi$.

Recall also the assumptions on functions h_i defining D_ψ. We suppose that $h_i(x|\phi) > 0, dx - a.e.$, and $\psi(t) = 0$ iff $t = 1$. Besides $\psi'(t) = 0$ iff $t = 1$.
Concerning assumptions A1 and AC, we have previously discussed the analytical properties of $\hat{D}_\varphi(p_\phi|p_{\phi_T})$ after Section 2.2 and in Section 2.3. In what concerns D_ψ, continuity and differentiability can be obtained merely by fulfilling Lebesgue theorems conditions. For example, if $h_i(x, \phi)$ is continuous and bounded uniformly away from 0 independently of ϕ,

then continuity is guaranteed as soon as ψ is continuous. If we also suppose that $\nabla_\phi h_i(x,\phi)$ exists, is continuous and is uniformly bounded independently of ϕ, then as soon as ψ is continuously differentiable, D_ψ becomes continuously differentiable. For assumption A2, there is no universal method. Still, in all the examples that will be discussed later, we use the fact that the inverse image of a closed set by a continuous function is closed. Boundedness is usually ensured using a *suitable* choice of ϕ^0. Finally, assumption A3 is checked using Lemma 2 proved in [Tseng, 2004] which we restate here.

Lemma 2.4.1 (Tseng [2004] Lemma 2)**.** *Suppose ψ to be a continuous non negative function such that $\psi(t) = 0$ iff $t = 1$. For any ϕ and ϕ' in Φ, if $h_i(x|\phi) \neq h_i(x|\phi')$ for some $i \in \{1, \cdots, n\}$ and some $x \in int(X)$ at which both $h_i(.|\phi)$ and $h_i(.|\phi')$ are continuous, then $D_\psi(\phi, \phi') > 0$.*

In section (2.6), we present three different examples; a two-component Gaussian mixture, a two-component Weibull mixture and a Cauchy model. We will see that the Cauchy example verifies assumption A3. However, the Gaussian mixture does not seem to verify it. Indeed, the same fact stays true for any mixture of the exponential family.
We start by providing some general facts about the sequence $(\phi^k)_k$ and its existence. We also prove convergence of the sequence $(\hat{D}_\varphi(p_{\phi^k}|p_{\phi_T}))_k$.

Remark 2.4.1. All results concerning algorithm (2.1.8) are proved even when assumption AC is not fulfilled. We give proofs using the subgradient of the estimated $\varphi-$divergence. In the case of the two-step algorithm (2.2.2, 2.2.3), it was not possible and thus remains an open problem. The difficulty resides in manipulating the *partial* subgradients with respect to λ and θ which cannot be handled in a similar way to the partial derivatives.

Remark 2.4.2. Convergence properties are proved without using the special form of the estimated $\varphi-$divergence. Thus, our theoretical approach applies to any optimization problem whose objective is to minimize a function $\phi \mapsto D(\phi)$. For example, our approach can be applied on density power divergences (Basu et al. [1998]), Bregman divergences, S-divergences (Ghosh et al. [2013]), etc.

Proposition 2.4.1. *We assume that recurrences (2.1.8) and (2.2.2, 2.2.3) are well defined in Φ. For both algorithms, the sequence $(\phi^k)_k$ verifies the following properties:*

(a) $\hat{D}_\varphi(p_{\phi^{k+1}}|p_{\phi_T}) \leq \hat{D}_\varphi(p_{\phi^k}|p_{\phi_T})$;

(b) $\forall k, \phi^k \in \Phi^0$;

(c) Suppose that assumptions A0 and A2 are fulfilled, then the sequence $(\phi^k)_k$ is defined and bounded. Moreover, the sequence $\left(\hat{D}_\varphi(p_{\phi^k}|p_{\phi_T})\right)_k$ converges.

The proof of this proposition is differed to Appendix 2.9.1. The interest of Proposition 2.4.1 is that the objective function is ensured, under mild assumptions, to decrease alongside the sequence $(\phi^k)_k$. This permits to build a stop criterion for the algorithm since in general there is no guarantee that the whole sequence $(\phi^k)_k$ converges. It may also continue to fluctuate in a neighborhood of an optimum. The following result provides a first characterization about the properties of the limit of the sequence $(\phi^k)_k$ as (generalized) a stationary point of the estimated $\varphi-$divergence. The proof is differed to Appendix 2.9.2.

Proposition 2.4.2. *Suppose that A1 is verified, and assume that Φ^0 is closed and $\{\phi^{k+1} - \phi^k\} \to 0$.*

(a) *For both algorithms (2.1.8) and (2.2.2,2.2.3), if AC is verified, then the limit of every convergent subsequence is a stationary point of* $\hat{D}_\varphi(.|p_{\phi_T})$*;*

(b) *For the first algorithm (2.1.8), if* $\hat{D}_\varphi(.|p_{\phi_T})$ *is not differentiable, then the limit of every convergent subsequence is a "generalized" stationary point of* $\hat{D}_\varphi(.|p_{\phi_T})$*, i.e. zero belongs to the subgradient of* $\hat{D}_\varphi(.|p_{\phi_T})$ *calculated at the limit point;*

Assumption $\{\phi^{k+1} - \phi^k\} \to 0$ used in Proposition 2.4.2 is not easy to be checked unless one has a close formula of ϕ^k. This is the case of the EM algorithm applied on a Gaussian mixture, see Tseng [2004] Section 5. In general, we prove that $\{\phi^{k+1} - \phi^k\} \to 0$ by imposing an identifiability assumption over the proximal term, see Chrétien and Hero [1998] Lemma 5 or Tseng [2004] Lemma 1. The following proposition is a mere adaptation of such results to the context of $\varphi-$divergences and the two-step algorithm. The proof is differed to Appendix 2.9.3. We will present later a result which does not need such assumption.

Proposition 2.4.3. *For both algorithms defined by (2.1.8) and (2.2.2,2.2.3), assume A1, A2 and A3 verified, then* $\{\phi^{k+1} - \phi^k\} \to 0$*. Thus, by proposition 2 (according to whether AC is verified or not) implies that any limit point of the sequence* ϕ^k *is a (generalized)*[7] *stationary point of* $\hat{D}_\varphi(.|p_{\phi_T})$*.*

We can go further in exploring the properties of the sequence $(\phi^k)_k$, but we need to impose more assumptions. The following corollary provides a convergence result of the *whole* sequence and not only some subsequence. The convergence is also towards a local minimum as soon as the estimated divergence is locally strictly convex. The proof of the following result is differed to Appendix 2.9.4.

Corollary 2.4.1. *Under the assumptions of Proposition 3, the set of accumulation points of* $(\phi^k)_k$ *is a connected compact set. Moreover, if* $\hat{D}(p_\phi, p_{\phi_T})$ *is strictly convex in a neighborhood of a limit point*[8] *of the sequence* $(\phi^k)_k$*, then the whole sequence* $(\phi^k)_k$ *converges to a local minimum of* $\hat{D}(p_\phi, p_{\phi_T})$*.*

Proposition 2.4.3 although provides a general solution to prove that $\{\phi^{k+1} - \phi^k\} \to 0$, the identifiability assumption over the proximal term is hard to be fulfilled. It is not verified in the most simple mixtures such as a two component Gaussian mixture, see Section (2.6.1). This was the reason behind our next result. We prove that we do not need to assume identifiability of the proximal term in order to prove that any convergent subsequence of $(\phi^k)_k$ is a (generalized) stationary point of the estimated $\varphi-$divergence.
A similar idea was employed in [Chrétien and Hero, 2008] who studied a proximal algorithm for the log-likelihood function with a relaxation parameter[9]. Their work however requires that the log-likelihood has $-\infty$ limit as $\|\phi\| \to \infty$ which is not verified on several mixture models (e.g. the Gaussian mixture model). Our result treat the problem from another approach based on the introduction of the set Φ^0.

Proposition 2.4.4. *Assume A1, AC and A2 verified. For the algorithm defined by (2.1.8), any convergent subsequence converges to a stationary point of the objective function* $\phi \to \hat{D}(p_\phi, p_{\phi_T})$*. If AC is dropped, then 0 belongs to the subgradient of* $\phi \mapsto \hat{D}(p_\phi, p_{\phi_T})$ *at the limit point.*

[7] The case where AC is not verified is only proved for the first algorithm (2.1.8)

[8] This assumption can be replaced by local strict convexity since *a priori*, we have no idea where might find a limit point of the sequence $(\phi^k)_k$.

[9] A sequence of decreasing positive numbers multiplied by the proximal term.

The proof is differed to Appendix 2.9.5 We could not perform the same idea on the two-step algorithm (2.2.2,2.2.3) without assuming that the difference between two consecutive terms of either the sequence of weights $(\lambda^k)_k$ or the sequence of form parameters $(\theta^k)_k$ converges to zero. The proof is differed to Appendix 2.9.6.

Proposition 2.4.5. *Assume A1 and A2 verified. For the algorithm defined by (2.2.2,2.2.3). If $\|\theta^{k+1} - \theta^k\| \to 0$, then any convergent subsequence $(\phi^{N(k)})_k$ converges to a stationary point of the objective function $\phi \to \hat{D}(p_\phi, p_{\phi^T})$.*

Remark 2.4.3. The previous proposition demands a condition on the distance between two consecutive members of the sequence $(\theta^k)_k$ which is *a priori* weaker than the same condition on the whole sequence $\phi^k = (\lambda^k, \theta^k)$. Still, as the regularization term D_ψ does not verify the identifiability condition A3, it stays an open problem for a further work. It is interesting to notice that condition $\|\theta^{k+1} - \theta^k\| \to 0$ can be replaced by $\|\lambda^{k+1} - \lambda^k\| \to 0$, but we then need to change the order of steps (2.2.2) and (2.2.3). A condition over the proportions seems to be *simpler*.

Remark 2.4.4. We can define an algorithm which converges to a global infimum of the estimated $\varphi-$divergence (see paragraph 2.5.1). The idea is very simple. We need to multiply the proximal term by a sequence of positive numbers which decreases to zero, say $1/k$. The justification of such variant can be deduced from Theorem 3.2.4 in [Chrétien and Hero, 2008]. The problem with this approach is that it depends heavily on the fact that the supremum on each step of the algorithm is calculated exactly. This does not happen in general unless function $\hat{D}_\varphi(p_\phi, p_{\phi^T}) + \beta_k D_\psi(\phi, \phi^k)$ is strictly convex. Although in our approach, we use similar assumption to prove the consecutive decreasing of $\hat{D}_\varphi(p_\phi, p_{\phi^T})$, we can replace the infimum calculus in (2.1.8) by two things. We require at each step that we find a local infimum of $\hat{D}_\varphi(p_\phi, p_{\phi^T}) + D_\psi(\phi, \phi^k)$ whose evaluation with $\phi \mapsto \hat{D}_\varphi(p_\phi, p_{\phi^T})$ is less than the previous term of the sequence ϕ^k. If we can no longer find any local maxima verifying the claim, the procedure stops with $\phi^{k+1} = \phi^k$. This ensures the availability of all proofs presented in this paper with no further changes.

2.5 Case Studies and Variants of the algorithm

2.5.1 An algorithm with theoretically global infimum attainment

We present a variant of algorithm (2.1.8) which ensures *theoretically* convergence to a global infimum of the objective function $\hat{D}_\varphi(p_\phi, p_{\phi^T})$ as long as there exists a convergent subsequence. The idea is the same as Theorem 3.2.4 in [Chrétien and Hero, 2008]. Define ϕ^{k+1} by:

$$\phi^{k+1} = \arg\inf_{\phi} \hat{D}_\varphi(p_\phi, p_{\phi^T}) + \beta_k D_\psi(\phi, \phi^k).$$

The proof of convergence is very simple and does not depend on the differentiability of any of the two functions $\hat{D}_\varphi$ or D_ψ. We only assume A1 and A2 to be verified. Let $(\phi^{N(k)})_k$ be a convergent subsequence. Let ϕ^∞ be its limit. This is guaranteed by the compactness of Φ^0 and the fact that the whole sequence $(\phi^k)_k$ resides in Φ^0 (see Proposition 2.4.1-b). Suppose also that the sequence $(\beta_k)_k$ converges to 0 as k goes to infinity.
Let ϕ by a vector of Φ which has a value of $\hat{D}_\varphi(p_\phi, p_{\phi^T})$ strictly inferior to the value of the same function at ϕ^∞, i.e.

$$\hat{D}_\varphi(p_\phi, p_{\phi^T}) < \hat{D}_\varphi(p_{\phi^\infty}, p_{\phi^T}). \tag{2.5.1}$$

By definition of $\phi^{N(k)}$, we have:

$$\hat{D}_\varphi(p_{\phi^{N(k)}}, p_{\phi^T}) + \beta_{N(k)-1} D_\psi(\phi^{N(k)}, \phi^{N(k)-1}) \leq \hat{D}_\varphi(p_\phi, p_{\phi^T}) + \beta_{N(k)-1} D_\psi(\phi, \phi^{N(k)}),$$

which holds for every ϕ in the whole set Φ. Using the non negativity of the term $\beta_{N(k)-1} D_\psi(\phi^{N(k)}, \phi^{N(k)-1})$, one can write:

$$\hat{D}_\varphi(p_{\phi^{N(k)}}, p_{\phi^T}) \leq \hat{D}_\varphi(p_\phi, p_{\phi^T}) + \beta_{N(k)} D_\psi(\phi, \phi^{N(k)}). \tag{2.5.2}$$

As we pass to the limit on k, recall firstly that $(\beta_k)_k$ converges to 0, so that any subsequence $(\beta_{N(k)})_k$ also converges to 0. Secondly, the continuity assumption on D_ψ implies that, since $\phi^{N(k)} \to \phi^\infty$, $D_\psi(\phi, \phi^{N(k)})$ converges to $D_\psi(\phi, \phi^\infty)$. By compactness of Φ^0 and Proposition 2.4.1-b, we have $\phi^\infty \in \Phi^0$. The continuity again of D_ψ will imply that the quantity $D_\psi(\phi, \phi^\infty)$ is finite. Finally, inequality (2.5.2) now implies that:

$$\hat{D}_\varphi(p_{\phi^\infty}, p_{\phi^T}) \leq \hat{D}_\varphi(p_\phi, p_{\phi^T})$$

which contradicts with the choice of ϕ verifying (2.5.1). Hence, ϕ^∞ is a global infimum on Φ.

The problem with this approach is that it depends heavily on the fact that the supremum on each step of the algorithm is calculated exactly. This does not happen in general unless function $\hat{D}_\varphi(p_\phi, p_{\phi^T}) + \beta_k D_\psi(\phi, \phi^k)$ is convex or that we dispose of an algorithm which can solve perfectly non convex optimization problems[10]. Although in our approach, we use similar assumption to prove the consecutive decreasing of $\hat{D}_\varphi(p_\phi, p_{\phi^T})$, we can replace the infimum calculus in (2.1.8) by two things. We demand that at each step that we find a local infimum of $\hat{D}_\varphi(p_\phi, p_{\phi^T}) + D_\psi(\phi, \phi^k)$ whose evaluation with $\phi \mapsto \hat{D}_\varphi(p_\phi, p_{\phi^T})$ is less than the previous term of the sequence ϕ^k. If we can no longer find any local infima verifying the claim, the procedure stops with $\phi^{k+1} = \phi^k$. This ensures the availability of all the proofs presented in this paper with no change.

2.5.2 The EM algorithm in the context of mixture models

In the case of mixture models (2.2.1), the EM recurrence can be rewritten in two parts; a part where the maximization is on the proportions, and a part on the parameters describing the form of classes. We code the s classes directly by their indices $\{1, \cdots, s\}$. Starting from (2.1.1), one may insert directly the constraint on the λ's into the optimized function as follows:

$$\begin{aligned}(\lambda_1^{k+1}, \cdots, \lambda_{s-1}^{k+1}, \theta_1^{k+1}, \cdots, \theta_s^{k+1}) = \underset{\lambda_1 \geq 0, \cdots, \lambda_{s-1} \geq 0, (\theta_1, \cdots \theta_s) \in \Theta}{\arg\sup} & \sum_{i=1}^{n} \sum_{j=1}^{s-1} \log\left(\lambda_j p(y_i|\theta_j)\right) h_i(j|\phi^k) \\ & + \sum_{i=1}^{n} \log\left(\left(1 - \sum_{j=1}^{s-1} \lambda_j\right) p(y_i|\theta_j)\right) h_i(j|\phi^k)\end{aligned}$$

where

$$h_i(x_j|\phi^k) = \frac{\lambda_j^k f_j(y_i|\theta_j^k)}{\sum_j \lambda_j^k f_j(y_i|\theta_j^k)}.$$

[10] In this case, there is no meaning in applying an iterative proximal algorithm. We would have used the optimization algorithm directly on the objective function $\hat{D}_\varphi(p_\phi, p_{\phi^T})$

Now, the property of the logarithmic function in transforming the product into a sum is the cornerstone in the simplification of the optimization.

$$(\lambda_1^{k+1},\cdots,\lambda_{s-1}^{k+1},\theta_1^{k+1},\cdots,\theta_s^{k+1}) = \underset{\lambda_1\geq 0,\cdots,\lambda_{s-1}\geq 0,(\theta_1,\cdots\theta_s)\in\Theta}{\arg\sup} \sum_{i=1}^{n}\sum_{j=1}^{s-1}\log(\lambda_j)h_i(j|\phi^k)+$$

$$\sum_{i=1}^{n}\log\left(1-\sum_{j=1}^{s-1}\lambda_j\right)h_i(s|\phi^k)+\sum_{i=1}^{n}\sum_{j=1}^{s-1}\log\left(\lambda_j p(y_i|\theta_j)\right)h_i(j|\phi^k)+\sum_{i=1}^{n}\log\left(p(y_i|\theta_j)\right)h_i(s|\phi^k).$$

The optimized function is, thus, written as the sum of two independent functions in the sense that the first one contains only proportions parameters whereas the second contains other parameters. Since the parameters (proportions and the others) are independent from each others[11], one can rewrite the previous optimization problem as the sum of two optimization problems:

$$\begin{aligned}(\lambda_1^{k+1},\cdots,\lambda_{s-1}^{k+1}) &= \underset{\lambda_1\geq 0,\cdots,\lambda_{s-1}\geq 0}{\arg\sup} \sum_{i=1}^{n}\sum_{j=1}^{s-1}\log(\lambda_j)h_i(j|\phi^k)+\sum_{i=1}^{n}\log\left(1-\sum_{j=1}^{s-1}\lambda_j\right)h_i(s|\phi^k);\\ (\theta_1^{k+1},\cdots,\theta_s^{k+1}) &= \underset{(\theta_1,\cdots\theta_s)\in\Theta}{\arg\sup} \sum_{i=1}^{n}\sum_{j=1}^{s-1}\log\left(\lambda_j p(y_i|\theta_j)\right)h_i(j|\phi^k)+\sum_{i=1}^{n}\log\left(p(y_i|\theta_j)\right)h_i(s|\phi^k).\end{aligned}$$

The first step can be explicitly calculated. Solving the gradient equation gives:

$$\begin{aligned}\frac{1}{\lambda_j}\sum_{i=1}^{n}h_i(j|\phi^k)-\frac{1}{\sum_{l=1}^{s-1}\lambda_l}\sum_{i=1}^{n}h_i(s|\phi^k) &= 0 \quad \forall j\in\{1,\cdots,s-1\};\\ \frac{\sum_{i=1}^{n}h_i(s|\phi^k)}{\sum_{i=1}^{n}h_i(j|\phi^k)}\lambda_j &= 1-\sum_{l=1}^{s-1} \quad \forall j\in\{1,\cdots,s-1\} \qquad (2.5.3)\\ \lambda_1+\cdots+\left(1+\frac{\sum_{i=1}^{n}h_i(s|\phi^k)}{\sum_{i=1}^{n}h_i(j|\phi^k)}\right)\lambda_j+\lambda_{s-1} &= 1 \quad \forall j\in\{1,\cdots,s-1\}; \qquad (2.5.4)\end{aligned}$$

Equation (2.5.3) implies that:

$$\forall j\in\{1,\cdots,s-1\},\quad \frac{1}{\sum_{i=1}^{n}h_i(j|\phi^k)}\lambda_j=\frac{1}{\sum_{i=1}^{n}h_i(j|\phi^k)}\lambda_1. \qquad (2.5.5)$$

Rewriting equation (2.5.4) for $j=1$ and using previous identities gives:

$$\begin{aligned}1 &= \left(1+\frac{\sum_{i=1}^{n}h_i(s|\phi^k)}{\sum_{i=1}^{n}h_i(1|\phi^k)}\right)\lambda_1+\sum_{l=2}^{s-1}\frac{\sum_{i=1}^{n}h_i(l|\phi^k)}{\sum_{i=1}^{n}h_i(1|\phi^k)}\lambda_1;\\ 1 &= \lambda_1\left[1+\frac{n-\sum_{i=1}^{n}h_i(1,\phi^k)}{\sum_{i=1}^{n}h_i(1,\phi^k)}\right];\\ \lambda_1 &= \frac{1}{n}\sum_{i=1}^{n}h_i(1,\phi^k).\end{aligned}$$

In the second line, we used the fact that $h_i(s|\phi^k)=1-\sum_{l=1}^{s-1}h_i(l|\phi^k)$. Finally, we use (2.5.5) to deduce that:

$$\lambda_1=\frac{1}{n}\sum_{i=1}^{n}h_i(1,\phi^k)\forall j\in\{1,\cdots,s-1\}.$$

[11] There is no common constraint between them.

Now, the EM recurrence is given by:

$$\begin{aligned}\lambda_j^{k+1} &= \frac{1}{n}\sum_{i=1}^{n} h_i(x_j|\phi^k) \quad j \in \{1,\cdots,s-1\};\\ \theta^{k+1} &= \arg\sup_{\theta} \sum_{i=1}^{n}\sum_{j=1}^{s} \log\left(f_j(y_i|\theta_j)\right) h_i(x_j|\phi^k).\end{aligned}$$

This was the idea behind our algorithm defined by (2.2.2,2.2.3). Furthermore, the second part of the optimization can be simplified more than that. We may write an optimization corresponding to each class since the optimized function is a sum of terms each of which depends only of the parameter vector θ_j defining the corresponding class. The EM algorithm can be rewritten as follows:

$$\begin{aligned}\lambda_j^{k+1} &= \frac{1}{n}\sum_{i=1}^{n} h_i(x_j|\phi^k) j \in \{1,\cdots,s-1\}\\ \theta_j^{k+1} &= \arg\sup_{\theta_j} \sum_{i=1}^{n} \log\left(f_j(y_i|\theta_j)\right) h_i(x_j|\phi^k) \quad j \in \{1,\cdots,s\}\end{aligned}$$

This suggests to go further in algorithm (2.2.2,2.2.3) and use the same idea of directional optimization on the second part (2.2.3). The convergence results can be extended to this variant without any additional assumptions.

2.6 Theoretical study of convergence on some mixtures with application to the EM algorithm

In this section, we present three examples where we check assumptions A0-A3 and AC and study the convergence properties of the sequence ϕ^k. We only consider for the estimated divergence the two dual formula presented in paragraphs 1.3.1 and 1.5.1. Other $\varphi-$divergence-based estimators; Beran's and Basu-Linsday's approaches, can be treated in a similar way to the kernel-based MDφDE.

2.6.1 two-component Gaussian mixture

We suppose that the model $(p_\phi)_{\phi\in\Phi}$ is a mixture of two Gaussian densities, and suppose that we are only interested in estimating the means $\mu = (\mu_1,\mu_2) \in \mathbb{R}^2$ and the proportions $\lambda = (\lambda_1,\lambda_2) \in [\eta, 1-\eta]^2$. The use of η is to avoid cancellation of any of the two components and to keep the hypothesis about the conditional densities h_i true, i.e. $h_i(x|\phi) > 0$ for $x = 1, 2$. We also suppose to simplify the calculus that the components variances are reduced ($\sigma_i = 1$). The model takes the form:

$$p_{\lambda,\mu}(x) = \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_1)^2} + \frac{1-\lambda}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_2)^2}, \tag{2.6.1}$$

where $\Phi = [\eta, 1-\eta]^s \times \mathbb{R}^s$. Here $\phi = (\lambda, \mu_1, \mu_2)$. The distance-like function D_ψ is defined by:

$$D_\psi(\phi,\phi^k) = \sum_{i=1}^{n} \psi\left(\frac{h_i(1|\phi)}{h_i(1|\phi^k)}\right) h_i(1|\phi^k) + \sum_{i=1}^{n} \psi\left(\frac{h_i(2|\phi)}{h_i(2|\phi^k)}\right) h_i(2|\phi^k),$$

where:

$$h_i(1|\phi) = \frac{\lambda e^{-\frac{1}{2}(y_i-\mu_1)^2}}{\lambda e^{-\frac{1}{2}(y_i-\mu_1)^2} + (1-\lambda)e^{-\frac{1}{2}(y_i-\mu_2)^2}}, \quad h_i(2|\phi) = 1 - h_i(1|\phi).$$

It is clear that functions h_i are of class $\mathcal{C}^1$ on $(\text{int}(\Phi))$, and as a consequence, D_ψ is also of class $\mathcal{C}^1$ on $(\text{int}(\Phi))$.

If we are using the dual estimator of the $\varphi-$divergence given by (1.3.5), then assumption A0 can be verified using the maximum theorem of Berge [1963]. There is still a great difficulty in studying the properties (closedness or compactness) of the set Φ^0. Moreover, all convergence properties of the sequence ϕ^k require the continuity of the estimated $\varphi-$divergence $\hat{D}_\varphi(p_\phi, p_{\phi^T})$ with respect to ϕ. In the context of paragraph 2.3, $\hat{D}_\varphi(p_\phi, p_{\phi^T}) = \sup_{\alpha\in\Phi} f(\alpha, \phi)$ for the Gaussian mixture cannot be treated directly using any of the two presented approaches. We propose to assume that Φ is compact, i.e. assume that the means are included in an interval of the form $[\mu_{\min}, \mu_{\max}]$. Now $\hat{D}_\varphi(p_\phi, p_{\phi^T}) = \sup_{\alpha\in\Phi} f(\alpha, \phi)$ is a lower$-\mathcal{C}^1$ function since $f(\alpha, \phi)$ is of class $\mathcal{C}^1(\Phi)$ using Lebesgue theorems. Thus, using Theorem 2.3.1, $\hat{D}_\varphi(p_\phi, p_{\phi^T})$ is continuous and differentiable almost everywhere with respect to ϕ.

The compactness assumption of Φ implies directly the compactness of Φ^0. Indeed

$$\begin{aligned}\Phi^0 &= \left\{\phi \in \Phi, \hat{D}_\varphi(p_\phi, p_{\phi^T}) \leq \hat{D}_\varphi(p_{\phi^0}, p_{\phi^T})\right\} \\ &= \hat{D}_\varphi(p_\phi, p_{\phi^T})^{-1}\left((-\infty, \hat{D}_\varphi(p_{\phi^0}, p_{\phi^T})]\right).\end{aligned}$$

Φ^0 is then the inverse image by a continuous function of a closed set, so it is closed in Φ. Hence, it is compact.

Conclusion 1. *Using Propositions 2.4.4 and 2.4.1, if $\Phi = [\eta, 1-\eta] \times [\mu_{\min}, \mu_{\max}]^2$, the sequence $(\hat{D}_\varphi(p_{\phi^k}, p_{\phi^T}))_k$ defined through formula (1.3.5) converges and there exists a subsequence $(\phi^{N(k)})$ which converges to a stationary point of the estimated divergence. Moreover, every limit point of the sequence $(\phi^k)_k$ is a stationary point of the estimated divergence.*

If we are using the kernel-based dual estimator given by (1.5.3) with a Gaussian kernel density estimator, then function $\phi \mapsto \hat{D}_\varphi(p_\phi, p_{\phi^T})$ is continuously differentiable over Φ even if the means μ_1 and μ_2 are not bounded. For example, take $\varphi = \varphi_\gamma$ defined by (1.1.3). There is one condition which relates the window of the kernel, say w, with the value of γ; $\gamma(w^2-1) > -1$. For $\gamma = 2$ (the Pearson's χ^2), we need that $w^2 > 1/2$. For $\gamma = 1/2$ (the Hellinger), there is no condition on w.

Closedness of Φ^0 is proved similarly to the previous case. Boundedness is however must be treated differently since Φ is not necessarily compact and is supposed to be $\Phi = [\eta, 1-\eta]^s \times \mathbb{R}^s$. For simplicity take $\varphi = \varphi_\gamma$. The idea is to choose ϕ^0 an initialization for the proximal algorithm in a way that Φ^0 does not include unbounded values of the means. Continuity of $\phi \mapsto \hat{D}_\varphi(p_\phi, p_{\phi^T})$ permits to calculate the limits when either (or both) of the means tends to infinity. If both means goes to infinity, then $p_\phi(x) \to 0, \forall x$. Thus, for $\gamma \in (0,\infty) \setminus \{1\}$, we have $\hat{D}_\varphi(p_\phi, p_{\phi^T}) \to \frac{1}{\gamma(\gamma-1)}$. For $\gamma < 0$, the limit is infinity. If only one of the means tends to ∞, then the corresponding component vanishes from the mixture. Thus, if we choose ϕ^0 such that:

$$\hat{D}_\varphi(p_{\phi^0}, p_{\phi^T}) < \min\left(\frac{1}{\gamma(\gamma-1)}, \inf_{\lambda,\mu} \hat{D}_\varphi(p_{(\lambda,\infty,\mu)}, p_{\phi^T})\right) \text{ if } \gamma \in (0,\infty)\setminus\{1\}; \quad (2.6.2)$$

$$\hat{D}_\varphi(p_{\phi^0}, p_{\phi^T}) < \inf_{\lambda,\mu} \hat{D}_\varphi(p_{(\lambda,\infty,\mu)}, p_{\phi^T}) \quad \text{if } \gamma < 0, \quad (2.6.3)$$

then the algorithm starts at a point of Φ whose function value is inferior to the limits of $\hat{D}_\varphi(p_\phi, p_{\phi^T})$ at infinity. By Proposition 2.4.1, the algorithm will continue to decrease the value of $\hat{D}_\varphi(p_\phi, p_{\phi^T})$ and never goes back to the limits at infinity. Besides, the definition of Φ^0 permits to conclude that if ϕ^0 is chosen according to condition (2.6.2,2.6.3), then Φ^0 is bounded. Thus, Φ^0 becomes compact. Unfortunately the value of $\inf_{\lambda,\mu} \hat{D}_\varphi(p_{(\lambda,\infty,\mu)}, p_{\phi^T})$ can be calculated but numerically. We will see next that in the case of Likelihood function, a similar condition will be imposed for the compactness of Φ^0, and there will be no need for any numerical calculus.

Conclusion 2. *Using Propositions 2.4.4 and 2.4.1, under condition (2.6.2, 2.6.3) the sequence* $(\hat{D}_\varphi(p_{\phi^k}, p_{\phi^T}))_k$ *defined through formula (1.5.3) converges and there exists a subsequence* $(\phi^{N(k)})$ *which converges to a stationary point of the estimated divergence. Moreover, every limit point of the sequence* $(\phi^k)_k$ *is a stationary point of the estimated divergence.*

In the case of the likelihood $\varphi(t) = -\log(t) + t - 1$, the set Φ^0 can be written as:

$$\begin{aligned} \Phi^0 &= \left\{\phi \in \Phi, J(\phi) \geq J(\phi^0)\right\} \\ &= J^{-1}\left([J(\phi^0), +\infty)\right), \end{aligned}$$

where J is the log-likelihood function. Function J is clearly of class $\mathcal{C}^1$ on $(\text{int}(\Phi))$. We prove that Φ^0 is closed and bounded which is sufficient to conclude its compactness, since the space $[\eta, 1-\eta]^s \times \mathbb{R}^s$ provided with the euclidean distance is complete.

Closedness. The set Φ^0 is the inverse image by a continuous function (the log-likelihood). Therefore it is closed in $[\eta, 1-\eta]^s \times \mathbb{R}^s$.

Boundedness. By contradiction, suppose that Φ^0 is unbounded, then there exists a sequence $(\phi^l)_l$ which tends to infinity. Since $\lambda^l \in [\eta, 1-\eta]$, then either of μ_1^l or μ_2^l tends to infinity. Suppose that both μ_1^l and μ_2^l tend to infinity, we then have $J(\phi^l) \to -\infty$. Any finite initialization ϕ^0 will imply that $J(\phi^0) > -\infty$ so that $\forall \phi \in \Phi^0, J(\phi) \geq J(\phi^0) > -\infty$. Thus, it is impossible for both μ_1^l and μ_2^l to go to infinity.

Suppose that $\mu_1^l \to \infty$, and that μ_2^l converges[12] to $\mu 2$. The limit of the likelihood has the form:

$$L(\lambda, \infty, \phi_2) = \prod_{i=1}^n \frac{(1-\lambda)}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - \mu_2)^2},$$

which is bounded by its value for $\lambda = 0$ and $\mu_2 = \frac{1}{n}\sum_{i=1}^n y_i$. Indeed, since $1 - \lambda \leq 1$, we have:

$$L(\lambda, \infty, \phi_2) \leq \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - \mu_2)^2}.$$

The right hand side of this inequality is the likelihood of a Gaussian model $\mathcal{N}(\mu_2, 0)$, so that it is maximized when $\mu_2 = \frac{1}{n}\sum_{i=1}^n y_i$. Thus, if ϕ^0 is chosen in a way that $J(\phi^0) > J\left(0, \infty, \frac{1}{n}\sum_{i=1}^n y_i\right)$, the case when μ_1 tends to infinity and μ_2 is bounded would never be allowed. For the other case where $\mu_2 \to \infty$ and μ_1 is bounded, we choose ϕ^0 in a way that $J(\phi^0) > J\left(1, \frac{1}{n}\sum_{i=1}^n y_i, \infty\right)$. In conclusion, with a choice of ϕ^0 such that:

$$J(\phi^0) > \max\left[J\left(0, \infty, \frac{1}{n}\sum_{i=1}^n y_i\right),\ J\left(1, \frac{1}{n}\sum_{i=1}^n y_i, \infty\right)\right] \tag{2.6.4}$$

[12] Normally, μ_2^l is bounded; still, we can extract a subsequence which converges.

the set Φ^0 is bounded.
This condition on ϕ^0 is very natural and means that we need to begin at a point at least better than the extreme cases where we only have one component in the mixture. This can be easily verified by choosing a random vector ϕ^0, and calculate the corresponding log-likelihood value. If $J(\phi^0)$ does not verify the previous condition, we draw again another random vector until satisfaction.

Conclusion 3. *Using Propositions 2.4.4 and 2.4.1, under condition (2.6.4) the sequence $(J(\phi^k))_k$ converges and there exists a subsequence $(\phi^{N(k)})$ which converges to a stationary point of the likelihood function. Moreover, every limit point of the sequence $(\phi^k)_k$ is a stationary point of the likelihood.*

Assumption A3 is not fulfilled (this part applies for all aforementioned situations). As mentioned in the paper of [Tseng, 2004], for the two Gaussian mixture example, by changing μ_1 and μ_2 by the same amount and suitably adjusting λ, the value of $h_i(x|\phi)$ would be unchanged. We explore this more thoroughly by writing the corresponding equations. Let's suppose, by absurd, that for distinct ϕ and ϕ' we have $D_\psi(\phi|\phi') = 0$. By definition of D_ψ, it is given by a sum of non negative terms, which implies that all terms need to be equal to zero. The following lines are equivalent $\forall i \in \{1, \cdots, n\}$:

$$\begin{aligned}
h_i(0|\lambda,\mu_1,\mu_2) &= h_i(0|\lambda',\mu_1',\mu_2') \\
\frac{\lambda e^{-\frac{1}{2}(y_i-\mu_1)^2}}{\lambda e^{-\frac{1}{2}(y_i-\mu_1)^2} + (1-\lambda)e^{-\frac{1}{2}(y_i-\mu_2)^2}} &= \frac{\lambda' e^{-\frac{1}{2}(y_i-\mu_1')^2}}{\lambda' e^{-\frac{1}{2}(y_i-\mu_1')^2} + (1-\lambda')e^{-\frac{1}{2}(y_i-\mu_2')^2}} \\
\log\left(\frac{1-\lambda}{\lambda}\right) - \frac{1}{2}(y_i-\mu_2)^2 + \frac{1}{2}(y_i-\mu_1)^2 &= \log\left(\frac{1-\lambda'}{\lambda'}\right) - \frac{1}{2}(y_i-\mu_2')^2 + \frac{1}{2}(y_i-\mu_1')^2
\end{aligned}$$

Looking at this set of n equations as an equality of two polynomials on y of degree 1 at n points, we deduce that as we dispose of two distinct observations, say, y_1 and y_2, the two polynomials need to have the same coefficients. Thus the set of n equations is equivalent to the following two equations:

$$\left\{\begin{array}{rcl} \mu_1 - \mu_2 &=& \mu_1' - \mu_2' \\ \log\left(\frac{1-\lambda}{\lambda}\right) + \frac{1}{2}\mu_1^2 - \frac{1}{2}\mu_2^2 &=& \log\left(\frac{1-\lambda'}{\lambda'}\right) + \frac{1}{2}{\mu_1'}^2 - \frac{1}{2}{\mu_2'}^2 \end{array}\right. \tag{2.6.5}$$

These two equations with three variables have an infinite number of solutions. Take for example $\mu_1 = 0, \mu_2 = 1, \lambda = \frac{2}{3}, \mu_1' = \frac{1}{2}, \mu_2' = \frac{3}{2}, \lambda' = \frac{1}{2}$.

Remark 2.6.1. The previous conclusion can be extended to any two-component mixture of exponential families having the form:

$$p_\phi(y) = \lambda e^{\sum_{i=1}^{m_1}\theta_{1,i}y^i - F(\theta_1)} + (1-\lambda)e^{\sum_{i=1}^{m_2}\theta_{2,i}y^i - F(\theta_2)}.$$

One may write the corresponding n equations. The polynomial of y_i has a degree of at most $\max(m_1, m_2)$. Thus, if one disposes of $\max(m_1, m_2) + 1$ distinct observations, the two polynomials will have the same set of coefficients. Finally, if $(\theta_1, \theta_2) \in \mathbb{R}^{d-1}$ with $d > \max(m_1, m_2)$, then assumption A3 does not hold.

This conclusion holds for both algorithms (2.1.8) or (2.2.2,2.2.3). Unfortunately, we have no an information about the difference between consecutive terms $\|\phi^{k+1} - \phi^k\|$ except

for the case of $\psi(t) = \varphi(t) = -\log(t) + t - 1$ which corresponds to the classical EM recurrence:

$$\lambda^{k+1} = \frac{1}{n}\sum_{i=1}^{n} h_i(0|\phi^k), \quad \mu_1^{k+1} = \frac{\sum_{i=1}^n y_i h_i(0|\phi^k)}{\sum_{i=1}^n h_i(0|\phi^k)} \quad \mu_1^{k+1} = \frac{\sum_{i=1}^n y_i h_i(1|\phi^k)}{\sum_{i=1}^n h_i(1|\phi^k)}.$$

Tseng [2004] has shown that we can prove directly that $\phi^{k+1} - \phi^k$ converges to 0.

2.6.2 Two-component Weibull mixture

Let p_ϕ be a two-component Weibull mixture:

$$p_\phi(x) = 2\lambda\alpha_1(2x)^{\alpha_1-1}e^{-(2x)^{\alpha_1}} + (1-\lambda)\frac{\alpha_2}{2}\left(\frac{x}{2}\right)^{\alpha_2-1} e^{-\left(\frac{x}{2}\right)^{\alpha_2}} \tag{2.6.6}$$

We have $\Phi = (0,1) \times \mathbb{R}_+^* \times \mathbb{R}_+^*$. Similarly to the Gaussian example, we will study convergence properties in light of our theoretical approach. We will only be interested in divergences with the class of Cressie-Read functions $\varphi = \varphi_\gamma$ given by (1.1.3).
The weight functions h_i are given by:

$$h_i(1|\phi) = \frac{2\lambda\alpha_1(2x)^{\alpha_1-1}e^{-(2x)^{\alpha_1}}}{2\lambda\alpha_1(2x)^{\alpha_1-1}e^{-(2x)^{\alpha_1}} + (1-\lambda)\frac{\alpha_2}{2}\left(\frac{x}{2}\right)^{\alpha_2-1} e^{-\left(\frac{x}{2}\right)^{\alpha_2}}}, \quad h_i(2|\phi) = 1 - h_i(1|\phi).$$

It is clear the functions h_i are of class $\mathcal{C}^1(\text{int}(\Phi))$ and so does $\phi \mapsto D_\psi(\phi, \phi')$ for any $\phi' \in \Phi$.
If we are using the dual estimator defined by (1.3.5), then continuity can be treated similarly to the case of the Gaussian example. Here, however, the continuity and differentiability of the optimized function $f(\alpha, \phi)$, where $\hat{D}_\varphi(p_\phi, p_{\phi^T}) = \sup_\alpha f(\alpha, \phi)$, are more technical. We list the following three results without any proof, because it suffices to study the integral term in the formula. Suppose, without loss of generality, that $\phi_1 < \phi_2$ and $\alpha_1 < \alpha_2$.

1. For $\gamma > 1$, which includes the Pearson's χ^2 case, the dual representation is *not* well defined since $\sup_\alpha f(\alpha, \phi) = \infty$;
2. For $\gamma \in (0,1)$, function $f(\alpha, \phi)$ is continuous.
3. For $\gamma < 0$, function $f(\alpha, \phi)$ is continuous and well defined for $\phi_1 < \frac{\gamma-1}{\gamma}\alpha_1$ and $\alpha_2 \geq \phi_2$. Otherwise $f(\alpha, \phi) = -\infty$, but the supremum $\sup_\alpha f(\alpha, \phi)$ is still well defined.

In both cases 2 and 3, differentiability of function $f(\alpha, \phi)$ holds only on a subset of $\Phi \times \Phi$ which cannot be written as $A \times B$, and thus the theoretical approaches presented in Section 2.3 are not suitable. In order to end this part, we emphasize the fact that, similarly to the Gaussian example, even continuity of the estimated divergence $\hat{D}_\varphi(p_\phi, p_{\phi^T})$ with respect to ϕ cannot be treated by our theoretical approaches unless we suppose that Φ is compact. If Φ is compact, function $f(\alpha, \phi)$ becomes level-bounded and Theorem 2.3.2 applies and we can deduce that the estimated divergence is continuous. Differentiability is far more subtle if we use Theorem 2.3.1.
Similar conclusion as Conclusion 1 can be stated here with no changes except for the fact that assumption AC is not fulfilled. This entails that our conclusion will be about the subgradient of the estimated divergence.

Remark 2.6.2 (Strict continuity of $\hat{D}_\varphi$ under boundedness assumption of the shape parameters)**.** When $\gamma < 0$, we can prove strict continuity of $\phi \mapsto \hat{D}_\varphi(p_\phi, p_{\phi^T})$ using Theorem 2.3.3. We need to calculate Y_∞ defined by:

$$Y_\infty(\phi_0) = \bigcup_{\alpha \in \arg\sup_\beta f(\beta,\phi_0)} M_\infty(\alpha,\phi_0), \quad \text{for } M_\infty(\alpha,\phi_0) = \{a | (0,a) \in \partial^\infty f(\alpha,\phi_0)\}.$$

Let $\alpha_0 \in \arg\sup_\beta f(\beta,\phi_0)$. Since the value of $f(\beta,\phi)$ is $-\infty$ on the set $\{(\alpha,\phi) \in \Phi | \alpha_2 < \phi_2\}$, then its supremum over β is attained outside of it. Consequently, (α_0,ϕ_0) belongs to the set $\{(\alpha,\phi) \in \Phi | \alpha_2 \geq \phi_2\}$ where the integral in function f is of class $\mathcal{C}^1$ which implies that f is also of class $\mathcal{C}^1$. This entails that $f(\alpha,\phi)$ is strictly continuous at (α_0,ϕ_0) which is, by Theorem 9.13 in [Rockafellar and Wets, 1998], equivalent to $\partial^\infty f(\alpha_0,\phi_0) = \{0\}$. Now, we may conclude that $M_\infty(\alpha,\phi_0) = \{0\}$, and hence $Y_\infty = \{0\}$. All ingredients of Theorem 2.3.1 are ready, and we conclude that the dual representation of the divergence $\phi \mapsto \hat{D}_\varphi(p_\phi, p_{\phi^T})$ is strictly continuous.
If we could prove that the set $Y(\phi)$ contains only one element given a vector ϕ, then the differentiability of the estimated divergence would be obtained using point (b) of Theorem 2.3.3. This demands however a great effort since the characterization of the set $Y(\phi)$ demands an investigation of the form of the estimated divergence and the model used.

If we are using the kernel-based dual estimator given by (1.5.3) with a Gaussian kernel density estimator, then things are a lot simplified. We need only to treat the integral term. From an analytic point of view, the study of continuity depends on the kernel used; more specifically its tail behavior. If we take a Gaussian kernel, then we have:

- For $\gamma > 1$, it is necessary that $\min(\phi_1,\phi_2) > 2$, otherwise the estimated divergence is infinity. Thus, it is necessary for either of the true values of the shapes to be inferior to 2 in order for the estimation to be valid;
- For $\gamma \in (0,1)$, then the estimated divergence is $\mathcal{C}^1(\text{int}(\Phi))$;
- For $\gamma < 0$, it is necessary that $\min(\phi_1,\phi_2) < 1 - \frac{1}{\gamma}$ and $\max(\phi_1,\phi_2) < 2$. If these conditions do not hold, then the estimated divergence is minimized at $-\infty$ at any vector of parameters which does not verify the previous condition.

In the first case, if we use a heavier-tailed kernel such as the Cauchy Kernel, the estimated divergence becomes $\mathcal{C}^1(\text{int}(\Phi))$. In the third case, if we use a compact-supported kernel such as the Epanechnikov's kernel, the condition is reduced to only $\min(\phi_1,\phi_2) < 1 - \frac{1}{\gamma}$. Similar conditions to (2.6.2,2.6.3) can be obtained and we have the same conclusion as Conclusion 2.
In the case of the Likelihood $\varphi(t) = -\log(t) + t - 1$, we illustrate the convergence of the EM algorithm through our theoretical approach. Assumptions A1 and AC are clearly verified since both the log-likelihood and the proximal term are sums of continuously differentiable functions, and integrals do not intervene here. The set Φ^0 is given by:

$$\begin{aligned} \Phi^0 &= \left\{\phi \in \Phi, J(\phi) \geq J(\phi^0)\right\} \\ &= J^{-1}\left([J(\phi^0),\infty)\right) \\ &= \left\{\phi \in \Phi, L(\phi) \geq L(\phi^0)\right\} \end{aligned}$$

where $L(\phi)$ is the likelihood of the model, and $J(\phi) = \log(L(\phi))$ is the log-likelihood function. We will show that under similar conditions to the Gaussian mixture, the set Φ^0 is compact.

Closedness of Φ^0. Since the shape parameter is supposed to be positive, continuity of the log-likelihood would imply only that Φ^0 is closed in $[0,1] \times \mathbb{R}_+^* \times \mathbb{R}_+^*$, a space which is not closed and hence is not complete. We therefore, propose to extend the definition of shape parameter on 0. From a statistical point of view, this extension is not reasonable since the density function of Weibull distribution with a shape parameter equal to 0 is the zero function which is not a probability density. Besides, identifiability problems would appear for a mixture model. Nevertheless, our need is only for analytical purpose. We will add suitable conditions on ϕ^0 in order to avoid such subtlety keeping in hand the closedness property.
We suppose now that the shape parameter can have values in $\mathbb{R}_+$. The set Φ^0 is now the inverse image of $[L(\phi^0), \infty)$ by the likelihood function[13] which is continuous on $[0,1] \times \mathbb{R}_+ \times \mathbb{R}_+$. Hence, it is closed in the space $[0,1] \times \mathbb{R}_+ \times \mathbb{R}_+$ embedded with the euclidean norm which is complete. It suffices then to prove that Φ^0 is bounded.
Boundedness of Φ^0. We will make similar arguments to the case of the Gaussian mixture example. We need to calculate the limit at infinity when the shape parameter of either of the two components tends to infinity. If both α_1 and α_2 goes to infinity, the log-likelihood tends to $-\infty$. Hence any choice of a finite ϕ^0 can avoid this case. Suppose now that α_1 goes to infinity whereas α_2 stays bounded. The corresponding limit of the log-likelihood functions is given by:

$$J(\lambda, \infty, \alpha_2) = \sum_{i=1}^{n} \log \left((1-\lambda) \frac{\alpha_2}{2} \left(\frac{y_i}{2} \right)^{\alpha_2 - 1} e^{-\left(\frac{y_i}{2}\right)^{\alpha_2}} \right)$$

if there is no observation y_i equal to $\frac{1}{2}$. In fact, if there is $y_i = \frac{1}{2}$, the limit is $+\infty$ and the set Φ^0 cannot be bounded. However, it is improbable to get such an observation since the probability of getting an observation equal to $\frac{1}{2}$ is zero. The case when α_2 goes to infinity whereas α_1 stays bounded is treated similarly.
To avoid the two previous scenarios, one should choose the initial point of the algorithm ϕ^0 in a way that it verifies:

$$J(\phi^0) > \max \left(\sup_{\lambda, \alpha_2} J(\lambda, \infty, \alpha_2), \sup_{\lambda, \alpha_1} J(\lambda, \alpha_1, \infty) \right). \tag{2.6.7}$$

Since all vectors of Φ^0 have a log-likelihood value greater than $J(\phi^0)$, the previous choice permits the set Φ^0 to avoid non finite values of ϕ. Thus it becomes bounded whenever ϕ_0 is chosen according to condition (2.6.7). Finally, the calculus of both terms $\sup_{\lambda,\alpha_1} J(\lambda, \alpha_1, \infty)$ and $\sup_{\lambda,\alpha_2} J(\lambda, \alpha_2, \infty)$ is not feasible but numerically. Those, however, can be simplified a little. One can notice by writing these terms without the logarithm (as a product), the term which has λ is maximized when it is equal to 1. The remaining of the calculus is a maximization of the likelihood function of a Weibull model[14].
We conclude that the set Φ^0 is compact under condition (2.6.7). Finally, it is important to notice that condition (2.6.7) permits also to avoid the border values which corresponds to $\alpha_1 = 0$ or $\alpha_2 = 0$. Indeed, when either of the shape parameters is zero, the corresponding component vanishes and the corresponding log-likelihood value is less than the upper bound in condition (2.6.7). The same conclusion as Conclusion 3 can be stated here for the Weibull mixture model.

[13] We do not use this time the log-likelihood function since it is not defined when both shape parameters are zero.

[14] In a Weibull model, the calculus of the MLE cannot be done but numerically when the parameter of interest is the shape parameter.

Notice that the verification of assumption A3 is a hard task here because it results in a set of n non-linear equations.

2.6.3 Pearson's χ^2 algorithm for a Cauchy model

Let $\{(x_i, y_i), i = 0, \cdots, n\}$ be an n-sample drawn from the joint probability law defined by the density function:

$$f(x, y|a, x_0) = \frac{a(y - x_0)^2 e^x}{\pi \left(a^2 + (y - x_0)^2 e^x\right)^2}, \quad x \in [0, \infty), y \in \mathbb{R}$$

where $a \in [\varepsilon, \infty)$, with $\varepsilon > 0$, denotes a scale parameter and $x_0 \in \mathbb{R}$ denotes a location parameter. We define an exponential probability law with parameter $\frac{1}{2}$ on the labels. It is given by the density function:

$$q(x) = \frac{1}{2} e^{-x/2}.$$

Now, the model defined on the observed data becomes a Cauchy model with two parameters:

$$p_{(a,x_0)}(y) = \int_0^\infty f(x, y|a, x_0)dx = \frac{a}{\pi(a^2 + (y - x_0)^2)}, \quad a \geq \varepsilon > 0, x_0 \in \mathbb{R}.$$

The goal of this example is to show how we prove assumptions A1-3 and AC in order to explore the convergence properties of the sequence ϕ^k generated by either of the algorithms (2.1.8) and (2.2.2,2.2.3). We also discuss the analytical properties of the dual representation of the divergence.
In this example, we only focus on the dual representation of the divergence given by (1.3.5) because the resulting MDφDE is robust against outliers (so does the MLE). Thus there is no need to use a robust estimator such as the kernel-based MDφDE which needs a choice of a suitable kernel and window.

Cauchy model with zero location

We suppose here that $x_0 = 0$, and we are only interested in estimating the scale parameter a. The Pearson's χ^2 divergence is given by:

$$D(p_a, p_{a^*}) = \frac{1}{2} \int \left[\frac{p_a(y)}{p_{a^*}} - 1\right]^2 p_{a^*}(y)dy.$$

Let's rewrite the dual representation of the Chi square divergence:

$$\hat{D}(p_a, p_{a^*}) = \sup_{b \geq \varepsilon} \left\{ \int_{\mathbb{R}} \frac{p_b^2(x)}{p_a(x)} dx - \frac{1}{2n} \sum_{i=1}^n \frac{p_b^2(y_i)}{p_a^2(y_i)} \right\} - \frac{1}{2}.$$

A simple calculus shows:

$$\int_{\mathbb{R}} \frac{p_b^2(x)}{p_a(x)} dx = \frac{(a^2 + b^2)\pi}{2ab}.$$

This implies a simpler form for the dual representation of the divergence:

$$\hat{D}(p_a, p_{a^*}) = \sup_{b \geq \varepsilon} \left\{ \frac{(a^2 + b^2)}{2ab} - \frac{1}{2n} \sum_{i=1}^n \frac{a^2(b^2 + y_i^2)^2}{b^2(a^2 + y_i^2)^2} \right\} - \frac{1}{2}. \tag{2.6.8}$$

Let $f(a,b)$ denote the optimized function in the above formula. We calculate the first derivative with respect to b:

$$\frac{\partial f}{\partial b}(a,b) = -\frac{\pi a}{2b^2} + \frac{\pi}{2a} - \frac{1}{2n}\sum_{i=1}^{n} \frac{a^2}{(a^2+y_i^2)^2}\left(2b - \frac{2y_i^4}{b^3}\right).$$

Notice that as $a \geq \varepsilon$ the term $\frac{\pi}{2a}$ stays bounded away from infinity uniformly. Therefore, it suffices then that b exceeds a finite value b_0 in order that the derivative becomes negative. Hence, there exists b_0 such that $b \mapsto f(a,b)$ becomes decreasing independently of a. On the other hand $\forall a > 0, \lim_{b\to\infty} f(a,b) = -\infty$. It results that all values of the function $b \mapsto f(a,b)$ for $b > b_0$ does not have any use in the calculus of the supremum in (2.6.8), since, by the decreasing property if $b \mapsto f(a,b)$, they all should have values less than the value at b_0. We may now rewrite the dual representation of the Chi square divergence as :

$$\hat{D}(p_a, p_{a^*}) = \sup_{b\in[\varepsilon,b_0]} \left\{ \frac{(a^2+b^2)}{2ab} - \frac{1}{2n}\sum_{i=1}^{n} \frac{a^2(b^2+y_i^2)^2}{b^2(a^2+y_i^2)} \right\} - \frac{1}{2}. \tag{2.6.9}$$

We have now two pieces of information about $f(a,b)$. The first is that it is level-bounded locally in b uniformly in a (see paragraph (2.3.2)). The second is that we are exactly in the context of lower$-\mathcal{C}^1$ functions (2.3.1). First of all, function f is $\mathcal{C}^1([\varepsilon,\infty)\times[\varepsilon,\infty))$ function, so that part (a) of Theorem 2.3.3 is verified and the function $a \mapsto \hat{D}(p_a, p_{a^*})$ is strictly continuous. To prove it is continuously differentiable, we need to prove that the set

$$Y(a) = \bigcup_{b\in\arg\max_{b'} f(a,b)} \left\{ \frac{\partial f}{\partial a}(a,b) \right\}$$

contains but one element. From a theoretic point of view, two possible methods are available: Prove that either there is a unique maximum for a fixed a, or that the derivative with respect to a at all maxima *does not depend* on a (they have the same value). In our example, function $b \mapsto f(a,b)$ is not concave. We may also plot it using any mathematical tool provided that we already have the data set. We tried out a simple example and generated a 10-sample of the standard Cauchy distribution ($a = 1$), see table (2.2). We used Mathematica to draw a 3D figure of function f, see figure (2.2).

y_i	0.534	-18.197	0.726	-0.439	-1.945	0.0119	12.376	-0.953	0.698	0.818

Table 2.2: A 10-sample Cauchy dataset.

It is clear that for a fixed a, the function $b \mapsto f(a,b)$ has two maxima which may both be global maxima. For example for $a = 0.9$, one gets figure (2.3). It is clearer now that conditions of Theorem 2.3.3 are not fulfilled, and we cannot prove that function $\hat{D}(p_a, p_{a^*})$ is continuously differentiable every where.

It is however not the end of the road. We still have the results presented in paragraph (2.3.1). Function $\hat{D}(p_a, p_{a^*})$ is lower-$\mathcal{C}^1$. Therefore, it is strictly continuous and almost everywhere continuously differentiable. Hence, we may hope that the limit points of the sequence $(\phi^k)_k$ for algorithm (2.1.8) are in the set of points where the dual representation of the Chi square divergence is $\mathcal{C}^1$, or be more reasonable and state any further result on the sequence in terms of the subgradient of $\hat{D}(p_a, p_{a^*})$.

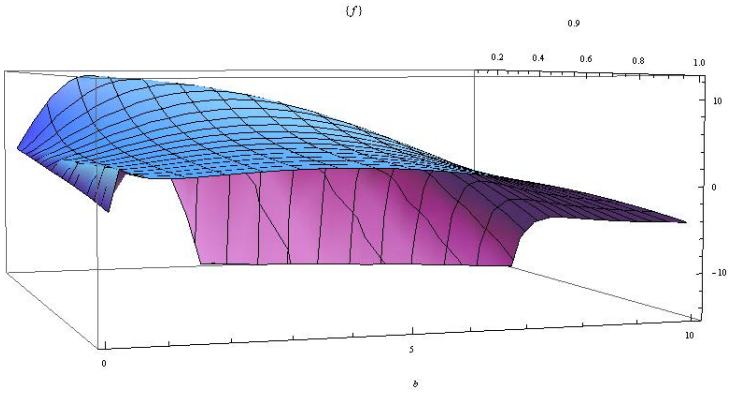

Figure 2.2: A 3D plot of function $f(a, b)$ for a 10-sample of the standard Cauchy distribution.

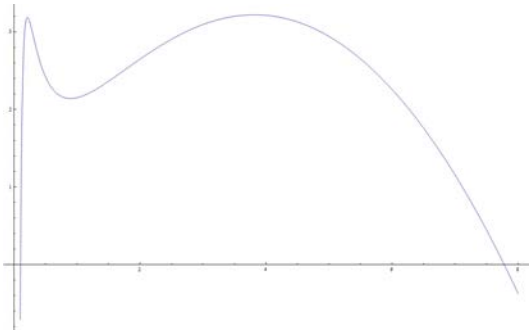

Figure 2.3: A 2D plot of function $f(0.9, b)$ for a 10-sample of the standard Cauchy distribution.

Compactness of Φ^0. We check when the set $\Phi^0 = \{a|\hat{D}(p_a, p_{a^*}) \leq \hat{D}(p_{a_0}, p_{a^*})\}$ is closed and bounded in $[\varepsilon, \infty)$ for an initial point a_0. **Closedness** is proved using continuity of $\hat{D}(p_a, p_{a^*})$. Indeed,

$$\Phi^0 = \hat{D}^{-1}(p_a, p_{a^*}) \left((-\infty, \hat{D}(p_{a_0}, p_{a^*})] \right).$$

Boundedness is proved by contradiction. Suppose that Φ^0 is unbounded, then there exists a sequence $(a^l)_l$ of points of Φ^0 which goes to infinity. Formula (2.6.9) shows that b stays in a bounded set during the calculus of the supremum. Hence the continuity of $\hat{D}(p_a, p_{a^*})$ implies:

$$\lim_{a \to \infty} \hat{D}(p_a, p_{a^*}) = +\infty.$$

This shows that by choosing any finite a_0, the set Φ^0 becomes bounded. Indeed, the relation defining Φ^0 implies that $\forall l, \hat{D}(p_{a^l}, p_{a^*}) \leq \hat{D}(p_{a_0}, p_{a^*}) < \infty$, and a contradiction is reached by taking the limit of each part of this inequality. Hence Φ^0 is closed and bounded in the space $[\varepsilon, \infty)$ which is complete provided with the euclidean distance. We conclude that Φ^0 is compact[15].
In this simple example, we only can use algorithm (2.1.8) since there is only one parameter of interest. Proposition 2.4.4 can be used to deduce convergence of any convergent subsequence to a generalized stationary point of $\hat{D}(p_a, p_{a^*})$.

[15] If we are to use a result which concerns the differentiability of $\hat{D}(p_a, p_{a^*})$, one should consider the case when Φ^0 shares a boundary with Φ. A possible solution to avoid this is to consider an initial point a^0 such that $\hat{D}(p_\varepsilon, p_{a^*}) > \hat{D}(p_{a_0}, p_{a^*})$. This expels the the boundary from the possible values of Φ^0.

To deduce more results about the sequence $(a^k)_k$, we may try and verify assumption A3 using Lemma 2.4.1. Let's write functions h_i.

$$h_i(x|a) = \frac{f(x, y_i|a)}{p_a(y_i)} = \frac{y_i^2 e^x (a^2 + y_i^2)}{(a^2 + e^x y_i^2)^2}.$$

Clearly, for any $i \in \{1, \cdots, n\}$ and $a \geq \varepsilon$, function $x \mapsto h_i(x|a)$ is continuous. Let $a, b \geq \varepsilon$ such that $a \neq b$. Suppose that:

$$\forall i, \quad h_i(x|a) = h_i(x|b) \qquad \forall x \geq 0.$$

This entails that:

$$a^2b^4 - a^4b2 + (b^4 - a^4)y_i^2 + \left(a^2e^{2x} + 2b^2e^x - b^2e^{2x} - 2a^2e^x\right) y_i^4 = 0, \qquad i = 1, \cdots, n.$$

This is a polynomial on y_i of degree 4 which coincides with the zero polynomial on n points. If there exists 5 distinct observations[16], then the two polynomials will have the same coefficients. Hence, we have $b^4 - a^4 = 0$. This implies that $a = b$ since they are both positive real numbers. We conclude that $D_\psi(a, b) = 0$ whenever $a = b$ which is equivalent to assumption A3. Proposition 2.4.3 can now be applied to deduce that sequence (a^k) defined by (2.1.8) (with ϕ^k replaced by a^k) is well defined and bounded. Furthermore, it verifies $a^{k+1} - a^k \to 0$, and the limit of any convergent subsequence is a generalized stationary point of $\hat{D}(p_a, p_{a^*})$. The existence of such subsequence is guaranteed by the compactness of Φ^0 and the fact that $\forall k, a^k \in \Phi^0$.

Cauchy model with both parameters

The model is now defined by:

$$p_{(a,x_0)}(y) = \frac{a}{\pi(a^2 + (y - x_0)^2)}, \quad a \geq \varepsilon > 0, x_0 \in \mathbb{R}.$$

Formula (2.6.8) of the dual representation of the Chi square divergence becomes:

$$\hat{D}(p_{a,x_0}, p_{a^*,x_0}) = \sup_{b \geq \varepsilon, x_1 \in \mathbb{R}} \left\{ \frac{(a^2 + b^2 + (x_1 - x_0)^2)}{2ab} - \frac{1}{2n} \sum_{i=1}^{n} \frac{a^2(b^2 + (y_i - x_1)^2)^2}{b^2(a^2 + (y_i - x_0)^2)^2} \right\} - \frac{1}{2}. \tag{2.6.10}$$

Let $f(a, b, x_0, x_1)$ be the optimized function in the previous formula. This time, it does not seem easy to prove that the supremum can be calculated on a compact set. We, therefore, work on the second approach to study continuity of $\hat{D}(p_{a,x_0}, p_{a^*,x_0})$, i.e. level-boundedness approach (see paragraph (2.3.2)). For a, let $(a-\tilde{a}, a+\tilde{a}) \subset [\varepsilon, \infty)$ be an open neighborhood around a, and for x_0, let $(x_0 - \tilde{x}, x_0 + \tilde{x})$ be an open neighborhood around x_0. It is clear that as either $b \to \infty$ or $x_1 \to \pm\infty$, we have $f(a, b, x_0, x_1) \to -\infty$ since the first term in f is of order b (resp. x_1^2) whereas the second term in f is of order b^2 (resp. x_1^4) as long as a is bounded away from zero and x_0 is supposed to be bounded. Finally, when both b and x_1 goes to infinity, the important terms in calculating the limit are of order $b - b^2$ and $\frac{x_1^2}{b} - \left(\frac{x_1^2}{b}\right)^2$. Hence the limit is *a fortiori* $-\infty$. We conclude that:

$$f(a, b, x_0, x_1) \xrightarrow[\|(b,x_1)\| \to \infty]{} -\infty$$

[16] If one uses the point $x = 0$, the result follows directly without supposing the existence of distinct observations.

Now that f is level-bounded in (b, x_1) locally uniformly in (a, x_0), and since f is readily continuous (so it is upper semicontinuous), all ingredients for Theorem 2.3.3 part (a) are ready. Hence $\hat{D}(p_{a,x_0}, p_{a^T,x_0^T})$ is strictly continuous[17].

Now that $\hat{D}(p_{a,x_0}, p_{a^*,x_0})$ is continuous, we may use analogous arguments to those given in the previous paragraph to prove closedness and boundedness of Φ^0. Boundedness is treated a bit differently since the supremum is no longer calculated over a bounded set. By definition of the supremum, one can write:

$$\sup_{b\geq\varepsilon, x_1\in\mathbb{R}} \left\{ \frac{((a^l)^2 + b^2 + (x_1 - x_0^l)^2)}{2a^l b} - \frac{1}{2n}\sum_{i=1}^{n} \frac{(a^l)^2(b^2 + (y_i - x_1)^2)^2}{b^2((a^l)^2 + (y_i - x_0^l)^2)^2} \right\} - \frac{1}{2} \geq$$
$$\left\{ \frac{((a^l)^2 + b'^2 + (x_1' - x_0^l)^2)}{2a^l b'} - \frac{1}{2n}\sum_{i=1}^{n} \frac{(a^l)^2(b'^2 + (y_i - x_1')^2)^2}{b'^2((a^l)^2 + (y_i - x_0^l)^2)^2} \right\} - \frac{1}{2}$$

for any $b' \geq \varepsilon$ and $x_1' \in \mathbb{R}$. As the sequence (a^l, x_0^l) goes to infinity, the second hand of the previous inequality tends to infinity. Hence the limit of the left hand is also infinity. Thus, $\lim_{l\to\infty} \hat{D}(p_{a^l,x_0^l}, p_{a^T,x_0^T}) = \infty$. We conclude that by choosing any finite initialization, the set Φ^0 becomes bounded. As we could not give any argument about the differentiability of $\hat{D}(p_{a,x_0}, p_{a^*,x_0})$, the only theoretical results we may state are about the subgradient of $\hat{D}(p_{a,x_0}, p_{a^*,x_0})$.

Finally, we prove that $D_\psi((a, x_0), (b, x_1)) > 0$ as $(a, x_0) \neq (b, x_1)$. We use Lemma 2.4.1. Let $x = 0$; a point at which h_i is (right)[18] continuous. We need to prove that there exists i such that if $h_i(0|a, x_0) = h_i(0|b, x_1)$ then $a = b$ and $x_0 = x_1$. We have $h_i(0|a, x_0) = h_i(0|b, x_1)$ is equivalent to

$$\frac{(y_i - x_0)^2(a^2 + (y_i - x_0)^2)}{(a^2 + (y_i - x_0)^2)^2} = \frac{(y_i - x_1)^2(b^2 + (y_i - x_1)^2)}{(b^2 + (y_i - x_1)^2)^2}, \quad \forall i \in \{1, \cdots, n\},$$

or equivalently

$$\left(b^2 + x_0^2 + x_1^2 + 4x_1x_0\right) y_i^2 - 2\left(b^2x_0 + x_0x_1^2 + x_1x_0^2\right) y_i + b^2x_0^2 + x_1^2x_0^2 =$$
$$\left(a^2 + x_1^2 + x_0^2 + 4x_0x_1\right) y_i^2 - 2\left(a^2x_1 + x_1x_0^2 + x_0x_1^2\right) y_i + a^2x_1^2 + x_0^2x_1^2.$$

Suppose that there are at least three distinct observations. The previous identities can be rewritten as an identity between two polynomial of degree 2 in y_i. Since they are equal at three distinct roots, they must be identical and all coefficients are equal with their corresponding ones. The coefficient of y_i^2 suffices to deduce that is $a = b$. Identify now the coefficients of y_i to get that $b^2x_0 = a^2x_1$. Since $a, b > 0$, then $x_0 = x_1$.

We finally conclude using Proposition 2.4.3 that if we use algorithm (2.1.8) or algorithm (2.2.2,2.2.3), the distance between two consecutive terms of the sequence $(a^k, x_0^k)_k$ tends to 0 and any limit point of the sequence is a generalized stationary point of $\hat{D}(p_{a,x_0}, p_{a^T,x_0^T})$.

2.7 Simulation study

We summarize the results of 100 experiments on 100-samples (with and without outliers) from two-component Gaussian and Weibull mixtures by giving the average of the error

[17] In order to get the same results we have on lower-$\mathcal{C}^1$ functions, we need to prove that $\hat{D}(p_{a,x_0}, p_{a^T,x_0^T})$ is also regular in the sense of Clarke at all points of its domain; a result which we get by a theorem of Rademarcher, see [Rockafellar and Wets, 1998] Chapter 9.

[18] In the proof of the lemma, we use continuity to deduce a certain result in a neighborhood of a point at which function h_i is continuous. Here, right continuity still gives us a neighborhood of the form $[x, x + \epsilon)$ which suffices to complete the proof.

committed with the corresponding standard deviation. The error criterion is mainly the total variation distance (TVD) defined by (1.7.2). We also provide for the Gaussian mixture the values of the χ^2 divergence between the estimated model and the true mixture, defined by (1.7.1), since it gave infinite values for the Weibull experiment.
We used different $\varphi-$divergences to estimate the parameters and compared the performances of the two dual formulas of estimating a $\varphi-$divergence (1.3.5) and (1.5.3). We also included the MDPD of Basu et al. [1998] defined by (1.3.8) for a tradeoff parameter $a = 0.5$. This parameter resulted in very good results throughout all simulations carried in the previous chapter. Other estimators of the divergence could also be considered in a future work, see the simulations of the previous chapter for a detailed comparison between these estimators. For the Gaussian mixture, we used the Pearson's χ^2 and the Hellinger divergences, whereas in the Weibull mixture, we used the Neymann's χ^2 and the Hellinger divergences. For the proximal term, we used $\psi(t) = \frac{1}{2}(\sqrt{t}-1)^2$. We illustrate also the performance of the EM method in the light of our method, i.e. using initializations verifying conditions (2.6.4) for the Gaussian mixture and conditions (2.6.7) for the Weibull one. When outliers were added, these initializations did not give always good results and the convergence of the proportion was towards the border $\eta = 0.1$ or $1-\eta = 0.9$. In such situations, the EM algorithm was initialized using another starting point manually.
We used the Nelder-Mead algorithm (see [Nelder and Mead, 1965]) for all optimization calculus. The method proved to be more efficient in our context than other optimization algorithms although having a slow convergence. Such method is derivative-free and applies even if the the objective function is not differentiable which may be the case of the estimated divergence defined through (1.3.5). The Nelder-Mead algorithm is known to give good results in problems with dimension at least 2 and does not perform well in dimension 1. We thus used Brent's method in such cases. It is also a derivative-free method which works only in a compact subset from $\mathbb{R}$. The calculus was done under the statistical tool [R Core Team, 2015]. In what concerns numerical integrations, see Section 1.7.

2.7.1 The two-component Gaussian mixture revisited

We consider the Gaussian mixture (2.6.1) presented earlier with true parameters $\lambda = 0.35, \mu_1 = 2, \mu_2 = 1.5$ and fixed variances $\sigma_1 = \sigma_2 = 1$. Since we are using an error function criterion, label-switching problems do not interfere. Figure (2.4) shows the values of the estimated divergence for both formulas (1.3.5) and (1.5.3) on a logarithmic scale at each iteration of the algorithm. The 1-step algorithm refers to algorithm (2.1.8), whereas 2-step refers to algorithm (2.2.2,2.2.3). We omitted the initial point in order to produce a clear image of the decrease of the objective function. For the kernel-based dual formula, we used a Gaussian kernel. Results are given in table (2.3).
We used the same data simulated in paragraph 1.7.2, so that contamination was done by adding in the original sample to the 5 lowest values random observations from the uniform distribution $\mathcal{U}[-5,-2]$. We also added to the 5 largest values random observations from the uniform distribution $\mathcal{U}[2,5]$. Results are presented in table (2.4).
It is clear that the kernel-based MDφDE is more robust than the EM algorithm and the classical MDφDE for both the Pearson's χ^2 and the Hellinger divergences. Differences between the two divergences are not significant for both estimation methods of the divergence. Besides, in comparison with the results obtained with a direct optimization in paragraph 1.7.2, we find no significant differences. The proximal point algorithm wored as well on the density power divergence. The MDPD produced robust estimates with minor differences with respect to the kernel-based MDφDE in favor of the former.

Estimation method		Error criterion	
		$\sqrt{\chi^2}$	TVD
Chi square			
Algorithm (2.1.8)	MDφDE	0.108, sd = 0.052	0.061, sd = 0.029
	kernel-based MDφDE	0.118 , sd = 0.052	0.066 ,sd= 0.027
Algorithm (2.2.2,2.2.3)	MDφDE	0.108, sd = 0.052	0.061, sd = 0.029
	kernel-based MDφDE	0.118, sd = 0.051	0.066 ,sd= 0.027
Hellinger			
Algorithm (2.1.8)	MDφDE	0.108, sd = 0.052	0.050 , sd=0.025
	kernel-based MDφDE	0.113, sd = 0.044	0.064 ,sd=0.025
Algorithm (2.2.2,2.2.3)	MDφDE	0.108, sd = 0.052	0.061, sd = 0.029
	kernel-based MDφDE	0.113, sd = 0.045	0.064 ,sd=0.025
MDPD $a = 0.5$ - Algorithm (2.1.8)		0.117, sd = 0.049	0.065, sd = 0.025
MDPD $a = 0.5$ - Algorithm (2.2.2,2.2.3)		0.117, sd = 0.047	0.065, sd = 0.025
EM		0.113, sd = 0.044	0.064 , sd = 0.025

Table 2.3: The mean value of errors committed in a 100-run experiment with the standard deviation. No outliers are considered here. The divergence criterion is the Chi square divergence or the Hellinger. The proximal term is calculated with $\psi(t) = \frac{1}{2}(\sqrt{t} - 1)^2$.

Estimation method		Error criterion	
		χ^2	TVD
Chi square			
Algorithm (2.1.8)	MDφDE	0.334, sd = 0.097	0.146,sd=0.036
	kernel-based MDφDE	0.149 , sd = 0.059	0.084 ,sd=0.033
Algorithm (2.2.2,2.2.3)	MDφDE	0.333, sd = 0.097	0.149, sd = 0.033
	kernel-based MDφDE	0.149 , sd = 0.059	0.084, sd=0.033
Hellinger			
Algorithm (2.1.8)	MDφDE	0.321, sd = 0.096	0.146, sd=0.034
	kernel-based MDφDE	0.155 , sd = 0.059	0.087 ,sd=0.033
Algorithm (2.2.2,2.2.3)	MDφDE	0.322, sd = 0.097	0.147, sd = 0.034
	kernel-based MDφDE	0.156 , sd = 0.059	0.087 ,sd=0.033
MDPD $a = 0.5$ - Algorithm (2.1.8)		0.129, sd = 0.049	0.065, sd = 0.025
MDPD $a = 0.5$ - Algorithm (2.2.2,2.2.3)		0.138, sd = 0.053	0.078, sd = 0.030
EM		0.335, sd = 0.102	0.150, sd = 0.034

Table 2.4: Error committed in estimating the parameters of a 2-component Gaussian mixture with 10% outliers. The divergence criterion is the Chi square divergence or the Hellinger. The proximal term is calculated with $\psi(t) = \frac{1}{2}(\sqrt{t} - 1)^2$.

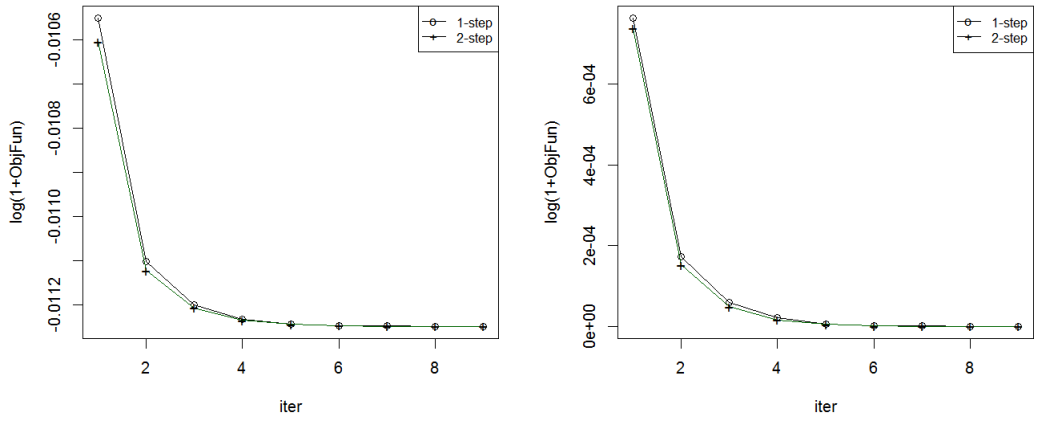

Figure 2.4: Decrease of the (estimated) Hellinger divergence between the true density and the estimated model at each iteration in the Gaussian mixture. The figure to the left is the curve of the values of the kernel-based dual formula (1.5.3). The figure to the right is the curve of values of the classical dual formula (1.3.5). Values are taken at a logarithmic scale $\log(1+x)$.

2.7.2 The two-component Weibull mixture model revisited

We consider the Weibull mixture (2.6.6) with shapes $\phi_1 = 0.5, \phi_2 = 3$ and $\lambda = 0.35$ which are supposed to be unknown during the estimation procedure. Here, we denote $\phi = (\phi_1, \phi_2)$ ($\alpha = (\alpha_1, \alpha_2)$, respectively) the shapes of the Weibull mixture model $p_{(\lambda,\phi)}$ ($p_{(\lambda,\alpha)}$, respectively). We used the same data simulated in paragraph 1.7.4, so that contamination was done by replacing 10 observations of each sample chosen randomly by 10 i.i.d. observations drawn from a Weibull distribution with shape $\nu = 0.9$ and scale $\sigma = 3$. Results are presented in tables (2.5) and (2.6).
Manipulating the optimization procedure for the Neymann's χ^2 was difficult because of the numerical integration calculus and the fact that for a subset of Φ (or $\Phi \times \Phi$ according to whether we use the estimator (1.3.5) or the estimator (1.5.3)) the integral term produces infinity, see paragraph 2.6.2 for more details. We therefore needed to keep the optimization from approaching the border in order to avoid numerical problems. For the Hellinger divergence, there is no particular remark.
For the case of the estimated divergence (1.3.5), if $\gamma = -1$, i.e. the Neymann χ^2, we need that $\alpha_1 < \phi_1/2$, otherwise the integral term is equal to infinity. In order to avoid numerical complications, we optimized over $\alpha_1 \leq 0.05 + \phi_1/2$. The value 0.05 ensures a small deviation from the border.
For the case of the estimated divergence (1.5.3), we used a Gaussian kernel for the Hellinger divergence. For the Neymann's χ^2 divergence, we used the Epanechnikov's kernel to avoid problems at infinity. Besides, it permits to integrate only over $[0, \max(Y) + w]$, where w is the window of the kernel, instead of $[0, \infty)$. In order to avoid problems near zero, it is necessary that $\min(\phi_1, \phi_2) < 1 - \frac{1}{\gamma} = 2$.

Comments on the tables: Experimental results show a clear robustness of the estimators calculated using the kernel-based MDφDE in comparison to other estimators using

the Hellinger divergence. When we are under the model, all estimation methods have the same performance. On the other hand, using the Neymann χ^2 divergence, results are different in the presence of outliers. The classical MDφDE calculated using formula (1.3.5) shows better robustness than other estimators, but is still not as good as the robustness of the kernel-based MDφDE using the Hellinger. Lack of robustness of the kernel-based MDφDE is not very surprising since the influence function of the kernel-based MDφDE is unbounded when we use the Neymann χ^2 divergence in simple models such as the Gaussian model, see Example 1.6.2.
In what concerns the proximal algorithm, there is no significant difference between the results obtained using the 1-step algorithm (2.1.8) and the ones obtained using the 2-step algorithm (2.2.2,2.2.3) using the Hellinger divergence. Differences appear when we used the Neymann χ^2 divergence with the classical MDφDE. This shows again the difficulty in handling the supermal form of the dual formal (1.3.5). Finally, in comparison to the results obtained with a direct optimization in paragraph 1.7.4, there is no significant differences.
The proximal-point algorithm worked as well using the density power divergence. The MDPD produced robust and efficient estimates which are slightly better than the results obtained by the kernel-based MDφDE using the Hellinger divergence and clearly better than the results obtained using the Neymann χ^2.

Estimation method		Error criterion TVD
Neymann Chi square		
Algorithm (2.1.8)	MDφDE	0.114 , sd = 0.032
	kernel-based MDφDE	0.057, sd = 0.028
Algorithm (2.2.2,2.2.3)	MDφDE	0.131, sd = 0.042
	kernel-based MDφDE	0.056, sd = 0.026
Hellinger		
Algorithm (2.1.8)	MDφDE	0.059, sd = 0.024
	kernel-based MDφDE	0.057, sd = 0.029
Algorithm (2.2.2,2.2.3)	MDφDE	0.061, sd = 0.026
	kernel-based MDφDE	0.057, sd = 0.029
MDPD $a = 0.5$ - Algorithm (2.1.8)		0.056, sd = 0.029
MDPD $a = 0.5$ - Algorithm (2.2.2,2.2.3)		0.056, sd = 0.029
EM		0.059, sd = 0.024

Table 2.5: The mean value of errors committed in a 100-run experiment of a two-component Weibull mixture with the standard deviation. No outliers are considered. The divergence criterion is the Neymann's χ^2 divergence or the Hellinger. The proximal term is taken with $\psi(t) = \frac{1}{2}(\sqrt{t}-1)^2$.

Estimation method		Error criterion TVD
Neymann Chi square		
Algorithm (2.1.8)	MDφDE	0.085, sd = 0.036
	kernel-based MDφDE	0.138, sd = 0.066
Algorithm (2.2.2,2.2.3)	MDφDE	0.096, sd = 0.057
	kernel-based MDφDE	0.127, sd = 0.056
Hellinger		
Algorithm (2.1.8)	MDφDE	0.120, sd = 0.034
	kernel-based MDφDE	0.068, sd = 0.034
Algorithm (2.2.2,2.2.3)	MDφDE	0.121, sd = 0.034
	kernel-based MDφDE	0.068, sd = 0.034
MDPD $a = 0.5$ - Algorithm (2.1.8)		0.060, sd = 0.029
MDPD $a = 0.5$ - Algorithm (2.2.2,2.2.3)		0.061, sd = 0.029
EM		0.129, sd = 0.046

Table 2.6: The mean value of errors committed in a 100-run experiment of a two-component Weibull mixture with the standard deviation. 10% outliers are considered. The divergence criterion is the Neymann's χ^2 divergence or the Hellinger. The proximal term is taken with $\psi(t) = \frac{1}{2}(\sqrt{t} - 1)^2$.

2.8 Conclusions

We presented in this chapter a proximal-point algorithm whose objective was the minimization of (an estimate of) a $\varphi-$divergence. The set of algorithms proposed here contains by construction the EM algorithm. We provided in several examples a proof of convergence of the EM algorithm in the spirit of our approach. We also showed how we may prove convergence for the two estimates of the $\varphi-$divergence (1.3.5) and (1.5.3). We reestablished similar results to the ones in Tseng [2004] in the context of $\varphi-$divergences, and provided a new result by relaxing the identifiability condition on the proximal term. Although our simulations do not permit to confirm the practical gain in comparison to direct methods, they are sufficient to conclude that the proximal algorithm works. The two-step algorithm (2.2.2,2.2.3) showed only slight deterioration in performance comparing to the original one (2.1.8) which is very encouraging especially that the dimension of the optimization is reduced at each step. Simulations have shown again the robustness of $\varphi-$divergences against outliers in comparison to the MLE.

2.9 Appendix: Proofs

2.9.1 Proof of Proposition 2.4.1

Proof. We prove (a). **For the first algorithm** defined by (2.1.8), we have by definition of the arginf:

$$\hat{D}_\varphi(p_{\phi^{k+1}}, p_{\phi_T}) + D_\psi(\phi^{k+1}, \phi^k) \leq \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}) + D_\psi(\phi^k, \phi^k).$$

We use the fact that $D_\psi(\phi^k, \phi^k) = 0$ for the right hand and that $D_\psi(\phi^{k+1}, \phi^k) \geq 0$ for the left hand side of the previous inequality. Hence $\hat{D}_\varphi(p_{\phi^{k+1}}, p_{\phi_T}) \leq \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T})$.
For the simplified algorithm defined by (2.2.2, 2.2.3), recurrence (2.2.2) and the definition of the arginf give:

$$\begin{aligned} \hat{D}_\varphi(p_{\lambda^{k+1},\theta^k}, p_{\phi_T}) + D_\psi((\lambda^{k+1}, \theta^k), \phi^k) &\leq \hat{D}_\varphi(p_{\lambda^k,\theta^k}, p_{\phi_T}) + D_\psi((\lambda^k, \theta^k), \phi^k) \\ &\leq \hat{D}_\varphi(p_{\lambda^k,\theta^k}, p_{\phi_T}). \end{aligned} \tag{2.9.1}$$

The second inequality is obtained using the fact that $D_\psi(\phi, \phi) = 0$. Using recurrence (2.2.3), we get:

$$\begin{aligned} \hat{D}_\varphi(p_{\lambda^{k+1},\theta^k}, p_{\phi_T}) + D_\psi((\lambda^{k+1}, \theta^k), \phi^k) &\geq \hat{D}_\varphi(p_{\lambda^{k+1},\theta^{k+1}}, p_{\phi_T}) + D_\psi((\lambda^{k+1}, \theta^{k+1}), \phi^k) \quad (2.9.2) \\ &\geq \hat{D}_\varphi(p_{\lambda^{k+1},\theta^{k+1}}, p_{\phi_T}). \quad (2.9.3) \end{aligned}$$

The second inequality is obtained using the fact that $D(\phi|\phi') \geq 0$. The conclusion is reached by combining the two inequalities (2.9.1) and (2.9.3).
We prove (b). Using the decreasing property previously proved in (a), we have by recurrence $\forall k, \hat{D}_\varphi(p_{\phi^{k+1}}, p_{\phi_T}) \leq \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}) \leq \cdots \leq \hat{D}_\varphi(p_{\phi^0}, p_{\phi_T})$. The result follows for both algorithms directly by definition of Φ^0.
We prove (c). By induction on k. For $k = 0$, clearly $\phi^0 = (\lambda^0, \theta^0)$ is well defined (a choice we make[19]). Suppose for some $k \geq 0$ that $\phi^k = (\lambda^k, \theta^k)$ exists. **For the first algorithm** defined by (2.1.8), we prove that the infimum is attained in Φ^0. Let $\phi \in \Phi$ be any vector

[19] The choice of the initial point of the sequence may influence the convergence of the sequence. See the example of the Gaussian mixture in paragraph (2.6.1).

at which the value of the optimized function has a value less than its value at ϕ^k, i.e. $\hat{D}_\varphi(p_\phi, p_{\phi_T}) + D_\psi(\phi, \phi^k) \leq \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}) + D_\psi(\phi^k, \phi^k)$. We have:

$$\begin{aligned}\hat{D}_\varphi(p_\phi, p_{\phi_T}) &\leq \hat{D}_\varphi(p_\phi, p_{\phi_T}) + D_\psi(\phi, \phi^k)\\ &\leq \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}) + D_\psi(\phi^k, \phi^k)\\ &\leq \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T})\\ &\leq \hat{D}_\varphi(p_{\phi^0}, p_{\phi_T}).\end{aligned}$$

The first line follows from the non negativity of D_ψ. As $\hat{D}_\varphi(p_\phi, p_{\phi_T}) \leq \hat{D}_\varphi(p_{\phi^0}, p_{\phi_T})$, then $\phi \in \Phi^0$. Thus, the infimum can be calculated for vectors in Φ^0 instead of Φ. Since Φ^0 is compact and the optimized function is lower semicontinuous (the sum of two lower semicontinuous functions), then the infimum exists and is attained in Φ^0. We may now define ϕ^{k+1} to be a vector whose corresponding value is equal to the infimum.

For the second algorithm defined by (2.2.2,2.2.3). Similarly, the infimum in (2.2.2) can be calculated on λ's such that $(\lambda, \theta^k) \in \Phi^0$. Indeed, suppose there exists a λ at which the value of the optimized function is less than its value at λ^k, i.e. $\hat{D}_\varphi(p_{\lambda,\theta^k}, p_{\phi_T}) + D_\psi((\lambda, \theta^k), \phi^k) \leq \hat{D}_\varphi(p_{\lambda^k,\theta^k}, p_{\phi_T}) + D_\psi((\lambda^k, \theta^k), \phi^k)$. We have:

$$\begin{aligned}\hat{D}_\varphi(p_{\lambda,\theta^k}, p_{\phi_T}) &\leq \hat{D}_\varphi(p_{\lambda,\theta^k}, p_{\phi_T}) + D_\psi((\lambda, \theta^k), \phi^k)\\ &\leq \hat{D}_\varphi(p_{\lambda^k,\theta^k}, p_{\phi_T}) + D_\psi((\lambda^k, \theta^k), \phi^k)\\ &\leq \hat{D}_\varphi(p_{\lambda^k,\theta^k}, p_{\phi_T})\\ &\leq \hat{D}_\varphi(p_{\phi^0}, p_{\phi_T}).\end{aligned}$$

This means that $(\lambda, \theta^k) \in \Phi^0$ and that the infimum needs not to be calculated for all values of λ in Φ, and can be restrained onto values which verify $(\lambda, \theta^k) \in \Phi^0$.

Define now $\Lambda_k = \{\lambda \in [0,1]^s | (\lambda, \theta^k) \in \Phi^0\}$. First of all, $\lambda^k \in \Lambda_k$ since $(\lambda^k, \theta^k) \in \Phi^0$. Therefore, Λ_k is not empty. Moreover, it is compact. Indeed, let $(\lambda^l)_l$ be a sequence of elements of Λ_k, then the sequence $((\lambda^l, \theta^k))_l$ is a sequence of elements of Φ^0. By compactness of Φ^0, there exists a subsequence which converges in Φ^0 to an element of the form $(\lambda^\infty, \theta^k)$ which clearly belongs to Λ_k. This proves that Λ_k is compact. Finally, since by assumption A0, the optimized function is lower semicontinuous so that it attains its infimum on the compact set Λ_k. We may now define λ^{k+1} as any vector verifying this infimum.

The second part of the proof treats the definition of θ^{k+1}. Let θ be any vector such that $(\lambda^{k+1}, \theta) \in \Phi$ and at which the value of the optimized function in (2.2.3) is less than its value at ϕ^k. We have

$$\begin{aligned}\hat{D}_\varphi(p_{\lambda^{k+1},\theta}, p_{\phi_T}) &\leq \hat{D}_\varphi(p_{\lambda^{k+1},\theta}, p_{\phi_T}) + D_\psi((\lambda^{k+1}, \theta), \phi^k)\\ &\leq \hat{D}_\varphi(p_{\lambda^{k+1},\theta^k}, p_{\phi_T}) + D_\psi((\lambda^{k+1}, \theta^k), \phi^k)\\ &\leq \hat{D}_\varphi(p_{\lambda^k,\theta^k}, p_{\phi_T}) + D_\psi((\lambda^k, \theta^k), \phi^k)\\ &\leq \hat{D}_\varphi(p_{\lambda^k,\theta^k}, p_{\phi_T})\\ &\leq \hat{D}_\varphi(p_{\phi^0}, p_{\phi_T})\end{aligned}$$

The third line comes from the previous definition of λ^{k+1} as an infimum of (2.2.2). This means that $(\lambda^{k+1}, \theta) \in \Phi^0$, and that the infimum in (2.2.3) can be calculated with respect to values θ which verifies $(\theta, \lambda^{k+1}) \in \Phi^0$. Define now $\Theta_k = \{\theta \in \mathbb{R}^{d-s} | (\lambda^{k+1}, \theta) \in \Phi^0\}$. One can prove analogously to Λ_k, that it is compact. The optimized function in (2.2.3) is,

by assumption A0, lower semicontinuous so that its infimum is attained on the compact Θ_k. We may now define θ^{k+1} as any vector verifying this infimum.
Convergence of the sequence $(\hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}))_k$ in both algorithms comes from the fact that it is non increasing and bounded. It is non increasing by virtue of (a). Boundedness comes from the lower semicontinuity of $\phi \mapsto \hat{D}_\varphi(p_\phi, p_{\phi_T})$. Indeed, $\forall k, \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}) \geq \inf_{\phi\in\Phi^0} \hat{D}_\varphi(p_\phi, p_{\phi_T})$. The infimum of a proper lower semicontinuous function on a compact set exists and is attained on this set. Hence, the quantity $\inf_{\phi\in\Phi^0} \hat{D}_\varphi(p_\phi, p_{\phi_T})$ exists and is finite. This ends the proof. □

2.9.2 Proof of Proposition 2.4.2

Proof. We prove (a). Let $(\phi^{n_k})_k$ be a convergent subsequence of $(\phi^k)_k$ which converges to ϕ^∞. First, $\phi^\infty \in \Phi^0$, because Φ^0 is closed and the subsequence (ϕ^{n_k}) is a sequence of elements of Φ^0 (proved in Proposition 2.4.1.b).
Let's show now that the subsequence (ϕ^{n_k+1}) also converges to ϕ^∞. We simply have:

$$\|\phi^{n_k+1} - \phi^\infty\| \leq \|\phi^{n_k} - \phi^\infty\| + \|\phi^{n_k+1} - \phi^{n_k}\|$$

Since $\phi^{k+1} - \phi^k \to 0$ and $\phi^{n_k} \to \phi^\infty$, we conclude that $\phi^{n_k+1} \to \phi^\infty$.
Let's start with the first algorithm (2.1.8). By definition of ϕ^{n_k+1}, it verifies the infimum in recurrence (2.1.8), so that the gradient of the optimized function is zero:

$$\nabla\hat{D}_\varphi(p_{\phi^{n_k+1}}, p_{\phi_T}) + \nabla D_\psi(\phi^{n_k+1}, \phi^{n_k}) = 0$$

Using the continuity assumptions A1 and AC of the gradients, one can pass to the limit with no problem:

$$\nabla\hat{D}_\varphi(p_{\phi^\infty}, p_{\phi_T}) + \nabla D_\psi(\phi^\infty, \phi^\infty) = 0$$

However, the gradient $\nabla D_\psi(\phi^\infty, \phi^\infty) = 0$ because (recall that $\psi'(1) = 0$):

$$\nabla D_\psi(\phi, \phi) = \sum_{i=1}^n \int_{\mathcal{X}} \frac{\nabla h_i(x|\phi)}{h_i(x|\phi)} \psi'\left(\frac{h_i(x|\phi)}{h_i(x|\phi)}\right) h_i(x|\phi)dx = \sum_{i=1}^n \int_{\mathcal{X}} \nabla h_i(x|\phi)\psi'(1)dx$$

Hence the gradient $\nabla D_\psi(\phi, \phi) = 0$. This implies that $\nabla\hat{D}_\varphi(p_{\phi^\infty}, p_{\phi_T}) = 0$.
For the second algorithm (2.2.2,2.2.3), by definition of λ^{n_k+1} and θ^{n_k+1}, they verify the infimum respectively in recurrences (2.2.2) and (2.2.3). Therefore, the gradient of the optimized function is zero for each step. In other words:

$$\begin{aligned}
\nabla_\lambda\hat{D}_\varphi(p_{\lambda^{n_k+1},\theta^{n_k}}, p_{\phi_T}) + \nabla_\lambda D_\psi((\lambda^{n_k+1}, \theta^{n_k}), \phi^{n_k}) &= 0 \\
\nabla_\theta\hat{D}_\varphi(p_{\lambda^{n_k+1},\theta^{n_k+1}}, p_{\phi_T}) + \nabla_\theta D_\psi((\lambda^{n_k+1}, \theta^{n_k+1}), \phi^{n_k}) &= 0
\end{aligned}$$

Since both (ϕ^{n_k+1}) and (ϕ^{n_k}) converge to the same limit ϕ^∞, then setting $\phi^\infty = (\lambda^\infty, \theta^\infty)$, we get λ^{n_k+1} and λ^{n_k} tends to λ^∞. We also have θ^{n_k+1} and θ^{n_k} tends to θ^∞. The continuity of the two gradients (assumptions A1 and AC) implies that:

$$\begin{aligned}
\nabla_\lambda\hat{D}_\varphi(p_{\lambda^\infty,\theta^\infty}, p_{\phi_T}) + \nabla_\lambda D_\psi((\lambda^\infty, \theta^\infty), \phi^\infty) &= 0 \\
\nabla_\theta\hat{D}_\varphi(p_{\lambda^\infty,\theta^\infty}, p_{\phi_T}) + \nabla_\theta D_\psi((\lambda^\infty, \theta^\infty), \phi^\infty) &= 0
\end{aligned}$$

However, $\nabla D_\psi(\phi, \phi) = 0$, so that $\nabla_\lambda\hat{D}_\varphi(p_{\phi^\infty}, p_{\phi_T}) = 0$ and $\nabla_\theta\hat{D}_\varphi(p_{\phi^\infty}, p_{\phi_T}) = 0$. Hence $\nabla\hat{D}_\varphi(p_{\phi^\infty}, p_{\phi_T}) = 0$.
We prove (b). For the first algorithm, we use again the definition of the arginf. As the

optimized function is not necessarily differentiable at the points of the sequence ϕ^k, a necessary condition for ϕ^{k+1} to be an infimum is that 0 belongs to the subgradient of the function on ϕ^{k+1}. Since $D_\psi(\phi, \phi^k)$ is assumed to be differentiable, the optimality condition is translated into:

$$-\nabla D_\psi(\phi^{k+1}, \phi^k) \in \partial \hat{D}_\varphi(p_{\phi^{k+1}}, p_{\phi_T}) \quad \forall k$$

Since $\hat{D}_\varphi(p_\phi, p_{\phi_T})$ is continuous, then its subgradient is outer semicontinuous (see [Rockafellar and Wets, 1998] Chap 8, proposition 7). We use the same arguments presented in (a) to conclude the existence of two subsequences $(\phi^{n_k})_k$ and $(\phi^{n_k+1})_k$ which converge to the same limit ϕ^∞. By definition of outer semicontinuity, and since $\phi^{n_k+1} \to \phi^\infty$, we have:

$$\limsup_{\phi^{n_k+1} \to \phi^\infty} \partial \hat{D}_\varphi(p_{\phi^{n_k+1}}, p_{\phi_T}) \subset \partial \hat{D}_\varphi(p_{\phi^\infty}, p_{\phi_T}) \tag{2.9.4}$$

We want to prove that $0 \in \limsup_{\phi^{n_k+1} \to \phi^\infty} \partial \hat{D}_\varphi(p_{\phi^{n_k+1}}, p_{\phi_T})$. By definition of limsup[20]:

$$\limsup_{\phi \to \phi^\infty} \partial \hat{D}_\varphi(p_\phi, p_{\phi_T}) = \left\{ u | \exists \phi^k \to \phi^\infty, \exists u^k \to u \text{ with } u^k \in \partial \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}) \right\}$$

In our scenario, $\phi = \phi^{n_k+1}$, $\phi^k = \phi^{n_k+1}$, $u = 0$ and $u^k = \nabla_1 D_\psi(\phi^{n_k+1}, \phi^{n_k})$. The continuity of $\nabla_1 D_\psi$ with respect to both arguments and the fact that the two subsequences ϕ^{n_k+1} and ϕ^{n_k} converge to the same limit, imply that $u^k \to \nabla_1 D_\psi(\phi^\infty, \phi^\infty) = 0$. Hence $u = 0 \in \limsup_{\phi^{n_k+1} \to \phi^\infty} \partial \hat{D}_\varphi(p_{\phi^{n_k+1}}, p_{\phi_T})$. By inclusion (2.9.4), we get our result:

$$0 \in \partial \hat{D}_\varphi(p_{\phi^\infty}, p_{\phi_T})$$

□

2.9.3 Proof of Proposition 2.4.3

Proof. The arguments presented are the same for both algorithms (2.1.8) and (2.2.2,2.2.3). By contradiction, let's suppose that $\phi^{k+1} - \phi^k$ does not converge to 0. There exists a subsequence such that $\|\phi^{N_0(k)+1} - \phi^{N_0(k)}\| > \varepsilon$, $\forall k \geq k_0$. Since $(\phi^k)_k$ belongs to the compact set Φ^0, there exists a convergent subsequence $(\phi^{N_1 \circ N_0(k)})_k$ such that $\phi^{N_1 \circ N_0(k)} \to \bar{\phi}$. The sequence $(\phi^{N_1 \circ N_0(k)+1})_k$ belongs to the compact set Φ^0, therefore we can extract a further subsequence $(\phi^{N_2 \circ N_1 \circ N_0(k)+1})_k$ such that $\phi^{N_2 \circ N_1 \circ N_0(k)+1} \to \tilde{\phi}$. Besides $\hat{\phi} \neq \tilde{\phi}$. Finally since the sequence $(\phi^{N_1 \circ N_0(k)})_k$ is convergent, a further subsequence also converges to the same limit $\bar{\phi}$. We have proved the existence of a subsequence of $(\phi^k)_k$ such that $\phi^{N(k)+1} - \phi^{N(k)}$ does not converge to 0 and such that $\phi^{N(k)+1} \to \tilde{\phi}$, $\phi^{N(k)} \to \bar{\phi}$ with $\bar{\phi} \neq \tilde{\phi}$. The real sequence $\hat{D}_\varphi(p_{\phi^k}, p_{\phi_T})_k$ converges as proved in Proposition 2.4.1-c. As a result, both sequences $\hat{D}_\varphi(p_{\phi^{N(k)+1}}, p_{\phi_T})$ and $\hat{D}_\varphi(p_{\phi^{N(k)}}, p_{\phi_T})$ converge to the same limit being subsequences of the same convergent sequence. In the proof of Proposition 2.4.1, we can deduce the following inequality:

$$\hat{D}(p_{\lambda^{k+1}, \theta^{k+1}}, p_{\phi_T}) + D_\psi((\lambda^{k+1}, \theta^{k+1}), \phi^k) \leq \hat{D}(p_{\lambda^k, \theta^k}, p_{\phi_T}) \tag{2.9.5}$$

which is also verified to any substitution of k by $N(k)$. By passing to the limit on k, we get $D_\psi(\tilde{\phi}, \bar{\phi}) \leq 0$. However, the distance-like function D_ψ is positive, so that it becomes zero. Using assumption A3, $D_\psi(\tilde{\phi}, \bar{\phi}) = 0$ implies that $\tilde{\phi} = \bar{\phi}$. This contradicts the hypothesis that $\phi^{k+1} - \phi^k$ does not converge to 0.
The second part of the proposition is a direct result of Proposition 2.4.2. □

[20] We use here the definition corresponding to the outer limit, see [Rockafellar and Wets, 1998] Chap 4, definition 1 or Chap 5-B.

2.9.4 Proof of Corollary 2.4.1

Proof. Since the sequence $(\phi)_k$ is bounded and verifies $\phi^{k+1} - \phi^k \to 0$, then Theorem 28.1 in [Ostrowski, 1966] implies that the set of accumulation points of $(\phi^k)_k$ is a connected compact set. It is not empty since Φ^0 is compact. Let ϕ^∞ be a limit point of $(\phi^k)_k$. The assumption about strict convexity of $\hat{D}(p_\phi, p_{\phi_T})$ in a neighborhood of ϕ^∞ implies that it is isolated in the sense that if there are another limit point $\tilde{\phi}$, then there is $\varepsilon > 0$ such that $\|\phi^\infty - \tilde{\phi}\| > \varepsilon$. Hence, the set of accumulation points can be written as the union of at least two disjoint open sets which contradicts the connectedness property. Thus, ϕ^∞ is the only limit point of the sequence (ϕ^k). To end the proof, we need to show that the whole sequence converge. By contradiction, if it does not converge, there exists then $\varepsilon > 0$ and an infinity of terms which verifies $\|\phi^{N_0(k)} - \phi^\infty\| > \varepsilon$. By compactness of Φ^0, one may extract a subsequence of $(\phi^{N_0(k)})_k$, say $(\phi^{N_1 \circ N_0(k)})_k$, which converges to some $\hat{\phi}$. Moreover, by continuity of the euclidean norm, $\|\phi^{N_1 \circ N_0(k)} - \phi^\infty\| \to \|\hat{\phi} - \phi^\infty\|$. Hence $\|\hat{\phi} - \phi^\infty\| \geq \varepsilon$. Contradiction is reached by uniqueness of the limit point of the sequence $(\phi^k)_k$. □

2.9.5 Proof of Proposition 2.4.4

Proof. If $(\phi^k)_k$ converges to, say, ϕ^∞, the result falls simply from Proposition 2.
If $(\phi^k)_k$ does not converge. Since Φ^0 is compact and $\forall k, \phi^k \in \Phi^0$ (proved in Proposition 1), there exists a subsequence $(\phi^{N_0(k)})_k$ such that $\phi^{N_0(k)} \to \tilde{\phi}$. Let's take the subsequence $(\phi^{N_0(k)-1})_k$. This subsequence does not necessarily converge; still it is contained in the compact Φ^0, so that we can extract a further subsequence $(\phi^{N_1 \circ N_0(k)-1})_k$ which converges to, say, $\bar{\phi}$. Now, the subsequence $(\phi^{N_1 \circ N_0(k)})_k$ converges to $\tilde{\phi}$, because it is a subsequence of $(\phi^{N_0(k)})_k$. We have proved until now the existence of two convergent subsequences $\phi^{N(k)-1}$ and $\phi^{N(k)}$ with *a priori* different limits. For simplicity and without any loss of generality, we will consider these subsequences to be ϕ^k and ϕ^{k+1} respectively.
Conserving previous notations, suppose that $\phi^{k+1} \to \tilde{\phi}$ and $\phi^k \to \bar{\phi}$. We use again inequality (2.9.5):

$$\hat{D}(p_{\phi^{k+1}}, p_{\phi_T}) + D_\psi(\phi^{k+1}, \phi^k) \leq \hat{D}(p_{\lambda^k, \theta^k}, p_{\phi_T})$$

By taking the limits of the two parts of the inequality as k tends to infinity, and using the continuity of the two functions, we have

$$\hat{D}(p_{\tilde{\phi}}, p_{\phi_T}) + D_\psi(\tilde{\phi}, \bar{\phi}) \leq \hat{D}(p_{\bar{\phi}}, p_{\phi_T})$$

Recall that under A1-2, the sequence $\left(\hat{D}_\varphi(p_{\phi^k}, p_{\phi_T})\right)_k$ converges, so that it has the same limit for any subsequence, i.e. $\hat{D}(p_{\tilde{\phi}}, p_{\phi_T}) = \hat{D}(p_{\bar{\phi}}, p_{\phi_T})$. We also use the fact that the distance-like function D_ψ is non negative to deduce that $D_\psi(\tilde{\phi}, \bar{\phi}) = 0$. Looking closely at the definition of this divergence (2.1.7), we get that if the sum is zero, then each term is also zero since all terms are non negative. This means that:

$$\forall i \in \{1, \cdots, n\}, \quad \int_{\mathcal{X}} \psi\left(\frac{h_i(x|\tilde{\phi})}{h_i(x|\bar{\phi})}\right) h_i(x|\bar{\phi}) dx = 0$$

The integrands are non negative functions, so they vanish almost ever where with respect to the measure dx defined on the space of labels.

$$\forall i \in \{1, \cdots, n\}, \quad \psi\left(\frac{h_i(x|\tilde{\phi})}{h_i(x|\bar{\phi})}\right) h_i(x|\bar{\phi}) = 0 \quad dx - a.e.$$

The conditional densities h_i are supposed to be positive[21], i.e. $h_i(x|\bar{\phi}) > 0, dx - a.e.$. Hence, $\psi\left(\frac{h_i(x|\tilde{\phi})}{h_i(x|\bar{\phi})}\right) = 0, dx - a.e.$. On the other hand, ψ is chosen in a way that $\psi(z) = 0$ iff $z = 1$, therefore :

$$\forall i \in \{1, \cdots, n\}, \quad h_i(x|\tilde{\phi}) = h_i(x|\bar{\phi}) \quad dx - a.e. \tag{2.9.6}$$

Since ϕ^{k+1} is, by definition, an infimum of $\phi \mapsto \hat{D}(p_\phi, p_{\phi_T}) + D_\psi(\phi, \phi^k)$, then the gradient of this function is zero on ϕ^{k+1}. It results that:

$$\nabla \hat{D}(p_{\phi^{k+1}}, p_{\phi_T}) + \nabla D_\psi(\phi^{k+1}, \phi^k) = 0, \quad \forall k$$

Taking the limit on k, and using the continuity of the derivatives, we get that:

$$\nabla \hat{D}(p_{\tilde{\phi}}, p_{\phi_T}) + \nabla D_\psi(\tilde{\phi}, \bar{\phi}) = 0 \tag{2.9.7}$$

Let's write explicitly the gradient of the second divergence:

$$\nabla D_\psi(\tilde{\phi}, \bar{\phi}) = \sum_{i=1}^{n} \int_{\mathcal{X}} \frac{\nabla h_i(x|\tilde{\phi})}{h_i(x|\bar{\phi})} \psi'\left(\frac{h_i(x|\tilde{\phi})}{h_i(x|\bar{\phi})}\right) h_i(x|\bar{\phi})$$

We use now the identities (2.9.6), and the fact that $\psi'(1) = 0$, to deduce that:

$$\nabla D_\psi(\tilde{\phi}, \bar{\phi}) = 0$$

This entails using (2.9.7) that $\nabla \hat{D}(p_{\tilde{\phi}}, p_{\phi_T}) = 0$.
Comparing the proved result with the notation considered at the beginning of the proof, we have proved that the limit of the subsequence $(\phi^{N_1 \circ N_0(k)})_k$ is a stationary point of the objective function. Therefore, The final step is to deduce the same result on the original convergent subsequence $(\phi^{N_0(k)})_k$. This is simply due to the fact that $(\phi^{N_1 \circ N_0(k)})_k$ is a subsequence of the convergent sequence $(\phi^{N_0(k)})_k$, hence they have the same limit.
When assumption AC is dropped, similar arguments to those used in the proof of Proposition 2-b. are employed. The optimality condition in (2.1.8) implies :

$$-\nabla D_\psi(\phi^{k+1}, \phi^k) \in \partial \hat{D}_\varphi(p_{\phi^{k+1}}, p_{\phi_T}) \quad \forall k$$

Function $\phi \mapsto \hat{D}_\varphi(p_\phi, p_{\phi_T})$ is continuous, hence its subgradient is outer semicontinuous and:

$$\limsup_{\phi^{k+1} \to \phi^\infty} \partial \hat{D}_\varphi(p_{\phi^{k+1}}, p_{\phi_T}) \subset \partial \hat{D}_\varphi(p_{\tilde{\phi}}, p_{\phi_T}) \tag{2.9.8}$$

By definition of limsup:

$$\limsup_{\phi \to \phi^\infty} \partial \hat{D}_\varphi(p_\phi, p_{\phi_T}) = \left\{ u | \exists \phi^k \to \phi^\infty, \exists u^k \to u \text{ with } u^k \in \partial \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}) \right\}$$

In our scenario, $\phi = \phi^{k+1}$, $\phi^k = \phi^{k+1}$, $u = 0$ and $u^k = \nabla_1 D_\psi(\phi^{k+1}, \phi^k)$. We have proved above in this proof that $\nabla_1 D_\psi(\tilde{\phi}, \bar{\phi}) = 0$ using only convergence of $(\hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}))_k$, inequality (2.9.5) and some properties of D_ψ. Assumption AC was not needed. Hence, $u^k \to 0$. This proves that, $u = 0 \in \limsup_{\phi^{k+1} \to \phi^\infty} \partial \hat{D}_\varphi(p_{\phi^{n_k+1}}, p_{\phi_T})$. Finally, using the inclusion (2.9.8), we get our result:

$$0 \in \partial \hat{D}_\varphi(p_{\tilde{\phi}}, p_{\phi_T})$$

□

[21] In the case of two Gaussian (or more generally exponential) components, this is justified by virtue of a suitable choice of the initial condition.

2.9.6 Proof of Proposition 2.4.5

Proof. We use the same lines from the previous proof to deduce the existence of two convergent subsequences $\phi^{N(k)-1}$ and $\phi^{N(k)}$ with *a priori* different limits. For simplicity and without any loss of generality, we will consider these subsequences to be ϕ^k and ϕ^{k+1} respectively. Suppose that $\phi^k \to \bar{\phi} = (\bar{\lambda}, \bar{\theta})$ and $\phi^{k+1} \to \tilde{\phi} = (\tilde{\lambda}, \tilde{\theta})$.
We first use inequality (2.9.5) as in the previous proposition, the convergence of the sequence $(\hat{D}_\varphi(p_{\lambda^k,\theta^k}, p_{\phi_T}))_k$ and some basic properties of D_ψ to deduce that:

$$\forall i \in \{1, \cdots, n\}, \quad h_i(x|\tilde{\phi}) = h_i(x|\bar{\phi}) \quad dx - a.e. \tag{2.9.9}$$

Let's calculate the gradient of the objective function with respect to λ and θ separately at the limit of $(\phi^{k+1})_k$. By definition of θ^{k+1} as an arginf in (2.2.3), we have:

$$\frac{\partial}{\partial\theta}\hat{D}_\varphi(p_{\lambda^{k+1},\theta^{k+1}}, p_{\phi_T}) + \frac{\partial}{\partial\theta}D_\psi((\lambda^{k+1}, \theta^{k+1}), \phi^k) = 0 \quad \forall k$$

Using the continuity of the derivatives (Assumptions A1 and AC), we may pass to the limit inside the gradients:

$$\frac{\partial}{\partial\theta}\hat{D}_\varphi(p_{\tilde{\lambda},\tilde{\theta}}, p_{\phi_T}) + \frac{\partial}{\partial\theta}D_\psi((\tilde{\lambda}, \tilde{\theta}), \bar{\phi}) = 0 \quad \forall k$$

As in the proof of Proposition 3, all terms in the gradient of D_ψ depend on $\psi'\left(\frac{h_i(x|\tilde{\lambda},\tilde{\theta})}{h_i(x|\bar{\phi})}\right)$ which is zero by virtue of (2.9.9). Hence $\frac{\partial}{\partial\theta}\hat{D}_\varphi(p_{\tilde{\lambda},\tilde{\theta}}, p_{\phi_T}) = 0$.
We prove now that $\frac{\partial}{\partial\lambda}\hat{D}_\varphi(p_{\tilde{\lambda},\tilde{\theta}}, p_{\phi_T}) = 0$. This is basically ensured by recurrence (2.2.2), identities (2.9.9), assumptions A1-AC and the fact that $\psi'(1) = 0$. Indeed, using recurrence (2.2.2), λ^{k+1} is an optimum so that the gradient of the objective function is zero:

$$\frac{\partial}{\partial\lambda}\hat{D}_\varphi(p_{\lambda^{k+1},\theta^k}, p_{\phi_T}) + \frac{\partial}{\partial\lambda}D_\psi((\lambda^{k+1}, \theta^k), \lambda^k, \theta^k) = 0, \quad \forall k$$

Since $\|\theta^{k+1} - \theta^k\| \to 0$, then $\bar{\theta} = \tilde{\theta}$. By passing to the limit in the previous identity and using the continuity of the derivatives, we have:

$$\frac{\partial}{\partial\lambda}\hat{D}_\varphi(p_{\tilde{\lambda},\bar{\theta}}, p_{\phi_T}) + \frac{\partial}{\partial\lambda}D_\psi((\tilde{\lambda}, \tilde{\theta}), \bar{\lambda}, \bar{\theta}) = 0$$

Since the derivative of D_ψ is a sum of terms which depend all on $\psi'(\frac{h_i(x|\tilde{\lambda},\bar{\theta})}{h_i(|\bar{\lambda},\bar{\theta})})$, and using identities (2.9.9), we conclude that $\psi'(\frac{h_i(|\tilde{\lambda},\bar{\theta})}{h_i(|\bar{\lambda},\bar{\theta})}) = \psi'(1) = 0$ and $\frac{\partial}{\partial\lambda}D_\psi((\tilde{\lambda}, \bar{\theta}), \bar{\lambda}, \bar{\theta}) = 0$. Finally, $\bar{\theta} = \tilde{\theta}$ implies that $\frac{\partial}{\partial\lambda}\hat{D}_\varphi(p_{\tilde{\lambda},\hat{\theta}}, p_{\phi_T}) = 0$.
We have proved that $\frac{\partial}{\partial\lambda}\hat{D}_\varphi(p_{\tilde{\lambda},\tilde{\theta}}, p_{\phi_T}) = 0$ and $\frac{\partial}{\partial\theta}\hat{D}_\varphi(p_{\tilde{\lambda},\tilde{\theta}}, p_{\phi_T}) = 0$, so the gradient is zero and the stated result is proved. □

Part II

Two-component Semiparametric Mixture Models When One Component is Unknown

Chapter 3

Semiparametric two-component mixture models where one component is defined through linear constraints on its distribution function

A two-component mixture model with an unknown component is defined by:

$$f(x) = \lambda f_1(x|\theta) + (1-\lambda) f_0(x), \qquad \text{for } x \in \mathbb{R}^r \tag{3.0.1}$$

for $\lambda \in (0,1)$ and $\theta \in \mathbb{R}^d$ to be estimated and the density f_0 is considered to be unknown. Such model appears in the study of gene expression data coming from microarray analysis. An application to two bovine gestation mode comparison is performed in Bordes et al. [2006]. The authors suppose that θ is known, f_0 is symmetric around an unknown μ and that $r = 1$. Xiang et al. [2014] studied a more general setup by considering θ unknown and applied model (3.0.1) on the Iris data by considering only the first principle component for each observed vector. Another application of model (3.0.1) in genetics can be found in Ma et al. [2011]. See also Patra and Sen [2016] for applications arising in astronomy and from microarray experiment.
Robin et al. [2007] used the semiparametric model (supposing that θ is known) in multiple testing procedures in order to estimate the posterior population probabilities and the local false rate discovery. Song et al. [2010] studied a similar setup where θ is unknown without further assumptions on f_0. They applied the semiparametric model in sequential clustering algorithms as a second step. After a sequential clustering algorithm finds the center of a cluster, the next step is to identify the observations belonging to this cluster. If we assume that the center of the cluster is known and that the distribution of observations not belonging to the cluster is unknown, the problem of identifying observations in the cluster is similar to the problem of estimating the mixing proportion in a special two-component mixture model. The mixing proportion can be considered as the proportion of observations belonging to the cluster. Finally, model (3.0.1) can also be regarded as a contamination model, see Titterington et al. [1985] or McLachlan and Peel [2005] for further applications of general mixture models.

Existing estimation methods for model (3.0.1) were proved or illustrated to work but only in specific situations and the only simulated example was a dataset generated by

a Gaussian mixture. The paper of Xiang et al. [2014] provides a comparison of several estimation methods for the semiparametric model. The compared methods give satisfactory results in most simulations, but no method performs uniformly good on all simulated mixtures. In all these simulations the authors consider θ to be given. We noticed that as we add θ to the set of unknown parameters, things become different. The performance of these methods depend on the proportion of the parametric component if it is high or low. They may have very poor performances sometimes even if the number of observations is high as we will demonstrate in the simulation section.

It is important for an estimation method to be applicable in contexts where the parametric component is not fully known. For example, the parametric component may be a signal whereas the unknown component is a noise. We need to extract the location of the signal or any other information concerning its shape and not only the proportion of the noise. We believe that the failure of the existing methods when the parametric component is not fully known comes from the degree of difficulty of the semiparametric model, i.e we do not possess sufficient information about the model in order to estimate it. Bordes and Vandekerkhove [2010] considered a symmetric assumption on the unknown component (see also Bordes et al. [2006]). This gave the model a structure and permitted to improve the estimation and made the study of the asymptotic properties of the resulting estimators tractable. Moreover, they were able to give sufficient conditions under which the semiparametric model is identifiable. Nevertheless, such assumption is very restrictive and cannot be applied for example in the context of distributions defined on a subset of $\mathbb{R}$. It appears that the addition of prior information should be helpful and may lead to a more understandable theory and better estimation results. We propose to add some information about f_0 in a way that we stay in between a (restrictive) fully parametric settings and a (complex) fully semiparametric one.

In this chapter, we introduce a method which permits to add a, relatively general, prior information about the unknown component in order to decrease the degree of difficulty of the model and be able to better estimate it. Such information needs to apply linearly on the distribution function of the unknown component such as moment-type information. For example, we may have an information relating the first and the second moments of f_0 such as $\int x f_0(x) = \alpha$ and $\int x^2 f_0(x)dx = m(\alpha)$, see Broniatowski and Keziou [2012] and the references therein. Such information adds some structure to the model without precising the value of the moments. More examples will be discussed later on.

Unfortunately, the incorporation of linear constraints on the distribution function cannot be done directly in existing methods because the optimization[1] will be carried over a (possibly) infinite dimensional space, and we need a new approach. Convex analysis offers a way using the Fenchel-Legendre duality to transform an optimization problem over an infinite dimensional space. On the other hand, $\varphi-$divergences offer a way by their convexity properties to exploit this duality result. The paper of Broniatowski and Keziou [2012] gives a complete study of this problem in the non mixture case, see also Decurninge [2015] Chap. 1 and Keziou [2003] Chap. 3. We will exploit these results to build upon a new estimation procedure which takes into account linear information over the unknown component's distribution.

[1]Not all existing methods, as we will see in the next paragraph, are defined through an optimization procedure. Hence, it becomes more difficult to introduce this kind of constraints inside the estimation procedure.

3.1 Semiparametric two-component mixture models in the literature

The literature on semiparametric mixture models contains several methods which permit to estimate efficiently the parameters with or without estimating the unknown component. All these methods were never tested on difficult situations except for datasets generated from a mixture of two Gaussian components with very close means (difference of means equal to 1.5) when the first component f_1 is *fully known*.
We present in this section the principle estimation methods in the literature. We present the method of Bordes and Vandekerkhove [2010] which is based on a symmetry constraint over the unknown component. We present also two EM-type algorithms introduced by Robin et al. [2007] and Song et al. [2010], and an SEM-type algorithm developed by Bordes et al. [2007]. There is also an interesting method developed in Song et al. [2010] based on the identifiability of a two-component mixture model when f_1 is Gaussian, called the $\pi-$maximizing algorithm. Finally, a method based on the Hellinger divergence was developed by Xiang et al. [2014]. However, the algorithm presented in their article is not clear and contains difficult integration calculus which cannot be calculated directly by a numerical method. The authors did not give any further explanations on how to do the calculus. We therefore prefer not to discuss it here.
We advise the reader to consult the simulation results in Xiang et al. [2014]. The article contains a comparison between some of these methods in a two-component Gaussian mixture. We will also be testing all these methods on further simulations and different models to explore their capacities in estimating the semiparametric mixture model.

3.1.1 Semiparametric two-component mixture models under a symmetry assumption

Bordes et al. [2006] proposed to study the semiparametric model (3.0.1) when $r = 1$, θ is given and the unknown component is supposed to be symmetric. Thus, the semiparametric model (3.0.1) can be rewritten as:

$$f(x|\lambda,\theta) = \lambda f_1(x|\theta) + (1-\lambda) f_0(x-\mu_0), \qquad \forall x \in \mathbb{R}. \tag{3.1.1}$$

Bordes et al. [2006] have studied the identifiability of this simplified model by imposing either a symmetry assumption over f_1, conditions over the characteristic function or conditions on the tail behavior of the components. Let us summarize their estimation procedure. Let $\mathbb{F}_0, \mathbb{F}_1$ and $\mathbb{F}$ be the cumulative distribution functions (cdf) of f_0, f_1 and f respectively. We have:

$$\mathbb{F}_0(x) = \frac{1}{1-\lambda}\mathbb{F}(x+\mu_0|\lambda,\theta) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(x+\mu_0|\theta).$$

Define the following functions:

$$\begin{aligned} H_1(x|\lambda,\theta,\mathbb{F}) &= \frac{1}{1-\lambda}\mathbb{F}(x+\mu_0|\lambda,\theta) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(x+\mu_0|\theta), \\ H_2(x|\lambda,\theta,\mathbb{F}) &= 1 - \frac{1}{1-\lambda}\mathbb{F}(\mu_0-x|\lambda,\theta) + \frac{\lambda}{1-\lambda}\mathbb{F}_1(\mu_0-x|\theta). \end{aligned}$$

By symmetry of f_0, we have $\mathbb{F}_0(x) = 1 - \mathbb{F}_0(-x)$, and thus $H_1(x|\lambda,\theta,\mathbb{F}) = H_2(x|\lambda,\theta,\mathbb{F})$. This means that if d is a distance mapping, then $d(H_1(.|\lambda,\theta,\mathbb{F}), H_2(.|\tilde{\lambda},\tilde{\theta},\mathbb{F})) = 0$ if $\lambda = \tilde{\lambda}$

and $\theta = \tilde{\theta}$. Otherwise the distance will be positive if d is chosen properly. In Bordes et al. [2006], the authors propose to use an L^q distance and define the estimation procedure by:

$$(\hat{\lambda}, \hat{\theta}) = \arg\inf_{\lambda,\theta} \left(\int \left| H_1(x|\lambda, \theta, \hat{\mathbb{F}}_n) - H_2(x|\lambda, \theta, \hat{\mathbb{F}}_n) \right|^q dx \right)^{1/q}$$

where $\hat{\mathbb{F}}_n$ is an estimator of $\mathbb{F}$. The authors proved consistency of this estimation procedure, but were unable to prove its asymptotic normality. Besides, Bordes and Vandekerkhove [2010] argue that the use of an L^q distance leads to numerical instability. They also propose to use the alternative distance $L^2(d\mathbb{F})$. The new procedure is now defined by:

$$(\hat{\lambda}, \hat{\theta}) = \arg\inf_{\lambda,\theta} \sum_{i=1}^{n} \left[H_1(x_i|\lambda, \theta, \hat{\mathbb{F}}_n) - H_2(x_i|\lambda, \theta, \hat{\mathbb{F}}_n) \right]^2. \tag{3.1.2}$$

Bordes and Vandekerkhove [2010] prove that the above estimator is consistent and asymptotically Gaussian. This method produces in practice good estimates even in difficult situations such as two-component Gaussian mixture when the two components are close provided that we restrict the proportion parameter inside an interval of the form $(\eta, 1-\eta)$ for a small η, say 0.1. It is however unusable in the context of distributions which are defined on a subset of $\mathbb{R}$. Besides, a generalization to the multivariate case does not seem simple.

Notice that, in the original approach presented by Bordes et al. [2006] or Bordes and Vandekerkhove [2010] we suppose that θ is given, so that the parametric component is fully known. We find however no problem in writing the same algorithm for the case when θ is considered unknown. In what concerns the theoretical results developed in these papers, we did not check their validity when θ is unknown. We mention the work of Maiboroda and Sugakova [2012] who use the approach of Bordes and Vandekerkhove [2010] and propose several methods to estimate the unknown component. They also study the theoretical properties of their estimators such as L-consistency and rates of convergence.

3.1.2 EM-type algorithms

This kind of algorithms is based on defining a vector of weights $w = (w_1, \cdots, w_n)$ for the observations and then estimate the unknown component using a weighted kernel estimator as follows:

$$\hat{f}_0(x|w) = \frac{1}{nh} \frac{1}{\sum 1 - w_i} \sum_{i=1}^{n} (1 - w_i) K\left(\frac{x - x_i}{h} \right).$$

The proportion is then, similarly to the EM algorithm, estimated by averaging these weights whereas the parameters of the known component are calculated by maximizing a weighted likelihood function. Such methods were proposed by several authors such as Song et al. [2010], Robin et al. [2007] and Ma et al. [2011]. They differ by how to calculate the weights. In tRobin et al. [2007] and Song et al. [2010], we calculate the weights by an iterative procedure in the same way as the EM algorithm does, i.e. as the quotient of the probability of being in the first component to the probability of being in the mixture. In Ma et al. [2011], the authors propose a weighted histogram to estimate the unknown component where the bins and their number are chosen before the estimation procedure (prior guess). They calculate after that the weights by maximizing a likelihood-like function related to the discretized model.

The method of Ma et al. [2011] has many drawbacks and practical issues. We need at first to precise a close interval for the values of the unknown component, then a good guess

for the number of bins. Besides, if one thinks about increasing the bins in order to give a closer estimate of f_0, it will cost much on the optimization step as each bin has its own parameter. Besides, in multivariate situations, the number of bins explodes easily and the optimization over the unknown weights becomes very difficult.
Other EM-type algorithms are very simple to implement and have good execution time when the number of observations is small, however, they do not perform very well in situations when the two components are close enough.
We present briefly the algorithms of Song et al. [2010] and Robin et al. [2007], and their analytical properties.

Robin et al. [2007] EM-type algorithm. The authors propose to estimate the weight vector w by the following iterative algorithm. For some initial value of λ and θ, say λ^0, θ^0, define at iteration $k+1$ for $k \geq 0$:

$$w_i^{(k+1)} = \frac{\lambda^{(k)} f_1(x_i|\theta^{(k)})}{\lambda^{(k)} f_1(x_i|\theta^{(k)}) + (1-\lambda^{(k)})\hat{f}_0(x_i|w^{(k)})}. \tag{3.1.3}$$

When the vector θ is known, the algorithm is proved to converge under mild conditions, see Theorem 1 in Robin et al. [2007]. We only need that the proportion to be inside the interval $(0,1)$ and that for all j and i the quantities $K((x_i - x_j)/h)$ are positive, which is *theoretically* fulfilled as long as we are *not* using a compacted-support kernel and there is no ties in the dataset[2]. On the other hand, the algorithm is proved to converge towards a fixed point of function ψ defined by:

$$\begin{aligned} \psi &= (\psi_1, \cdots, \psi_n) \\ \psi_j(w_1, \cdots, w_n) &= \frac{(1-\lambda)\sum_{i=1}^n (1-w_i)K((x_i-x_j)/h)/h}{(1-\lambda)\sum_{i=1}^n (1-w_i)K((x_i-x_j)/h)/h + \sum_{i=1}^n w_i f_1(x_j|\theta)}, \end{aligned}$$

provided that θ is known. In what concerns our objective, θ will be unknown, and this theoretical result may not hold. The algorithm is still applicable. Besides, the theoretical result supposes that the proportion is known. The authors propose to estimate the proportion by $\sum w_i/n$ as a natural choice, but state that this can easily lead the algorithm to converge to a proportion equals to 0 or 1. They propose another estimator suitable to the application they considered (estimation of the FDR) by adding an assumption that the distribution of the unknown component is defined on a semi-closed interval $(-\infty, a)$. Besides, the cdf of the parametric component must verify $\mathbb{F}_1(a) < 1$. This means that the semiparametric component distribution must have a lighter tail than the parametric component one. An R package, `kerfdr`, was written about this approach where the data is supposed to be bounded, see Guedj et al. [2009]. In our examples and simulations, the distribution of both components is the same; either $(0, \infty)$ or the whole real line $\mathbb{R}$. Thus, we cannot adapt this methodology.
We adapt the following algorithm based on the ideas of Robin et al. [2007] as follows. Let D be a kernel function and h be a pre-chosen window, and denote $D_i(x) = D(\frac{x-x_i}{h})$. Initialize the algorithm with $\lambda^{(0)}$, $\theta^{(0)}$ and a vector of weights for the observations say

[2]Which is simply ensured if the data is distributed from a continuous probability distribution.

$(w_1^{(0)}, \cdots, w_n^{(0)})$. At iteration k calculate

$$\begin{aligned}
\hat{p}_0^{(k)}(y_i) &= \frac{1}{\sum_j w_j^{(k-1)}} \sum_l w_l^{(k-1)} D_i(y_l); \\
\lambda^{(k)} &= \frac{1}{n} \sum_l w_l^{(k-1)}; \\
\theta^{(k)} &= \arg\max_\theta \sum_i w_i^{(k-1)} \log(p_1(y_i|\theta)); \\
w_i^{(k)} &= \frac{\lambda^{(k)} p_1(y_i|\theta^{(k)})}{\lambda^{(k)} p_1(y_i|\theta^{(k)}) + (1-\lambda^{(k)}) \hat{p}_0^{(k)}(y_i)}.
\end{aligned}$$

Repeat this iteration until convergence.

Song et al. [2010] EM-type algorithm. The authors propose to estimate the weight vector without the need to estimate the unknown component and merely using an estimate of the whole mixture. The authors propose two methods with no proofs of convergence. Let $\hat{f}_n$ be an estimator of the mixture density on the basis of an $n-$sample. The first recurrence proposed by the authors is given by:

$$w_i^{(k+1)} = \min\left[1, \quad \frac{\lambda^{(k)} f_1(x_i|\theta^{(k)})}{\hat{f}_n}\right]. \tag{3.1.4}$$

The authors argue that this formulation does not stabilize well and propose the alternative iteration:

$$w_i^{(k+1)} = \min\left[1, \quad \frac{2\lambda^{(k)} f_1(x_i|\theta^{(k)})}{\lambda^{(k)} f_1(x_i|\theta^{(k)}) + \hat{f}_n}\right], \tag{3.1.5}$$

and state that this new formulation is better without any theoretical justification. It is worth noting that these algorithms were only proposed in the context of a mixture of a two Gaussian components when one component is unknown. We, however, see no problem in using them in a more general context and even in the multivariate case since no particular constraint on f_1 is needed in order to write the iterative procedure.

3.1.3 Stochastic EM-type method

Bordes et al. [2007] have proposed a stochastic EM-type algorithm to estimate the parameters in model (3.1.1), i.e. under the symetry assumption. There is however no problem in using their algorithm in the general context of model (3.0.1). We give the general form for this algorithm which can also be used in a multivariate context as we will see in the simulation section.
The algorithm starts by giving an initial Bernoulli vector $Z^{(0)}$ which attributes a zero to coordinate Z_i if the observation x_i is drawn according to the unknown component f_0 and one if x_i is drawn according to the parametric component $f_1(.|\theta)$. We then calculate the initial proportion $\lambda^{(0)} = \sum_{i=1}^n Z_i^{(0)}/n$, and give an initial value for the vector θ, say $\theta^{(0)}$. At iteration $k+1$, calculate the kernel density estimator of the unknown component as follows:

$$\hat{f}_0(x|Z^{(k)}) = \frac{1}{h(1-\sum Z_i^{(k)})} \sum_{i=1}^n (1 - Z_i^{(k)}) K\left(\frac{x - x_i}{h}\right).$$

Calculate the weights by:

$$w_i^{(k)} = \frac{\lambda^{(k)} f_1(x_i|\theta^{(k)})}{\lambda^{(k)} f_1(x_i|\theta^{(k)}) + (1-\lambda^{(k)})\hat{f}_0(x_i|Z^{(k)})}.$$

Generate now a Bernoulli vector $Z^{(k+1)}$ with probabilities $(w_1^{(k)}, \cdots, w_n^{(k)})$, and calculate now the new proportion:

$$\lambda^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} Z_i^{(k+1)}.$$

Finally, we calculate the new vector of parameters $\theta^{(k+1)}$ by maximum likelihood using the observations for which $Z_i^{(k+1)}$ is equal to 1. We repeat this procedure until the sequence of proportions $\lambda^{(k)}$ stabilizes. It is preferable in the context of the stochastic EM algorithm not to keep the final iteration, but to average the results of the n_0 final iterations instead. For our simulations, we performed 5000 iterations and averaged the 4000 final iterations, see figure (3.1) for a better understanding.

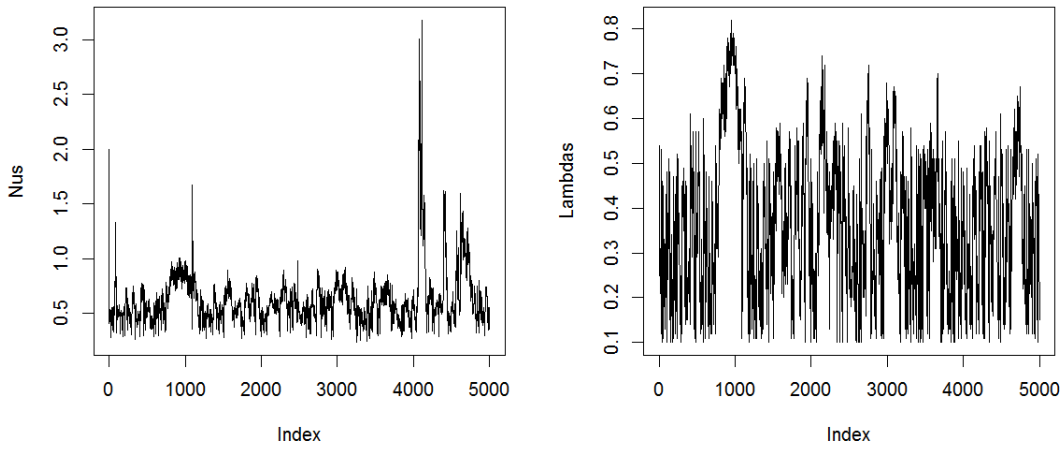

Figure 3.1: Fluctuations in a trajectory of the semiparametric SEM algorithm in a Weibull mixture.

This method has a good performance in regular situations when the two components do not overlap. Besides, the asymptotic behavior of the sequence of points which generates the algorithm remains an open problem.

3.1.4 $\pi-$maximizing method

Song et al. [2010] propose another kind of algorithm which is based on the identifiability condition of a mixture of two Gaussian components with known means. They state that no estimating procedure can distinguish between two sets of parameters (λ_1, θ_1) and (λ_2, θ_2) as long as they both verify:

$$\lambda_i f_1(x|\theta_i) < f(x), \forall x, \qquad \text{for } i = 1, 2.$$

This is because the unknown component whose form is not specified by any prior condition can take the form of a mixture to fill the gap between $\lambda f_1(x|\theta)$ and $f(x)$, see figure (3.2).

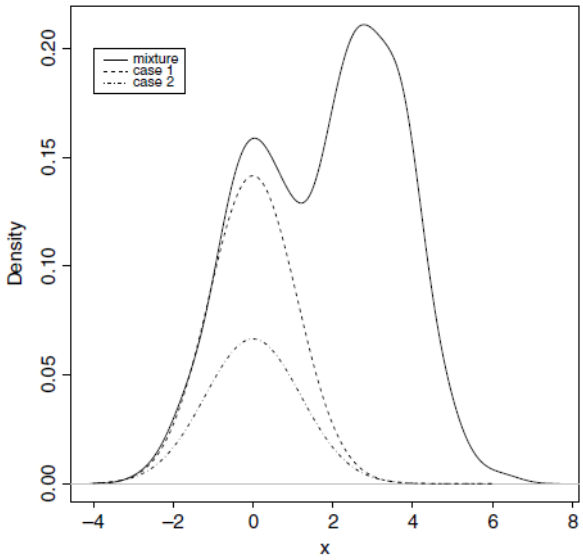

Figure 3.2: The solid line is the density estimation of the whole mixture. The dotted lines are two examples of normal densities that fit under the mixture density. These two cases can not be distinguished by an estimating procedure. Figure copied from Song et al. [2010].

Based on this idea, Song et al. [2010] propose the following estimation procedure:

$$\hat{\lambda} = \sup_{\theta} \min_{x_i} \frac{\hat{f}(x_i)}{f_1(x_i|\theta)} \tag{3.1.6}$$

$$\hat{\theta} = \arg\sup_{\theta} \min_{x_i} \frac{\hat{f}(x_i)}{f_1(x_i|\theta)}. \tag{3.1.7}$$

The authors prove that if $f_0(0) = 0$, f_1 is a centered Gaussian distribution with unknown variance σ^2 and that the tail of the mixture is not heavier than the tail of the Gaussian component, then:

$$\lambda^* = \sup_{\sigma>0} \inf_{x} \frac{f(x)}{f_1(x|\sigma)},$$

where the supremum is attained at $\sigma = \sigma^*$ the true value of the scale. This constitutes a simple theoretical justification to the estimation procedure (3.1.6, 3.1.7). The authors do not provide, however, any *real* proof of consistency or a generalization to other distributions other than the Gaussian. Notice that a proof of consistency does not seem to be simple mainly because we are dealing with a double optimization procedure.

3.2 Semiparametric models defined through linear constraints

Now that the idea of the semiparametric mixture model is presented, we proceed to propose our new model. We want to integrate linear information in the semiparametric model and propose an estimation procedure which permits to retrieve the true vector of parameters defining the model on the basis of a given i.i.d. sample $X_1, \cdots, X_n$ drawn from the mixture distribution P_T.
We prefer to proceed step by step. The previous paragraphs introduced semiparametric

mixture models. Now we will present models which can be defined through a linear information. These models are not necessarily mixtures of distributions. Besides, the constraints or the linear information defining the model will apply over the whole model, i.e. if the model is a mixture then the constraints apply over the whole mixture and not only over one component. We give in this section a brief idea of what the literature offers us to study such model. In the next section we will proceed to aggregate the two ideas, i.e. mixture models and semiparametric models defined through linear constraints, in order to introduce our semiparametric mixture model where a component is parametric (but not fully known) and a component is defined through linear constraints.

3.2.1 Definition and examples

Denote by M^+ the set of all probability measures (p.m.) defined on the same measurable space as P_T, i.e. $(\mathbb{R}^r, \mathscr{B}(\mathbb{R}^r))$.

Definition 3.2.1. *Let $X_1, \cdots, X_n$ be random variables drawn independently from the probability distribution P_T. A semiparametric model is a collection of probability measures $\mathcal{M}_\alpha(P_T)$, for $\alpha \in \mathcal{A} \subset \mathbb{R}^s$, absolutely continuous with respect to P_T which verifies a set of linear constraints, i.e.*

$$\mathcal{M}_\alpha(P_T) = \left\{ Q \in M^+ \ \text{ such that } \ Q \ll P_T, \ \int g(x) dQ(x) = m(\alpha) \right\}, \tag{3.2.1}$$

where $g : \mathbb{R}^r \to \mathbb{R}^\ell$ and $m : \mathcal{A} \to \mathbb{R}^\ell$ are specified vector-valued functions.

This semiparametric model was studied by many authors, see Broniatowski and Keziou [2012], Broniatowski and Decurninge [2016] (with dP replaced by dP^{-1} the quantile measure), Owen [1990] (in the empirical likelihood context) and Jiahua Chen [1993] (for finite population problems). It is possible in the above definition to make g depend on the parameter vector α, but we stay for the sake of simplicity with the assumption that g does not depend on α. The theoretical approach we present in this chapter remains valid if g depends on α with slight modification on the assumptions and more technicalities at the level of the proofs.

Example 3.2.1. A simple and standard example is a model defined through moment constraints. Let P_T be the Weibull distribution with scale a^* and shape b^*. We define $\mathcal{M}_\alpha$ with $\alpha = (a, b) \in (0, \infty)^2$ to be the set of all probability measures whose first three moments are given by:

$$\int x^i dQ(x) = a^i \Gamma(1 + i/b), \qquad i = 1, 2, 3.$$

The set $\mathcal{M}_{\alpha^*}$ is a "neighborhood" of probability measures of the Weibull distribution P_T. It contains all probability measures absolutely continuous with respect to the Weibull mixture P_T and which share the first three moments with it. The union of the sets $\mathcal{M}_\alpha$ contains all probability measures whose first three moments share the same analytic form as a Weibull distribution.

If the true distribution P_T verifies the set of ℓ constraints (3.2.1) for some α^*, then the set

$$\mathcal{M}_{\alpha^*}(P_T) = \left\{ Q \in M^+ \ \text{ such that } \ Q \ll P_T, \ \int g(x) dQ(x) = m(\alpha^*) \right\} \tag{3.2.2}$$

constitutes a "neighborhood" of probability measures of P_T. Generally, one would rather consider the larger "neighborhood" defined by

$$\mathcal{M} = \bigcup_{\alpha \in \mathcal{A}} \mathcal{M}_\alpha,$$

because the value of α^* is unknown and needs to be estimated. The estimation procedure aims at finding α^* the ("best") vector for which $P_T \in M_{\alpha^*}$. This is generally done by either solving the set of equations (3.2.1) defining the constraints for Q replaced by (an estimate of) P_T or by minimizing a suitable distance-like function between the set $\mathcal{M}$ and (an estimate of) P_T. In other words, we search for the "projection" of P_T on $\mathcal{M}$. Solving the set of equations (3.2.1) is in general a difficult task since it is a set of nonlinear equations. In the literature, similar problems were solved using the Fenchel-Legendre duality. Broniatowski and Keziou [2012] proposed to estimate the value of α^* using $\varphi-$divergences developing an efficient and a simple estimation method using the duality of Fenchel-Legendre. In the next chapter, we will see similar semiparametric models defined through linear constraints over the quantile measures. Broniatowski and Decurninge [2016] proposed also to use $\varphi-$divergences and basing on the duality of Fenchel-Legendre to estimate his "semiparametric linear quantile models", see Chapter 4.

In the next paragraph, we present the duality technique which will be essential in the development of our estimation method.

3.2.2 Estimation using $\varphi-$divergences and the duality technique

As mentioned in the previous paragraph, $\varphi-$divergences offer an efficient tool to handle the projection of a probability measure on a set of probability measures. This remains also valid for finite signed measures. We will explain how we may use $\varphi-$divergences to find the "best" vector α^* such that $P_T \in \mathcal{M}_{\alpha^*}$. The optimality of the solution is absolute if the distribution P_T verifies the constraints for some value α^*. Otherwise, the vector α^* is sub-optimum in the sens that optimality is considered merely from a point of view of $\varphi-$projections (see definitions below). The following definitions concern the notion of φ-projection of finite signed measures over a set of finite signed measures and are essential in order to clearly present the estimation procedure and introduce our new estimation methodology. The context of semiparametric models presented earlier can be extended to finite signed measures, see the theory in Broniatowski and Keziou [2012]. For our study, the use of finite *signed* measures and not only probability measures is essential as will be demonstrated in the next section.

Definition 3.2.2. *Let $\mathcal{M}$ be some subset of M, the space of finite signed measures. The φ-divergence between the set $\mathcal{M}$ and some finite signed measure P, noted as $D_\varphi(\mathcal{M}, P)$, is given by*

$$D_\varphi(\mathcal{M}, P) := \inf_{Q \in \mathcal{M}} D_\varphi(Q, P). \tag{3.2.3}$$

Furthermore, we define the $\varphi-$divergence between two subsets of M, say $\mathcal{M}$ and $\mathcal{N}$ by:

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

Definition 3.2.3. *Assume that $D_\varphi(\mathcal{M}, P)$ is finite. A measure $Q^* \in \mathcal{M}$ such that*

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

is called a φ-projection of P onto $\mathcal{M}$. This projection may not exist, or may not be defined uniquely.

Estimation of the semiparametric model using $\varphi-$divergences is summarized by the following optimization problem:

$$\alpha^* = \arg\inf_{\alpha\in\mathcal{A}} \inf_{Q\in\mathcal{M}_\alpha} D_\varphi(Q, P_T). \tag{3.2.4}$$

We are, then, searching for the projection of P_T on the set $\mathcal{M} = \cup_\alpha \mathcal{M}_\alpha$. We are more formally interested in the vector α^* for which the projection of P_T on $\mathcal{M}$ belongs to the set $\mathcal{M}_{\alpha^*}$. Notice that the sets $\mathcal{M}_\alpha$ need to be disjoint so that the projection cannot belong to several sets in the same time.
The magical property of $\varphi-$divergences stems from their characterization of the projection of a finite signed measure P onto a set $\mathcal{M}$ of finite signed measures, see Broniatowski and Keziou [2006] Theorem 3.4. Such characterization permits to transform the search of a projection in an infinite dimensional space to the search of a vector ξ in $\mathbb{R}^\ell$ through the duality of Fenchel-Legendre and thus simplify the optimization problem (3.2.4). Note that Theorem 3.4 from Broniatowski and Keziou [2006] provides a formal characterization of the projection, but we will only use it implicitly.
Let φ be a strictly convex function which verifies the same properties mentioned in the definition of a $\varphi-$divergence, see paragraph 1.1.1. The Fenchel-Legendre transform of φ, say ψ is defined by:

$$\psi(t) = \sup_{x\in\mathbb{R}} \{tx - \varphi(x)\}, \qquad \forall t \in \mathbb{R}.$$

We are concerned with the convex optimization problem

$$(\mathcal{P}) \qquad \inf_{Q\in\mathcal{M}_\alpha} D_\varphi(Q, P_T). \tag{3.2.5}$$

We associate to $(\mathcal{P})$ the following dual problem

$$(\mathcal{P}^*) \qquad \sup_{\xi\in\mathbb{R}^\ell} \xi^t m(\alpha) - \int \psi\left(\xi^t g(x)\right) dP_T(x). \tag{3.2.6}$$

We require that φ is differentiable. Assume furthermore that $\int |g_i(x)| dP_T(x) < \infty$ for all $i = 1, \ldots, \ell$ and there exists some measure Q_T a.c.w.r.t. P_T such that $D_\varphi(Q_T, P_T) < \infty$. According to Proposition 1.4 in Decurninge [2015] (see also Proposition 4.2 in Broniatowski and Keziou [2012]) we have a strong duality attainment, i.e. $(\mathcal{P}) = (\mathcal{P}^*)$. In other words,

$$\inf_{Q\in\mathcal{M}_\alpha} D_\varphi\left(Q, P_T\right) = \sup_{\xi\in\mathbb{R}^\ell} \xi^t m(\alpha) - \int \psi\left(\xi^t g(x)\right) dP_T(x). \tag{3.2.7}$$

The estimation procedure of the semiparametric model (3.2.4) is now simplified into the following finite-dimensional optimization problem

$$\begin{aligned} \alpha^* &= \arg\inf_{\alpha\in\mathcal{A}} D_\varphi(\mathcal{M}_\alpha, P_T) \\ &= \arg\inf_{\alpha\in\mathcal{A}} \sup_{\xi\in\mathbb{R}^\ell} \xi^t m(\alpha) - \int \psi\left(\xi^t g(x)\right) dP_T(x). \end{aligned}$$

This is indeed a feasible procedure since we only need to optimize a real function over $\mathbb{R}^\ell$. Examples of such procedures can be found in Broniatowski and Keziou [2012], Broniatowski and Decurninge [2016], Newey and Smith [2004] and the references therein. Robustness of this procedure was studied theoretically by Toma [2013] and was shown

numerically[3] in Broniatowski and Decurninge [2016].

Now that all notions and analytical tools are presented, we proceed to the main objective of this chapter; semiparametric mixtures models. The following section defines such models and presents a method to estimate them using $\varphi-$divergences. We study after that the asymptotic properties of the vector of estimates.

3.3 Semiparametric two-component mixture models when one component is defined through linear constraints

3.3.1 Definition and identifiability

Definition 3.3.1. *Let X be a random variable taking values in $\mathbb{R}^r$ distributed from a probability measure P. We say that $P(.|\phi)$ with $\phi = (\lambda, \theta, \alpha)$ is a two-component semiparametric mixture model subject to linear constraints if it can be written as follows:*

$$\begin{aligned} P(.|\phi) &= \lambda P_1(.|\theta) + (1-\lambda)P_0 \quad s.t. \\ P_0 \in \mathcal{M}_\alpha &= \left\{ Q \in M \ s.t. \ \int_{\mathbb{R}^r} dQ(x) = 1, \int_{\mathbb{R}^r} g(x) dQ(x) = m(\alpha) \right\} \end{aligned} \tag{3.3.1}$$

for $\lambda \in (0,1)$ the proportion of the parametric component, $\theta \in \Theta \subset \mathbb{R}^d$ a set of parameters defining the parametric component, $\alpha \in \mathcal{A} \subset \mathbb{R}^s$ is the constraints parameter vector and finally $m(\alpha) = (m_1(\alpha), \cdots, m_\ell(\alpha))$ is a vector-valued function determining the value of the constraints.

The identifiability of the model was not questioned in the context of Section 3.2 because it suffices that the sets $\mathcal{M}_\alpha$ are disjoint (the function $m(\alpha)$ is one-to-one). However, in the context of this semiparametric mixture model, identifiability cannot be achieved only by supposing that the sets $\mathcal{M}_\alpha$ are disjoint.

Definition 3.3.2. *We say that the two-component semiparametric mixture model subject to linear constraints is identifiable if it verifies the following assertion. For two triplets $(\lambda, \theta, \alpha)$ and $(\tilde{\lambda}, \tilde{\theta}, \tilde{\alpha})$ in $\Phi = (0,1) \times \times \times \mathcal{A}$, if*

$$\lambda P_1(.|\theta) + (1-\lambda)P_0 = \tilde{\lambda} P_1(.|\tilde{\theta}) + (1-\tilde{\lambda})\tilde{P}_0, \quad \text{with } P_0 \in \mathcal{M}_\alpha, \tilde{P}_0 \in \mathcal{M}_{\tilde{\alpha}}, \tag{3.3.2}$$

then $\lambda = \tilde{\lambda}, \theta = \tilde{\theta}$ and $P_0 = \tilde{P}_0$ (and hence $\alpha = \tilde{\alpha}$).

This is the same identifiability concept considered in Bordes et al. [2006] where the authors exploited their symmetry assumption over $\mathbb{P}_0$ and built a system of moments equations. They proved that if P_1 is also symmetric, then equation (3.3.2) has two solutions, otherwise it has three solutions. Their idea appears here in a natural way in order to prove the identifiability of our semiparametric mixture model (3.3.1).

Proposition 3.3.1. *For a given mixture distribution $P_T = P(.|\phi^*)$, suppose that the system of equations:*

$$\frac{1}{1-\lambda} m^* - \frac{\lambda}{1-\lambda} m_1(\theta) = m_0(\alpha)$$

[3] The results in Broniatowski and Decurninge [2016] show that his estimator is not robust against outliers, but robust against misspecification.

where $m^* = \int g(x)dP_T(x)$ *and* $m_1(\theta) = \int g(x)dP_1(x|\theta)$*, has a unique solution* $(\lambda^*, \theta^*, \alpha^*)$*. Then, equation (3.3.2) has a unique solution, i.e.* $\lambda = \tilde{\lambda}, \theta = \tilde{\theta}$ *and* $P_0 = \tilde{P}_0$*, and the semiparametric mixture model* $P_T = P(.|\phi^*)$ *is identifiable.*

The proof is differed to Appendix 3.7.1.

Example 3.3.1 (Semiparametric two-component Gaussian mixture)**.** Suppose that $P_1(.|\theta)$ is a Gaussian model $\mathcal{N}(\mu_1, 1)$. Suppose also that the set of constraints is defined as follows:

$$\mathcal{M}_{\mu_0^*} = \left\{ f_0 \text{ s.t. } \int f_0(x)dx = 1, \quad \int_{\mathbb{R}} x f_0(x)dx = \mu_0^*, \quad \int_{\mathbb{R}} x^2 f_0(x)dx = 1 + \mu_0^{*2} \right\}.$$

We would like to study the identifiability of the two-component semiparametric Gaussian mixture whose unknown component P_0 shares the first two moments with the Gaussian distribution $\mathcal{N}(\mu_0^*, 1)$ for a known μ_0^*. Using Proposition 3.3.1, it suffices to study the system of equations

$$\begin{aligned} \frac{1}{1-\lambda}\int x f(x|\mu_1^*, \mu_0^*, \lambda^*)dx - \frac{\lambda}{1-\lambda}\int x f_1(x|\mu_1)dx &= \mu_0^* \\ \frac{1}{1-\lambda}\int x^2 f(x|\mu_1^*, \mu_0^*, \lambda^*)dx - \frac{\lambda}{1-\lambda}\int x^2 f_1(x|\mu_1)dx &= 1 + \mu_0^{*2}. \end{aligned}$$

Recall that $\int x f(x|\mu_1^*, \mu_0^*, \lambda^*)dx = \lambda^*\mu_1^* + (1-\lambda^*)\mu_0^*$ and $\int x^2 f(x|\mu_1^*, \mu_0^*, \lambda^*)dx = 1 + \lambda^*\mu_1^{*2} + (1-\lambda^*)\mu_0^{*2}$. The first equation in the previous system entails that:

$$\lambda\mu_1 - \lambda\mu_0^* = \lambda^*\mu_1^* - \lambda^*\mu_0^*. \tag{3.3.3}$$

The second equation gives:

$$\lambda^*(1+\mu_1^{*2}) - \lambda^*(1+\mu_0^{*2}) = \lambda\mu_1^2 - \lambda\mu_0^{*2} \tag{3.3.4}$$

The nonlinear system of equations (3.3.3, 3.3.4) has a solution for $\mu_1 = \mu_1^*, \lambda = \lambda^*$. Suppose by contradiction that $\mu_1 \neq \mu_1^*$ and check if there are other solutions. The system (3.3.3, 3.3.4) implies:

$$\lambda = \frac{\lambda^*(1+\mu_1^{*2}) - \lambda^*(1+\mu_0^{*2})}{(\mu_1 - \mu_0^*)(\mu_1 + \mu_0^*)} = \frac{\lambda^*\mu_1^* - \lambda^*\mu_0^*}{\mu_1 - \mu_0^*}$$

This entails that

$$\mu_1 + \mu_0^* = \frac{\lambda^*\left[(1+\mu_1^{*2}) - (1+\mu_0^{*2})\right]}{\lambda^*\left[\mu_1^* - \mu_0^*\right]} = \mu_1^* + \mu_0^*$$

Hence, $\mu_1 = \mu_1^*$ which contradicts what we have assumed. Thus $\mu_1 = \mu_1^*, \lambda = \lambda^*$ is the only solution. We conclude that if μ_0^* is known and that we impose two moments constraints over f_0, then the semiparametric two-component Gaussian mixture model is identifiable. Notice that imposing only one condition on the first moment is not sufficient since any value of $\lambda \in (0,1)$ would produce a corresponding solution for μ_1 in equation (3.3.3). We therefore are in need for the second constraint. Notice also that if $\lambda = \lambda^*$, then $\mu_1 = \mu_1^*$. This means, by continuity of the equation over (λ, μ_1), if λ is initialized in a close neighborhood of λ^*, then μ_1 would be estimated near μ_1^*. This may represent a remedy if we could not impose but one moment constraint.

3.3.2 An algorithm for the Estimation of the semiparametric mixture model

We have seen in paragraph 3.2.2 that it is possible to use $\varphi-$divergences to estimate a semiparametric model as long as the constraints apply over $P(.|\phi)$, i.e. the whole mixture. In our case, the constraints apply only on a component of the mixture. It is thus reasonable to consider a "model" expressed through P_0 instead of P. We have:

$$P_0 = \frac{1}{1-\lambda}P(.|\phi) - \frac{\lambda}{1-\lambda}P_1(.|\theta).$$

Denote $P_T = P(.|\phi^*)$ with $\phi^* = (\lambda^*, \theta^*, \alpha^*)$ the distribution which generates the observed data. Denote also P_0^* to the true semiparametric component of the mixture P_T. The only information we hold about P_0^* is that it belongs to a set $\mathcal{M}_{\alpha^*}$ for some (possibly unknown) $\alpha^* \in \mathcal{A}$. Besides, it verifies:

$$P_0^* = \frac{1}{1-\lambda^*}P_T - \frac{\lambda^*}{1-\lambda^*}P_1(.|\theta^*). \tag{3.3.5}$$

We would like to retrieve the value of the vector $\phi^* = (\lambda^*, \theta^*, \alpha^*)$ provided a sample $X_1, \cdots, X_n$ drawn from P_T and that $P_0^* \in \cup_\alpha \mathcal{M}_\alpha$. Consider the set of signed measures:

$$\mathcal{N} = \left\{ Q = \frac{1}{1-\lambda}P_T - \frac{\lambda}{1-\lambda}P_1(.|\theta), \quad \lambda \in (0,1), \theta \in \Theta \right\}. \tag{3.3.6}$$

Notice that P_0^* belongs to this set for $\lambda = \lambda^*$ and $\theta = \theta^*$. On the other hand, P_0^* is supposed, for simplicity, to belong to the union $\cup_{\alpha\in\mathcal{A}}\mathcal{M}_\alpha$. We may now write,

$$P_0^* \in \mathcal{N} \bigcap \cup_{\alpha\in\mathcal{A}}\mathcal{M}_\alpha.$$

If we suppose now that the intersection $\mathcal{N} \bigcap \cup_{\alpha\in\mathcal{A}}\mathcal{M}_\alpha$ contains only one element (see paragraph 3.3.4 for a discussion) which would be a fortiori P_0^*, then it is very reasonable to consider an estimation procedure by calculating some "distance" between the two sets $\mathcal{N}$ and $\cup_{\alpha\in\mathcal{A}}\mathcal{M}_\alpha$. Such distance can be measured using a $\varphi-$divergence by (see Definition 3.2.2):

$$D_\varphi(\mathcal{M}, \mathcal{N}) = \inf_{Q\in\mathcal{N}} \inf_{P_0\in\mathcal{M}} D_\varphi(P_0, Q). \tag{3.3.7}$$

We may reparametrize this distance using the definition of $\mathcal{N}$. Indeed,

$$\begin{aligned} D_\varphi(\cup_\alpha\mathcal{M}_\alpha, \mathcal{N}) &= \inf_{Q\in\mathcal{N}} \inf_{P_0\in\cup_\alpha\mathcal{M}_\alpha} D_\varphi(P_0, Q) \\ &= \inf_{\lambda,\theta} \inf_{\alpha, P_0\in\mathcal{M}_\alpha} D_\varphi\left(P_0, \frac{1}{1-\lambda}P_T - \frac{\lambda}{1-\lambda}P_1(.|\theta)\right). \end{aligned} \tag{3.3.8}$$

If we still have P_0^* as the only signed measure which belongs to both $\mathcal{N}$ and $\cup_\alpha\mathcal{M}_\alpha$, then, the argument of the infimum in (3.3.8) is none other than $(\lambda^*, \theta^*, \alpha^*)$, i.e.

$$(\lambda^*, \theta^*, \alpha^*) = \arg\inf_{\lambda,\theta,\alpha} \inf_{P_0\in\mathcal{M}_\alpha} D_\varphi\left(P_0, \frac{1}{1-\lambda}P(.|\phi^*) - \frac{\lambda}{1-\lambda}P_1(.|\theta)\right). \tag{3.3.9}$$

It is important to notice that if $P_0^* \notin \cup\mathcal{M}_\alpha$, then the procedure still makes sense. Indeed, we are searching for the best measure of the form $\frac{1}{1-\lambda}P_T - \frac{\lambda}{1-\lambda}P_1(.|\theta)$ which verifies the constraints.

3.3.3 The algorithm in practice : Estimation using the duality technique and plug-in estimate

The Fenchel-Legendre duality permits to transform the problem of minimizing under linear constraints in a possibly infinite dimensional space into an unconstrained optimization problem in the space of Lagrangian parameters over $\mathbb{R}^{\ell+1}$, where $\ell+1$ is the number of constraints. We will apply the duality result presenter earlier in paragraph 3.2.2 on the inner optimization in equation (3.3.8). Redefine the function m as $m(\alpha) = (m_0(\alpha), m_1(\alpha), \cdots, m_\ell(\alpha))$ where $m_0(\alpha) = 1$. We have:

$$\begin{aligned}\inf_{Q\in\mathcal{M}_\alpha} D_\varphi\left(Q, \frac{1}{\lambda-1}P_T - \frac{\lambda}{1-\lambda}P_1(.|\theta)\right) &= \sup_{\xi\in\mathbb{R}^{l+1}} \xi^t m(\alpha) - \frac{1}{1-\lambda}\int \psi\left(\xi^t g(x)\right)\left(dP_T(x) - \lambda dP_1(x|\theta)\right)\\ &= \sup_{\xi\in\mathbb{R}^{l+1}} \xi^t m(\alpha) - \frac{1}{1-\lambda}\int \psi\left(\xi^t g(x)\right) dP_T(x)\\ &\quad + \frac{\lambda}{1-\lambda}\int \psi\left(\xi^t g(x)\right) dP_1(x|\theta).\end{aligned}$$

Inserting this result in (3.3.9) gives that:

$$\begin{aligned}(\lambda^*, \theta^*, \alpha^*) &= \arg\inf_{\phi} \inf_{Q\in\mathcal{M}_\alpha} D_\varphi\left(Q, \frac{1}{\lambda-1}P_T - \frac{\lambda}{1-\lambda}P_1(.|\theta)\right)\\ &= \arg\inf_{\phi} \sup_{\xi\in\mathbb{R}^{l+1}} \xi^t m(\alpha) - \frac{1}{1-\lambda}\int \psi\left(\xi^t g(x)\right) dP_T(x)\\ &\quad + \frac{\lambda}{1-\lambda}\int \psi\left(\xi^t g(x)\right) dP_1(x|\theta).\end{aligned}$$

The right hand side can be estimated on the basis of an $n-$sample drawn from P_T, say $X_1, \cdots, X_n$, by a simple plug-in of the empirical measure P_n. The resulting procedure can now be written as:

$$\begin{aligned}(\hat{\lambda}, \hat{\theta}, \hat{\alpha}) &= \arg\inf_{\lambda,\theta,\alpha} \sup_{\xi\in\mathbb{R}^{l+1}} \xi^t m(\alpha) - \frac{1}{1-\lambda}\frac{1}{n}\sum_{i=1}^{n} \psi\left(\xi^t g(X_i)\right)\\ &\quad + \frac{\lambda}{1-\lambda}\int \psi\left(\xi^t g(x)\right) dP_1(x|\theta). \qquad (3.3.10)\end{aligned}$$

This is a feasible procedure in the sens that we only need the data, the set of constraints and the model of the parametric component.

Example 3.3.2 (Chi square)**.** Let's take the case of the χ^2 divergence for which $\varphi(t) = (t-1)^2/2$. The Convex conjugate of φ is given by $\psi(t) = t^2/2 + t$. For $(\lambda, \theta, \alpha) \in \Phi$, we have:

$$\begin{aligned}\inf_{Q\in\mathcal{M}_\alpha} D_\varphi\left(Q, \frac{1}{\lambda-1}P_T - \frac{\lambda}{1-\lambda}P_1(.|\theta)\right) &= \sup_{\xi\in\mathbb{R}^{l+1}} \xi^t m(\alpha) - \frac{1}{1-\lambda}\int \left[\frac{1}{2}\left(\xi^t g(x)\right)^2 + \xi^t g(x)\right] dP_T(x)\\ &\quad + \frac{\lambda}{1-\lambda}\int \left[\frac{1}{2}\left(\xi^t g(x)\right)^2 + \xi^t g(x)\right] dP_1(x|\theta).\end{aligned}$$

It is interesting to note that the supremum over ξ can be calculated explicitly. Clearly, the optimized function is a polynomial of ξ and thus infinitely differentiable. The Hessian matrix is equal to $-\Omega$ where:

$$\Omega = \int g(x)g(x)^t\left(\frac{1}{1-\lambda}dP(x) - \frac{\lambda}{1-\lambda}dP_1(x|\theta)\right). \qquad (3.3.11)$$

If the measure $\frac{1}{1-\lambda}dP - \frac{\lambda}{1-\lambda}dP_1(.|\theta)$ is positive, then Ω is symmetric definite positive (s.d.p) and the Hessian matrix is symmetric definite negative. Consequently, the supremum over ξ is ensured to exist. If it is a signed measure, then the supremum might be infinity. We may now write:

$$\xi(\phi) = \Omega^{-1}\left(m(\alpha) - \int g(x)\left(\frac{1}{1-\lambda}dP(x) - \frac{\lambda}{1-\lambda}dP_1(x|\theta)\right)\right), \quad \text{if } \Omega \text{ is s.d.p}$$

For the empirical criterion, we define similarly Ω_n by:

$$\Omega_n = \frac{1}{n}\frac{1}{1-\lambda}\sum_{i=1}^{n} g(X_i)g(X_i)^t - \frac{\lambda}{1-\lambda}\int g(x)g(x)^t dP_1(x|\theta). \tag{3.3.12}$$

The solution to the corresponding supremum over ξ is given by:

$$\xi_n(\phi) = \Omega_n^{-1}\left(m(\alpha) - \frac{1}{n}\frac{1}{1-\lambda}\sum_{i=1}^{n} g(X_i) + \frac{\lambda}{1-\lambda}\int g(x)dP_1(x|\theta)\right), \quad \text{if } \Omega_n \text{ is s.d.p}$$

3.3.4 Uniqueness of the solution "under the model"

By a unique solution we mean that only one measure, which can be written in the form of $\frac{1}{1-\lambda}P_T - \frac{\lambda}{1-\lambda}P_1(.|\theta)$, verifies the constraints with a unique triplet $(\lambda^*, \theta^*, \alpha^*)$. The existence of a unique solution is essential in order to ensure that the procedure (3.3.9) is a reasonable estimation method. We provide next a result ensuring the uniqueness of the solution. The idea is based on the identification of the intersection of the set $\mathcal{N} \cap \mathcal{M}$. The proof is differed to Appendix 3.7.2.

Proposition 3.3.2. *Assume that $P_0^* \in \mathcal{M} = \cup_\alpha \mathcal{M}_\alpha$. Suppose also that:*

1. *the system of equations:*

$$\int g_i(x)\left(dP(x|\phi^*) - \lambda dP_1(x|\theta)\right) = (1-\lambda)m_i(\alpha), \qquad i = 1, \cdots, \ell \tag{3.3.13}$$

   has a unique solution $(\lambda^*, \theta^*, \alpha^*)$;

2. *the function $\alpha \mapsto m(\alpha)$ is one-to-one;*

3. *for any $\theta \in \Theta$ we have :*

$$\lim_{\|x\|\to\infty} \frac{dP_1(x|\theta)}{dP_T(x)} = c, \qquad \text{with } c \in [0, \infty) \setminus \{1\};$$

4. *the parametric component is identifiable, i.e. if $P_1(.|\theta) = P_1(.|\theta')$ dP_T-a.e. then $\theta = \theta'$,*

then, the intersection $\mathcal{N} \cap \mathcal{M}$ contains a unique measure P_0^*, and there exists a unique vector $(\lambda^*, \theta^*, \alpha^*)$ such that $P_T = \lambda^* P_1(.|\theta^*) + (1-\lambda)P_0^*$ where P_0^* is given by (3.3.5) and belongs to $\mathcal{M}_{\alpha^*}$. Moreover, provided assumptions 2-4, the conclusion holds if and only if assumption 1 is fulfilled.

There is no general result for a non linear system of equations to have a unique solution; still, it is necessary to ensure that $\ell \geq d+s+1$, otherwise there would be an infinite number of signed measures in the intersection $\mathcal{N} \bigcap \cup_{\alpha\in\mathcal{A}} \mathcal{M}_\alpha$.

Remark 3.3.1. Assumptions 3 and 4 of Proposition 3.3.2 are used to prove the identifiability of the "model" $\left(\frac{1}{1-\lambda}P_T - \frac{\lambda}{1-\lambda}P_1(.|\theta)\right)_{\lambda,\theta}$. Thus, according to the considered situation we may find simpler ones for particular cases (or even for the general case). Our assumptions remain sufficient but not necessary for the proof.

Example 3.3.3. One of the most popular models in clustering is the Gaussian multivariate mixture (GMM). Suppose that we have two classes. Linear discriminant analysis (LDA) is based on the hypothesis that the covariance matrix of the two classes is the same. Let X be a random variable which takes its values in $\mathbb{R}^2$ and is drawn from a mixture model of two components. In the context of LDA, the model has the form:

$$f(x,y|\lambda,\mu_1,\mu_2,\Sigma) = \lambda f_1(x,y|\mu_1,\Sigma) + (1-\lambda)f_1(x,y|\mu_2,\Sigma),$$

with:

$$f_1(x,y|\mu_1,\Sigma) = \frac{1}{2\pi\sqrt{|\det(\Sigma)|}}\exp\left[-\frac{1}{2}((x,y)^t-\mu_1)^t\Sigma((x,y)^t-\mu_1)\right], \quad \Sigma = \begin{pmatrix}\sigma^2 & \rho\\ \rho & \sigma^2\end{pmatrix}.$$

We would like to relax the assumption over the second component by keeping the fact that the covariance matrix is the same as the one of the first component. We will start by imposing the very natural constraints on the second component.

$$\begin{aligned}
\int x f_0(x,y)dxdy &= \mu_{2,1},\\
\int y f_0(x,y)dxdy &= \mu_{2,2},\\
\int x^2 f_0(x,y)dxdy &= \sigma^2,\\
\int y^2 f_0(x,y)dxdy &= \sigma^2,\\
\int xy f_0(x,y)dxdy &= \rho + \mu_{2,1}\mu_{2,2} - \mu_{2,1}^2 - \mu_{2,2}^2.
\end{aligned}$$

These constraints concern only the fact that the covariance matrix Σ is the same as the one of the Gaussian component (the parameteric one). In order to see whether this set of constraints is sufficient for the existence of a unique measure in the intersection $\mathcal{N}\cap\mathcal{M}$, we need to write the set of equations corresponding to (3.3.13) in Proposition 3.3.2.

$$\begin{aligned}
\int x\left[\frac{1}{1-\lambda}f(x,y) - \frac{\lambda}{1-\lambda}f_1(x,y|\mu_1,\sigma,\rho)\right]dxdy &= \mu_{2,1},\\
\int y\left[\frac{1}{1-\lambda}f(x,y) - \frac{\lambda}{1-\lambda}f_1(x,y|\mu_1,\sigma,\rho)\right]dxdy &= \mu_{2,2},\\
\int x^2\left[\frac{1}{1-\lambda}f(x,y) - \frac{\lambda}{1-\lambda}f_1(x,y|\mu_1,\sigma,\rho)\right]dxdy &= \sigma^2,\\
\int y^2\left[\frac{1}{1-\lambda}f(x,y) - \frac{\lambda}{1-\lambda}f_1(x,y|\mu_1,\sigma,\rho)\right]dxdy &= \sigma^2,\\
\int xy\left[\frac{1}{1-\lambda}f(x,y) - \frac{\lambda}{1-\lambda}f_1(x,y|\mu_1,\sigma,\rho)\right]dxdy &= \rho + \mu_{2,1}\mu_{2,2} - \mu_{2,1}^2 - \mu_{2,2}^2,
\end{aligned}$$

The number of parameters is 7, and we only have 5 equations. In order for the problem to have a unique solution, it is necessary to either add two other constraints or to consider

for example $\mu_1 = (\mu_{1,1}, \mu_{1,2})$ to be known[4]. Other solutions exist, but depend on the prior information. We may imagine an assumption of the form $\mu_{1,1} = a\mu_{1,2}$ and $\mu_{2,1} = b\mu_{2,2}$ for given constants a and b.
The gain from relaxing the normality assumption on the second component is that we are building a model which is not constrained to a Gaussian form for the second component, but rather to a form which suits the data. The price we pay is the number of relevant constraints which must be at least equal to the number of unknown parameters.

3.4 Asymptotic properties of the new estimator

3.4.1 Consistency

The double optimization procedure defining the estimator $\hat{\phi}$ defined by (3.3.10) does not permit us to use M-estimates methods to prove consistency. In Keziou [2003] Proposition 3.7 and in Broniatowski and Keziou [2009a] Proposition 3.4, the authors propose a method which can simply be generalized to any double optimization procedure since the idea of the proof slightly depends on the form of the optimized function. In order to restate this result here and give an exhaustive and a general proof, suppose that our estimator $\hat{\phi}$ is defined through the following double optimization procedure. Let H and H_n be two generic functions such that $H_n(\phi, \xi) \to H(\phi, \xi)$ in probability for any couple (ϕ, ξ). Define $\hat{\phi}$ and ϕ^* as follows:

$$\begin{aligned} \hat{\phi} &= \arg\inf_{\phi} \sup_{\xi} H_n(\phi, \xi); \\ \phi^* &= \arg\inf_{\phi} \sup_{\xi} H(\phi, \xi). \end{aligned}$$

We adapt the following notation:

$$\xi(\phi) = \arg\sup_{t} H(\phi, t), \qquad \xi_n(\phi) = \arg\sup_{t} H_n(\phi, t)$$

The following theorem provides sufficient conditions for consistency of $\hat{\phi}$ towards ϕ^*. This result will then be applied to the case of our estimator.
Assumptions:

A1. the estimate $\hat{\phi}$ exists (even if it is not unique);

A2. $\sup_{\xi,\phi} |H_n(\phi, \xi) - H(\phi, \xi)|$ tends to 0 in probability;

A3. for any ϕ, the supremum of H over ξ is unique and isolated, i.e. $\forall \varepsilon > 0, \forall \tilde{\xi}$ such that $\|\tilde{\xi} - \xi(\phi)\| > \varepsilon$, then there exists $\eta > 0$ such that $H(\phi, \xi(\phi)) - H(\phi, \tilde{\xi}) > \eta$;

A4. the infimum of $\phi \mapsto H(\phi, \xi(\phi))$ is unique and isolated, i.e. $\forall \varepsilon > 0, \forall \phi$ such that $\|\phi - \phi^*\| > \varepsilon$, there exists $\eta > 0$ such that $H(\phi, \xi(\phi)) - H(\phi^*, \xi(\phi^*)) > \eta$;

A5. for any ϕ in Φ, function $\xi \mapsto H(\phi, \xi)$ is continuous.

In assumption A4, we suppose the existence and uniqueness of ϕ^*. It does not, however, imply the uniqueness of $\hat{\phi}$. This is not a problem for our consistency result. The vector $\hat{\phi}$ may be any point which verifies the minimum of function $\phi \mapsto \sup_{\xi} H_n(\phi, \xi)$. Our consistency result shows that all vectors verifying the minimum of $\phi \mapsto \sup_{\xi} H_n(\phi, \xi)$

[4] or estimated by another procedure such as $k-$means.

converge to the unique vector ϕ^*. We also prove an asymptotic normality result which shows that even if $\hat{\phi}$ is not unique, all possible values should be in a neighborhood of radius $\mathcal{O}(n^{-1/2})$ centered at ϕ^*.
The following lemma establishes a uniform convergence result for the argument of the supremum over ξ of function $H_n(\phi, \xi)$ towards the one of function $H(\phi, \xi)$. It constitutes a first step towards the proof of convergence of $\hat{\phi}$ towards ϕ^*. The proof is differed to Appendix 3.7.3.

Lemma 3.4.1. *Assume A2 and A3 are verified, then*

$$\sup_{\phi} \|\xi_n(\phi) - \xi(\phi)\| \to 0, \qquad \textit{in probability.}$$

We proceed now to announce our consistency theorem. The proof is differed to Appendix 3.7.4.

Theorem 3.4.1. *Let $\xi(\phi)$ be the argument of the supremum of $\xi \mapsto H(\phi, \xi)$ for a fixed ϕ. Assume that A1-A5 are verified, then $\hat{\phi}$ tends to ϕ^* in probability.*

Let's now go back to our optimization problem (3.3.10) in order to simplify the previous assumptions. First of all, we need to specify functions H and H_n. Define function h as follows. Let $\phi = (\lambda, \theta, \alpha)$,

$$h(\phi, \xi, z) = \xi^t m(\alpha) - \frac{1}{1-\lambda}\psi\left(\xi^t g(z)\right) + \frac{\lambda}{1-\lambda}\int \psi\left(\xi^t g(x)\right) dP_1(x|\theta).$$

Functions H and H_n can now be defined through h by:

$$H(\phi, \xi) = P_T h(\phi, \xi, .), \qquad H_n(\phi, \xi) = P_n h(\phi, \xi, .).$$

In example 3.3.2, we considered the the case of the Pearson's χ^2. The supremum is infinity whenever the matrix Ω defined by (3.3.11) is s.d.p. It is thus interesting to define the *effective* set of parameters. Define the set Φ^+ by

$$\Phi^+ = \{\phi \in \Phi \text{ s.t. } \xi \mapsto H(\phi, \xi) \text{ is strictly concave}\}$$

Outside the set Φ^+, function $\xi \mapsto H(\phi, \xi)$ is not upper bounded.

Theorem 3.4.2. *Assume that A1, A4 and A5 are verified for Φ replaced by Φ^+. Suppose also that*

$$\sup_{\xi \in \mathbb{R}^{\ell+1}} \left| \int \psi\left(\xi^t g(x)\right) dP_T(x) - \frac{1}{n}\sum_{i=1}^{n} \psi\left(\xi^t g(X_i)\right) \right| \xrightarrow[\mathbb{P}]{n\to\infty} 0, \tag{3.4.1}$$

then the estimator defined by (3.3.10) is consistent.

The proof is differed to Appendix 3.7.4. Assumption A5 could be handled using Lebesgue's continuity theorem if one finds a P_T-integrable function $\tilde{h}$ such that $|\psi\left(\xi^t g(z)\right)| \leq \tilde{h}(z)$. This is, however, not possible in general unless we restrain ξ to a compact set. Otherwise, we need to verify this assumption according the situation we have in hand, see example 3.4.1 below for more details. The uniform limit (3.4.1) can be treated according to the divergence and the constraints which we would like to impose. A general method is to prove that the class of functions $\{x \mapsto \psi\left(\xi^t g(x)\right), \xi \in \mathbb{R}^{\ell+1}\}$ is a Glivenko-Cantelli class of functions, see van der Vaart [1998] Chap. 19 Section 2 and the examples therein for some possibilities.

Remark 3.4.1. Under suitable differentiability assumptions, the set Φ^+ defined earlier can be rewritten as:

$$\Phi^+ = \Phi \cap \left\{\phi : \quad J_{H(\phi,.)} \text{ is definite negative}\right\},$$

where $J_{H(\phi,.)}$ is the Hessian matrix of function $\xi \mapsto H(\phi, \xi)$ and is given by:

$$J_{H(\phi,.)} = -\int g(x)g(x)^t \psi''(\xi^t g(x)) \left(\frac{1}{1-\lambda} dP_T - \frac{\lambda}{1-\lambda} dP_1\right)(x). \tag{3.4.2}$$

The problem with using the set Φ^+ is that if we take a point ϕ in the interior of Φ, there is no guarantee that it would be an interior point of Φ^+. This will impose more difficulties in the proof of the asymptotic normality. We prove in the next proposition that this is however true for ϕ^*. Besides, the set Φ^+ is open as soon as $\int \Phi$ is not void. The proof is differed to Appendix 3.7.6.

Proposition 3.4.1. *Assume that function $\xi \mapsto H(\phi, \xi)$ is of class $\mathcal{C}^2$ for any $\phi \in \Phi^+$. Suppose that ϕ^* is an interior point of Φ, then there exists a neighborhood $\mathcal{V}$ of ϕ^* such that for any $\phi \in \mathcal{V}$, $J_{H(\phi,.)}$ is definite negative and thus $\xi(\phi)$ exists and is finite. Moreover, function $\phi \mapsto \xi(\phi)$ is continuously differentiable on $\mathcal{V}$.*

Corollary 3.4.1. *Assume that function $\xi \mapsto H(\phi, \xi)$ is of class $\mathcal{C}^2$ for any $\phi \in \Phi^+$. If Φ is bounded, then there exists a compact neighborhood $\bar{\mathcal{V}}$ of ϕ^* such that $\xi(\bar{\mathcal{V}})$ is bounded and $\{x \mapsto \psi\left(\xi^t g(x)\right), \xi \in \xi(\bar{\mathcal{V}})\}$ is a Glivenko-Cantelli class of functions.*

Proof. The first part of the corollary is an immediate result of Proposition 3.4.1 and the continuity of function $\phi \mapsto \xi(\phi)$ over $\text{int}(\Phi^+)$. The implicit functions theorem permits to conclude that $\xi(\phi)$ is continuously differentiable over $\text{int}(\Phi^+)$. The second part is an immediate result of Example 19.7 page 271 from van der Vaart [1998]. □

This corollary suggests that in order to prove the consistency of $\hat{\phi}$, it suffices to restrict the values of ϕ on Φ^+ and the values of ξ on $\xi(\text{int}(\Phi^+))$ in the definition of $\hat{\phi}$ (3.3.10). Besides, since $\{x \mapsto \psi\left(\xi^t g(x)\right), \xi \in \xi(\text{int}(\Phi^+))\}$ is a Glivenko-Cantelli class of functions, the uniform limit (3.4.1) is verified by the Glivenko-Cantelli theorem.

Remark 3.4.2. There is a great difference between the set Φ^+ where Ω is s.d.p. (J_H is s.d.n.) and the set where only $\frac{1}{1-\lambda} dP_T - \frac{\lambda}{1-\lambda} dP_1$ is a probability measure. Indeed, there is a strict inclusion in the sense that if $\frac{1}{1-\lambda} dP_T - \frac{\lambda}{1-\lambda} dP_1$ is a probability measure, then Ω is s.d.p., but the inverse is not right. Figure (3.3) shows this difference. Furthermore, it is clearly simpler to check for a vector ϕ if the matrix Ω is s.d.p. It suffices to calculate the integral[5] (even numerically) and then use some rule such as Sylvester's rule to check if it is definite negative, see the example below. However, in order to check if the measure $\frac{1}{1-\lambda} dP_T - \frac{\lambda}{1-\lambda} dP_1$ is positive, we need to verify it on all $\mathbb{R}^r$.

Remark 3.4.3. The previous remark shows the interest of adapting a methodology based on signed measures and not only positive ones. We have a larger and better space to search inside for the triplet $(\lambda^*, \theta^*, \alpha^*)$. For example, in figure (3.3), the optimization algorithm which tries to solve (3.3.10) gets stuck if we only search in the set of parameters for which $\frac{1}{1-\lambda} dP_T - \frac{\lambda}{1-\lambda} dP_1$ is a probability measure. This does not happen if we search in the set Φ^+. Moreover, even if the algorithm returns a triplet $(\hat{\lambda}, \hat{\theta}, \hat{\alpha})$ for which the semiparametric

[5] If function g is a polynomial, i.e. moment constraints, then the integral is a mere subtractions between the moments of P_T and the ones of P_1.

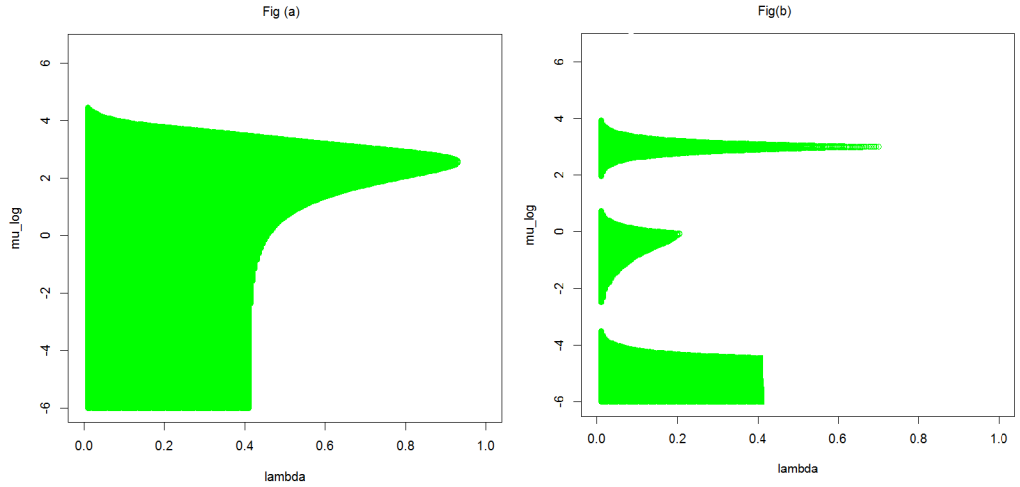

Figure 3.3: Differences between the set where $\frac{1}{1-\lambda}dP_T - \frac{\lambda}{1-\lambda}dP_1$ is positive (Fig (b)) and the set Φ^+ (Fig (a)) in a Weibull–Lognormal mixture.

component $P_0 = \frac{1}{1-\lambda}dP_T - \frac{\lambda}{1-\lambda}dP_1$ is not a probability measure, it should not mean that the procedure failed. This is because we are looking for the parameters and not to estimate P_0. Besides, it is still possible to threshold the negative values from the density and then regularize in order to integrate to one.

Example 3.4.1 (χ^2 case)**.** Consider the case of a two-component semiparapetric mixture model where P_0 is defined through its first three moments. In other words, the set of constraints $\mathcal{M}_\alpha$ is given by:

$$\mathcal{M}_\alpha = \left\{ Q : \int dQ(x) = 1,\ \int x dQ(x) = m_1(\alpha),\ \int x^2 dQ(x) = m_2(\alpha),\ \int x^3 dQ(x) = m_3(\alpha) \right\}.$$

We have already seen in example 3.3.2 that if $\psi(t) = t^2/2 + t$, the Pearson's χ^2 convex conjugate, then the optimization over ξ can be solved and the solution is given by:

$$\xi(\phi) = \Omega^{-1}\left(m(\alpha) - \int g(x)\left(\frac{1}{1-\lambda}dP(x) - \frac{\lambda}{1-\lambda}dP_1(x|\theta) \right) \right),\ \text{for } \phi \in \Phi^+.$$

Let M_i denotes the moment of order i of P_T. Denote also $M_i^{(1)}(\theta)$ the moment of order i of the parametric component $P_1(.|\theta)$.

$$M_i = \mathbb{E}_{P_T}[X^i], \qquad M_i^{(1)}(\theta) = \mathbb{E}_{P_1(.|\theta)}[X^i].$$

A simple calculus shows that:

$$\begin{aligned} \Omega &= \int g(x)g(x)^t \left(\frac{1}{1-\lambda}dP(x) - \frac{\lambda}{1-\lambda}dP_1(x|\theta) \right) \\ &= \left[\frac{1}{1-\lambda}M_{i+j-2} - \frac{\lambda}{1-\lambda}M_{i+j-2}^{(1)}(\theta) \right]_{i,j\in\{1,\cdots,4\}}. \end{aligned}$$

The solution holds for any $\phi \in \text{int}(\Phi^+)$. Continuity assumption A5 over $\xi \mapsto H(\phi, \xi)$ is

simplified here because function H is a polynomial of degree 2. We have:

$$H(\phi,\xi) = \xi^t m(\alpha) - \left[\frac{1}{2}\xi_1^2 + \xi_1 + (\xi_1\xi_2+\xi_2)\left(\frac{1}{1-\lambda}M_1 - \frac{\lambda}{1-\lambda}M_1^{(1)}(\theta)\right)\right.$$
$$+(\xi_2^2/2+\xi_1\xi_2+\xi_3)\left(\frac{1}{1-\lambda}M_2 - \frac{\lambda}{1-\lambda}M_2^{(1)}(\theta)\right)+(\xi_1\xi_4+\xi_2\xi_3+\xi_4)\left(\frac{1}{1-\lambda}M_3 - \frac{\lambda}{1-\lambda}M_3^{(1)}(\theta)\right)$$
$$+(\xi_3^2/2+\xi_2\xi_4)\left(\frac{1}{1-\lambda}M_4 - \frac{\lambda}{1-\lambda}M_4^{(1)}(\theta)\right)+\xi_3\xi_4\left(\frac{1}{1-\lambda}M_5 - \frac{\lambda}{1-\lambda}M_5^{(1)}(\theta)\right)$$
$$\left.+\xi_4^2/2\left(\frac{1}{1-\lambda}M_6 - \frac{\lambda}{1-\lambda}M_6^{(1)}(\theta)\right)\right].$$

Regularity of function $\phi \mapsto \xi(\phi)$ is directly tied by the regularity of the moments of $P_1(.|\theta)$ with respect to θ. If $M_i^{(1)}$ is continuous with respect to θ and $m(\alpha)$ is continuous with respect to α, then the existence of ϕ^* becomes immediate as soon as the set Φ is compact. If ϕ^* is an interior point of Φ, then Proposition 3.4.1 and Corollary 3.4.1 apply. Thus $\text{int}(\Phi^+)$ is non void and the class $\{x \mapsto \psi\left(\xi^t g(x)\right), \xi \in \xi(\text{int}(\Phi^+))\}$ is a Glivenko-Cantelli class of functions. Assumption A4 remains specific to the model we consider.
The previous calculus shows that our procedure for estimating $\hat{\phi}$ can be done efficiently and the complexity of the calculus does not depend on the dimension of the data. Besides, no numerical integration is needed.

3.4.2 Asymptotic normality

We will suppose that the model p_ϕ is $\mathcal{C}^2(\text{int}(\Phi^+))$ and that ψ is $\mathcal{C}^2(\mathbb{R})$. In order to simplify the formula below, we suppose that $\psi'(0) = 1$ and $\psi''(0) = 1$. These are not restrictive assumptions and can be relaxed. Recall that they are both verified in the class of Cressie-Read functions (1.1.3).
Define the following matrices:

$$J_{\phi^*,\xi^*} = \left(\frac{1}{(1-\lambda^*)^2}\left[-\mathbb{E}_{P_T}\left[g(X)\right] + \mathbb{E}_{P_1(.|\theta^*)}\left[g(X)\right]\right], \frac{\lambda^*}{1-\lambda^*}\int g(x)\nabla_\theta p_1(x|\theta^*)dx, \nabla m(\alpha^*)\right) \quad (3.4.3)$$
$$J_{\xi^*,\xi^*} = \mathbb{E}_{P_0^*}\left[g(X)g(X)^t\right]; \quad (3.4.4)$$
$$\Sigma = \left(J_{\phi^*,\xi^*}^t J_{\xi^*,\xi^*} J_{\phi^*,\xi^*}\right)^{-1}; \quad (3.4.5)$$
$$H = \Sigma J_{\phi^*,\xi^*}^t J_{\xi^*,\xi^*}^{-1}; \quad (3.4.6)$$
$$W = J_{\xi^*,\xi^*}^{-1} - J_{\xi^*,\xi^*}^{-1} J_{\phi^*,\xi^*}\Sigma J_{\phi^*,\xi^*}^t J_{\xi^*,\xi^*}^{-1}. \quad (3.4.7)$$

Recall the definition of Φ^+ and define similarly the set Φ_n^+

$$\Phi^+ = \left\{\phi : \int g(x)g(x)^t\left(\frac{1}{1-\lambda}dP_T - \frac{\lambda}{1-\lambda}dP_1\right)(x|\theta) \text{ is s.p.d.}\right\}; \quad (3.4.8)$$
$$\Phi_n^+ = \left\{\phi : \frac{1}{n}\frac{1}{1-\lambda}\sum_{i=1}^n g(X_i)g(X_i)^t - \frac{\lambda}{1-\lambda}\int g(x)g(x)^t dP_1(x|\theta) \text{ is s.p.d.}\right\} \quad (3.4.9)$$

These two sets are the feasible sets of parameters for the optimization problems (3.3.9) and (3.3.10) respectively. In other words, outside of the set Φ^+, we have $H(\phi,\xi(\phi)) = \infty$. Similarly, outside of the set Φ_n^+, we have $H_n(\phi,\xi_n(\phi)) = \infty$.

Theorem 3.4.3. *Suppose that:*

1. *$\hat{\phi}$ is consistent and $\phi^* \in$ int(Φ);*

2. *the function* $\alpha \mapsto m(\alpha)$ *is* $\mathcal{C}^2$*;*

3. $\forall \phi \in B(\phi^*, \tilde{r})$ *and any* $\xi \in \xi(B(\phi^*, \tilde{r}))$*, there exist functions* $h_{1,1}, h_{1,2} \in L^1(p_1(.|\theta))$ *such that* $\left\|\psi'\left(\xi^t g(x)\right) g(x)\right\| \leq h_{1,1}(x)$ *and* $\left\|\psi''\left(\xi^t g(x)\right) g(x) g(x)^t\right\| \leq h_{1,2}(x)$*;*

4. $\forall \xi \in \xi(B(\phi^*, \tilde{r}))$*, there exist functions* $h_{2,1}, h_{2,2} \in L^1(dx)$ *such that* $\left\|\psi\left(\xi^t g(x)\right) \nabla_\theta p_1(x|\theta)\right\| \leq h_2(x)$ *and* $\left\|\psi\left(\xi^t g(x)\right) J_{p_1(.|\theta)}\right\| \leq h_2(x)$*;*

5. *for any couple* $(\phi, \xi) \in B(\phi^*, \tilde{r}) \times \xi(B(\phi^*, \tilde{r}))$*, there exists a function* $h_3 \in L^1(dx)$ *such that* $\left\|\psi'\left(\xi^t g(x)\right) g(x) \nabla_\theta p_1(x|\theta)^t\right\| \leq h_3(x)$*;*

6. *finite second order moment of* g *under* P_T*, i.e.* $\mathbb{E}_{P_T}\left[g_i(X) g_j(X)\right] < \infty$ *for* $i, j \leq \ell$*;*

7. *matrices* J_{ξ^*,ξ^*} *and* $J^t_{\phi^*,\xi^*} J_{\xi^*,\xi^*} J_{\phi^*,\xi^*}$ *are invertible,*

then

$$\begin{pmatrix} \sqrt{n}\left(\hat{\phi} - \phi^*\right) \\ \sqrt{n}\xi_n(\hat{\phi}) \end{pmatrix} \underset{\mathcal{L}}{\to} \mathcal{N}\left(0, \frac{1}{(1-\lambda^*)^2} \begin{pmatrix} H \\ W \end{pmatrix} Var_{P_T}(g(X)) \begin{pmatrix} H^t & W^t \end{pmatrix}\right),$$

where H *and* P *are given by formulas (3.4.6) and (3.4.7).*

The proof is differed to Appendix 3.7.7. Assumption 3 entails the differentiability of function $H_n(\xi, \phi)$ up to second order with respect to ξ whatever the value of ϕ in a neighborhood of ϕ^*. Assumption 4 entails the differentiability of function $H_n(\xi, \phi)$ up to second order with respect to θ in a neighborhood of θ^* inside $\xi(B(\phi^*, \tilde{r}))$. Finally, assumption 5 implies the cross-differentiability of function $H_n(\xi, \phi)$ with respect to ξ and θ.
Differentiability assumptions in Theorem 3.4.3 can be relaxed in the case of the Pearson's χ^2 since all integrals in functions H_n and H can be calculated. Our result covers the general case and thus we need to ensure differentiability of the integrals using Lebesgue theorems which requires the existence of integrable functions which upperbound the integrands.

3.5 Simulation study

We perform several simulations in univariate and multivariate situations and show how prior information about the moments of the distribution of the semiparametric component P_0 can help us better estimate the set of parameters $(\lambda^*, \theta^*, \alpha^*)$ in regular examples, i.e. the components of the mixture can be clearly distinguished when we plot the probability density function. We also show how our approach permits to estimate even in difficult situations when the proportion of the parametric component is very low; such cases could *not* be estimated using existing methods.
Another important problem in existing methods is their quadratic complexity. For example, an EM-type method such as Robin et al. [2007]'s algorithm or its stochastic version introduced by Bordes et al. [2007] performs n^2+3n operations in order to complete a single iteration. An EM-type algorithm for semiparametric mixture models needs in average 100 iterations to converge and may attain 1000 iterations[6] for each sample. To conclude, the estimation procedure performs at least $100(n^2+3n)$ operations. In a signal-noise situations where the signal has a very low proportion around 0.05, we need a greater number of observations say $n = 10^5$. Such experiences cannot be performed using an EM-type method

[6] This was the case of the Weibull mixture.

such as Robin et al. [2007]'s algorithm or its stochastic version introduced by Bordes et al. [2007] unless one has a "super computer". The method of Bordes and Vandekerkhove [2010] shares similar complexity[7] $\mathcal{O}(n^2)$. Last but not least, the EM-type method of Song et al. [2010] and their $\pi-$maximizing one have the advantage over other methods, because we need only to calculate a kernel density estimator once and for all, then use it at each iteration[8]. Nevertheless, the method has still a complexity of order n^2.
Our approach, although has a double optimization procedure, it can be implemented when g is polynomial and φ corresponds to the Pearson's χ^2 in a way that it has a linear complexity $\mathcal{O}(n)$. First of all, using the χ^2 divergence, the optimization over ξ in (3.3.10) can be calculated directly. On the other hand, all integrals are mere calculus of empirical moments and moments of the parametric part, see Example 3.4.1. Empirical moments can be calculated once and for all whereas moments of the parametric part can be calculated using direct formulas available for a large class of probability distributions. What remains is the optimization over ϕ. In the simulations below, our method produced the estimates instantly even for a number of observations of order 10^7 whereas other existing methods needed from several hours (algorithms of Song et al. [2010]) to several days (for other algorithms). It is however important to notice that if the number of constraints is large enough, say a function of n, then we no longer have a linear complexity.
Because of the very long execution time of existing methods, we restricted the comparison to simulations in regular situations with $n < 10^4$. Experiments with greater number of observations were only treated using our method and the methods in Song et al. [2010]. In all tables presented hereafter, we performed 100 experiments and calculated the average of resulting estimators. We provided also the standard deviation of the 100 experiments in order to get a measure of preference in case the different estimation methods gave close results.
Our experiments cover the following models:

- A two-component Weibull mixture;
- A two-component Weibull - Lognormal mixture;
- A two-component Gaussian – Two-sided Weibull mixture;
- A two-component bivariate Gaussian mixture.

We apply the several estimation methods from Section 3.1. We have chosen a variety of values for the parameters especially the proportion. The second model may represent a problem from queue theory where the left component represents the impatient customers whereas the right component represents the regular customers. The third model stems from a signal-noise application where the signal is centered at zero whereas the noise is repartitioned at both sides. The fourth model appears in clustering and is only presented to show how our method performs in multivariate contexts.
In all our experiments, no numerical integration was used since they can be easily calculated as functions of the empirical moments of the data and the moments of the parametric component, see Example 3.4.1. Simulations were done using the R Core Team [2015]. Optimization was performed using the Nelder-Mead algorithm, see Nelder and Mead [1965]. For the $\pi-$maximizing algorithm of Song et al. [2010], we used the Brent's method because

[7] we need more than 24 hours to estimate the parameters of one sample with 10^5 observations.
[8] We were able to perform simulations with $n = 10^5$ observations but needed about 5 days on an i7 laptop clocked at 2.5 GHz with 8GB of RAM. For Robin et al. [2007]'s algorithm, a few iterations took about one day. One can imagine the time needed to estimate 100 samples with 10^5 observations in each sample.

the optimization was carried over one parameter.
For our procedure, we only used the χ^2 divergence, because the optimization over ξ can be calculated without numerical methods[9]. Recall that the optimized function over ξ is not always strictly concave and the Hessian matrix may be definite positive, see remark 3.4.1. It is thus important to check for each vector $\phi = (\lambda, \theta, \alpha)$ if the Hessian matrix is still definite negative for example using Sylvester's criterion. If it is not, we set the objective function to a value such as 10^2. Besides, since the resulting function $\phi \mapsto H_n(\phi, \xi_n(\phi))$ as a function of ϕ is not ensured to be strictly convex, we used 10 random initial feasible points inside the set Φ_n^+ defined by (3.4.9). We then ran the Nelder-Mead algorithm and chose the vector of parameters for which the objective function has the lowest value. We applied a similar procedure on the algorithm of Bordes and Vandekerkhove [2010] in order to ensure a *good and fair* optimization.

Remark 3.5.1. In the literature on the stochastic EM algorithm, it is advised that we iterate the algorithm for some time until it reaches a stable state, then continue iterating long enough and average the values obtained in the second part. The trajectories of the algorithm were very erratic especially for the estimation of the proportion. For us, we iterated for the stochastic EM-type algorithm of Bordes et al. [2007] 5000 times and averaged the 4000 final iterations.

Remark 3.5.2. Initialization of both the EM-type algorithm of Song et al. [2010] and the SEM-type algorithm of Bordes et al. [2007] was not very important, and we got the same results when the vector of weights was initialized uniformly or in a "good" way. The method of Robin et al. [2007] was more influenced by such initialization and we used most of the time a good starting points.

Remark 3.5.3. For the methods of Song et al. [2010], we need to estimate mixture's distribution using a kernel density estimator. For the data generated from a Weibull mixture and the data generated from a Weibull Lognormal mixture, we used a reciprocal inverse Gaussian kernel density estimator with a window equal to 0.01 according to our simulations in Chapter 1.

Remark 3.5.4. Matrix inversion was done manually using direct inversion methods, because the function `solve` in the statistical program R produced errors sometimes because the matrix was highly sensible at some point during the optimization. For matrices of dimension 4×4 and 5×5 we used block matrix inversion, see for example Lu and Shiou [2002]. The inverse of a 3×3 was calculated using a direct formula.

3.5.1 Data generated from a two-component Weibull mixture modeled by a semiparametric Weibull mixture

We consider a mixture of two Weibull components with scales $\sigma_1 = 0.5, \sigma_2 = 1$ and shapes $\nu_1 = 2, \nu_2 = 1$ in order to generate the dataset. In the semiparametric mixture model, the parametric component will be "the one to the right", i.e. the component whose true set of parameters is $(\nu_1 = 2, \sigma_1 = 0.5)$. We illustrate several values of the proportion $\lambda \in \{0.7, 0.3\}$, see figure (3.4). This constitutes a difficult example for both our method and existing methods such as EM-type methods or the $\pi-$maximizing algorithm of Song et al. [2010]. We therefore, simulated 10000 samples and fixed both scales during estimation. We estimate the proportion and the shapes of both components. For our method, the

[9] We noticed no great difference when using a Hellinger divergence.

variance of the estimator of ν_1 was high and we needed to use 4 moments to reduce it to an acceptable range. Of course, as the number of observations increases, the variance reduces. We, however, avoided greater number of observations because methods such as Robin et al. [2007] need very long execution time for even one sample. The method of Bordes and Vandekerkhove [2010] cannot be applied here since the support of the mixture is $\mathbb{R}_+$.
Moments of the Weibull distribution are given by:

$$\mathbb{E}[X^i] = \sigma^i \Gamma(1 + i/\nu), \qquad \forall i \in \mathbb{N}.$$

For our method , function g is a vector of polynomials that is $g(x) = (1, x, x^2, x^3)^t$ when we use only 3 moment constraints. It suffices to add x^4 when the number of constraints is 4. On the other hand, $m(\alpha) = (1, \sigma\Gamma(1 + 1/\nu), \sigma^2\Gamma(1 + 2/\nu), \sigma^3\Gamma(1 + 3/\nu))$. The results of our method are clearly better than existing methods which practically failed and could not see but one main component with shape in between the two shapes, see table (3.1). Although our method presents an inconvenient greater variance for ν_1, the Monte-Carlo mean of the hundred experiences is still unbiased. We believe that the use of other types of constraints would have resulted in better results without the need to add one more constraint.

Nb of observations	λ	sd(λ)	ν_1	sd(ν_1)	ν_2	sd(ν_2)
Mixture 1 : $n = 10^4$ $\lambda^* = 0.7$, $\nu_1^* = 2$, $\sigma_1^* = 0.5$(fixed), $\nu_2^* = 1$, $\sigma_2^* = 1$(fixed)						
Pearson's χ^2 3 moments	0.700	0.010	2.006	0.217	1.005	0.024
Pearson's χ^2 4 moments	0.701	0.010	2.014	0.086	1.013	0.024
Robin	0.654	0.101	1.591	0.085	—	—
Song EM-type	0.907	0.004	1.675	0.020	—	—
Song $\pi-$maximizing	0.782	0.006	1.443	0.012	—	—
Mixture 1 : $n = 10^4$ $\lambda^* = 0.3$, $\nu_1^* = 2$, $\sigma_1^* = 0.5$(fixed), $\nu_2^* = 1$, $\sigma_2^* = 1$(fixed)						
Pearson's χ^2 3 moments	0.304	0.016	2.191	0.887	0.998	0.013
Pearson's χ^2 4 moments	0.303	0.016	2.120	0.285	1.001	0.013
Robin	0.604	0.029	1.256	0.037	—	—
Song EM-type	0.806	0.005	1.185	0.018	—	—
Song $\pi-$maximizing	0.624	0.007	1.312	0.013	—	—

Table 3.1: The mean value with the standard deviation of estimates in a 100-run experiment on a two-component Weibull mixture.

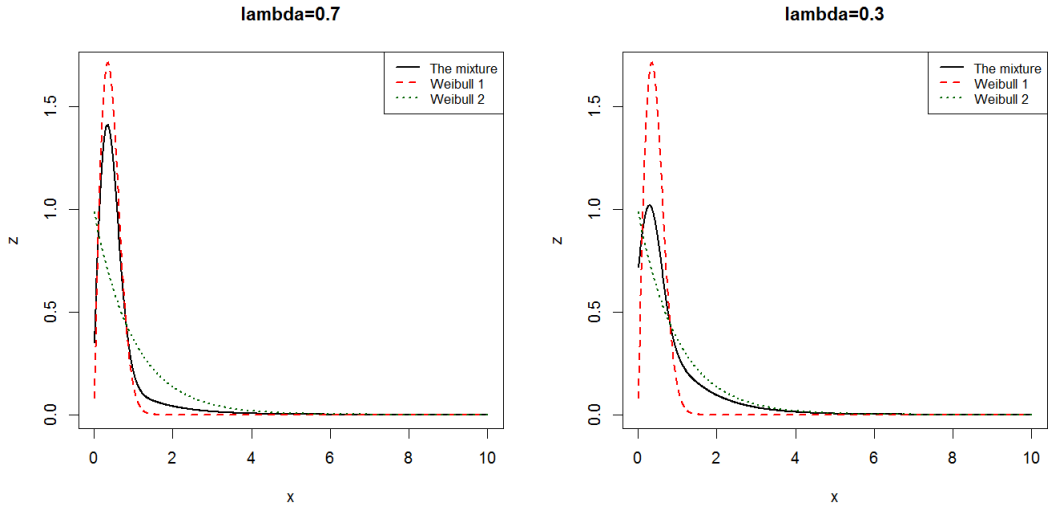

Figure 3.4: The Weibull mixtures, see table (3.1)

3.5.2 Data generated from a two-component Weibull-LogNormal mixture modeled by a semiparametric Weibull-LogNormal mixture

We consider a mixture of two components; a Weibull component with scale = 1 and shape = 1.5, and a log-normal component with meanlog = 3 and scale = 0.5, see figure (3.5) for the considered cases. In the results of table (3.2), the parametric part is considered to be the Lognormal distribution. In the results of table (3.3), the parametric part of the semiparametric model is considered to be the Weibull distribution.
The number of observations for each mixture depends on its subtlety. As the proportion of the parametric component becomes lower, we needed more observations to produce reasonable estimates. We chose the number of observations in a way that the standard deviation of estimated parameters does not exceed 1.
We used the first three moments such that we can estimate three parameters; the proportion of the parameteric component, the shape of the Weibull component and the mean-parameter of the Lognormal component. Thus, the scales of both components are supposed to be known during the estimation procedure. The moments of these two distributions are given by:

$$\begin{aligned} \text{Weibull:} \quad & \mathbb{E}[X^i] &&= \sigma^i \Gamma(1 + i/\nu); \\ \text{Lognormal:} \quad & \mathbb{E}[X^i] &&= e^{i\mu + i^2\sigma^2/2}. \end{aligned}$$

For our method , function g is a vector of polynomials that is $g(x) = (1, x, x^2, x^3)^t$. Function $m(\alpha)$ is given by $(1, \mathbb{E}[X], \mathbb{E}[X^2], \mathbb{E}[X^3])^t$ with the corresponding moments according to whether we consider the Weibull or the log-normal as the semiprametric component.
Our new method seems to produce high variance of the shape of the Weibull component. This should not be surprising, because the part which influences on the moments of the model is the Lognormal component. Its moments have an exponential form and small differences in the mean-parameter could compensate for a great differences in the shape of the Weibull component. The results are still satisfactory since we get to estimate an information of the semiparametric component at a great precision together with the proportion.

Nb of observations	λ	sd(λ)	μ	sd(μ)	ν	sd(ν)
Mixture 1 : $\lambda^* = 0.7$, $\mu^* = 3$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 1.5$, $\sigma_1^* = 1$(fixed)						
$n = 10^2$	0.384	0.117	2.654	0.153	0.488	0.018
$n = 10^3$	0.518	0.068	2.806	0.099	0.473	0.014
$n = 10^4$	0.605	0.044	2.903	0.069	0.531	0.326
$n = 10^5$	0.651	0.030	2.957	0.041	0.809	0.630
$n = 10^6$	0.682	0.018	2.979	0.022	1.638	0.813

Table 3.2: The mean value with the standard deviation of estimates produced by our procedure with three moments constraints in a 100-run experiment on a two-component Weibull– Lognormal mixture. The parametric component is the Lognormal distribution.

Nb of observations	λ	sd(λ)	ν	sd(ν)	μ	sd(μ)
Mixture 1 : $n = 10^3$, $\lambda^* = 0.3$, $\nu^* = 1.5$, $\sigma_1^* = 1$(fixed), $\mu^* = 3$, $\sigma_2^* = 0.5$(fixed)						
Pearson's χ^2	0.308	0.017	1.484	0.624	3.002	0.026
Robin	0.296	0.015	1.557	0.068	—	—
Song EM-type	0.291	0.015	1.614	0.087	—	—
Song $\pi-$maximizing	0.230	0.022	1.662	0.251	—	—
SEM	0.284	0.041	1.570	0.263	—	—
Mixture 2 : $n = 10^4$, $\lambda^* = 0.1$, $\nu^* = 1$, $\sigma_1^* = 1$(fixed), $\mu^* = 3$, $\sigma_2^* = 0.5$(fixed)						
Pearson's χ^2	0.103	0.006	1.284	0.677	3.001	0.007
Robin	0.095	0.003	1.049	0.031	—	—
Song EM-type	0.100	0.004	0.894	0.039	—	—
Song $\pi-$maximizing	0.085	0.005	1.024	0.055	—	—
SEM	0.094	0.015	1.054	0.228	—	—
Mixture 3 : $n = 10^4$, $\lambda^* = 0.05$, $\nu^* = 1$, $\sigma_1^* = 1$(fixed), $\mu^* = 3$, $\sigma_2^* = 0.5$(fixed)						
Pearson's χ^2	0.052	0.004	1.312	0.703	3.001	0.006
Song EM-type	0.050	0.003	0.855	0.068	—	—
Song $\pi-$maximizing	0.042	0.003	1.013	0.052	—	—
Mixture 4 : $n = 5 \times 10^4$, $\lambda^* = 0.05$, $\nu^* = 0.4$, $\sigma_1^* = 1$(fixed), $\mu^* = 3$, $\sigma_2^* = 0.5$(fixed)						
Pearson's χ^2	0.049	0.002	0.629	0.438	3.001	0.004
Song EM-type	0.064	0.001	0.345	0.004	—	—
Song $\pi-$maximizing	0.024	0.001	0.773	0.010	—	—

Table 3.3: The mean value with the standard deviation of estimates in a 100-run experiment on a two-component Weibull-log normal mixture. The parametric component is the Weibull distribution.

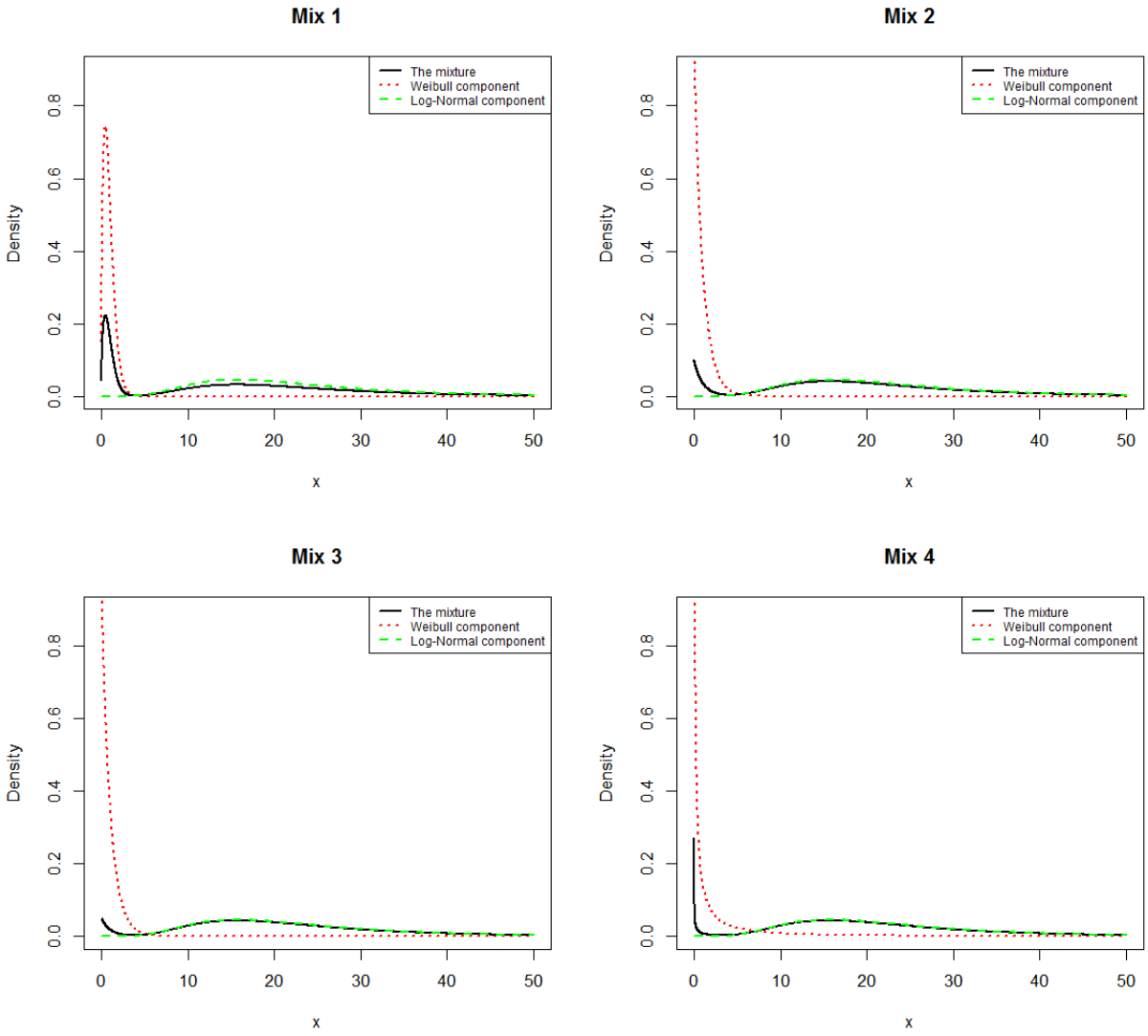

Figure 3.5: The Weibull – Lognormal mixtures, see tables (3.2,3.3)

3.5.3 Data generated from a two-sided Weibull Gaussian mixture modeled by a semiparametric two-sided Weibull Gaussian mixture

The (symmetric) two-sided Weibull distribution can be considered as a generalization of the Laplace distribution and can be defined through either its density or its distribution function as follows:

$$f(x|\nu,\sigma) = \frac{1}{2}\frac{\sigma}{\nu}\left(\frac{|x|}{\sigma}\right)^{\nu-1} e^{-\left(\frac{|x|}{\sigma}\right)^{\nu}}, \qquad \mathbb{F}(x|\nu,\sigma) = \begin{cases} 1-\frac{1}{2}e^{-\left(\frac{x}{\sigma}\right)^{\nu}} & x \geq 0 \\ e^{-\left(\frac{-x}{\sigma}\right)^{\nu}} & x < 0 \end{cases}$$

We can also define a skewed form of the two-sided Weibull distribution by attributing different scale and shape parameters to the positive and the negative parts, and then normalizing in a suitable way so that $f(x)$ integrates to one; see Chen and Gerlach [2013]. The moments of the symmetric two-sided Weibull distribution we consider here are given by:

$$\begin{aligned} \mathbb{E}[X^{2k}] &= \sigma^{2k}\Gamma(1+2k/\nu) \\ \mathbb{E}[X^{2k+1}] &= 0, \forall k \in \mathbb{N}. \end{aligned}$$

We simulate different samples from a two-component mixture with a parametric component f_1 a Gaussian $\mathcal{N}(\mu = 0, \sigma = 0.5)$ and a semiparametric component f_0 a (symmetric) two-sided Weibull distribution with parameters $\nu \in \{3, 2.5, 1.5\}$ and a scale $\sigma_0 \in \{1.5, 2\}$, see figure (3.6) for different choices of the proportion. We perform different experiments to estimate the proportion and the mean of the parametric part (the Gaussian) and the shape of the semiparametric component. The values of the scale of the two components are considered to be known during estimation. We consider the following two sets of constraints:

$$\begin{aligned} \mathcal{M}_{1:3} &= \left\{ f_0 : \int_{\mathbb{R}} f_0(x)dx = 1, \mathbb{E}_{f_0}[X] = 0, \mathbb{E}_{f_0}[X^2] = \sigma_0^2\Gamma(1+2/\nu), \mathbb{E}_{f_0}[X^3] = 0, \nu > 0 \right\}; \\ \mathcal{M}_{2:4} &= \left\{ f_0 : \int_{\mathbb{R}} f_0(x)dx = 1, \mathbb{E}_{f_0}[X^2] = \sigma_0^2\Gamma(1+2/\nu), \mathbb{E}_{f_0}[X^3] = 0, \right. \\ &\qquad \left. \mathbb{E}_{f_0}[X^4] = \sigma_0^4\Gamma(1+4/\nu), \nu > 0 \right\}. \end{aligned}$$

The first set imposes that the semiparametric component is centered around zero whereas the second one does not impose it.
The first set of constraints is not really suitable for estimation especially when the number of observations is high enough. The reason is simple and is based on the original idea behind our procedure, see paragraph 3.3.2. The first and the third moment constraints are practically the same constraint. Indeed, the number of models of the form $\frac{1}{1-\lambda}f(.) - \frac{\lambda}{1-\lambda}f_1(x|\theta)$ verifying the constraints of $\mathcal{M}_{1:3}$ is infinite because the first and the third constraints give rise to the following equations:

$$\begin{aligned} \lambda\mu &= 0 \\ \lambda\mu\left(\mu^2 + 3\sigma_1^2\right) &= 0. \end{aligned}$$

The zero in the right hand side comes from the fact that the first and the third true moments of the whole mixture are zero. These are two equations in λ and μ (since σ_1 is supposed to be known) with infinite number of solutions $(\mu, \lambda) \in \{0\} \times [0, 1]$. This entails

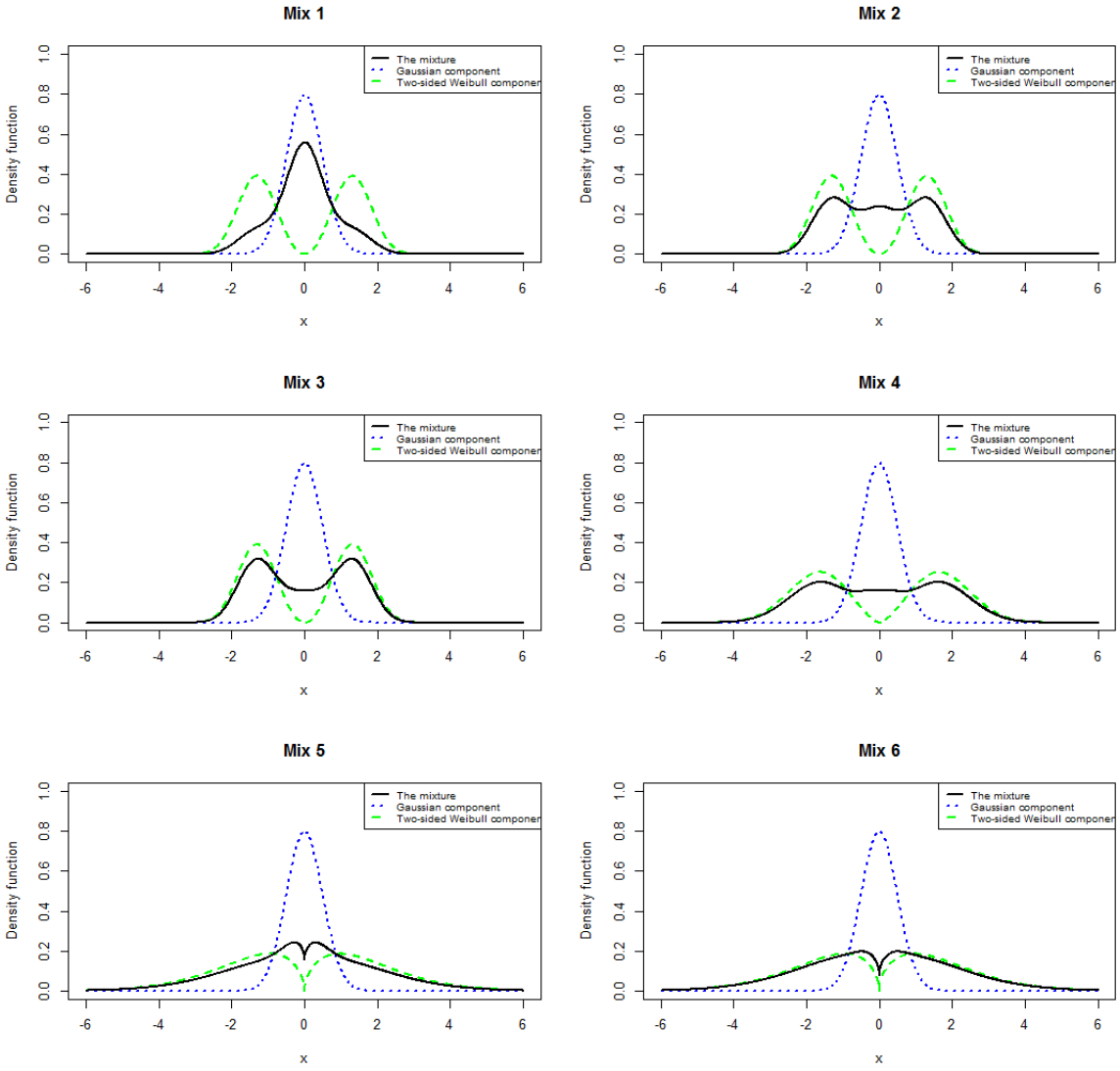

Figure 3.6: Mixtures of two-sided Weibull – Gaussian with low and high proportion of the parametric part. See table (3.4)

that theoretically, there is an infinite number of models of the form $\frac{1}{1-\lambda}f(.) - \frac{\lambda}{1-\lambda}f_1(x|\theta)$ in the intersection $\mathcal{N} \cap M_{1:3}$. Still, the empirical version of these equations is

$$\begin{aligned} \lambda\mu &= \frac{1}{n}\sum_{i=1}^{n} X_i; \\ \lambda\mu\left(\mu^2 + 3\sigma_1^2\right) &= \frac{1}{n}\sum_{i=1}^{n} X_i^3. \end{aligned}$$

As the number of observations is very small, the right hand side of both equations is biased enough from zero and it is highly possible that the number of solutions becomes not only finite but reduced to one. As the number of observations increases, the law of large numbers implies directly that the right hand side becomes arbitrarily close to zero and the set of solutions becomes infinite. This is exactly what happened in the simulation results in table (3.4) below. The algorithm favored the value zero for the estimate of

the proportion as the true proportion of the parametric component became close to zero, whereas the estimates of the mean took values very dispersed centered around zero but with a high standard deviation. The set of constraints $\mathcal{M}_{2:4}$ gave clear better results even for very low proportions. On the other hand, our method outperforms other semiparametric algorithms without prior information especially when the proportion of the parameteric component is low. This shows once more the interest of incorporating a prior information in the estimation procedure.

Estimation method	λ	sd(λ)	μ	sd(μ)	ν	sd(ν)
Mixture 1 :$n = 100$ $\lambda^* = 0.7$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 3$, $\sigma_1^* = 1.5$(fixed)						
Pearson's χ^2 under $\mathcal{M}_{1:3}$	0.713	0.064	-0.0003	0.085	4.315	0.118
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.764	0.067	-0.012	0.342	2.893	0.731
Bordes symmetry Triangular Kernel	0.309	0.226	0.240	0.609	$\mu_2 = -0.220$	sd(μ_2) =0.398
Bordes symmetry Gaussian Kernel	0.211	0.133	0.106	0.533	$\mu_2 = -0.035$	sd(μ_2) =0.203
Robin et al.	0.488	0.137	-0.005	0.114	—	—
EM-type Song et al.	0.762	0.040	-0.005	0.092	—	—
$\pi-$maximizing Song et al.	0.717	0.156	-0.161	2.301	—	—
Stochastic EM	0.539	0.083	-0.005	0.112	—	—
Mixture 2 :$n = 100$ $\lambda^* = 0.3$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 3$, $\sigma_1^* = 1.5$(fixed)						
Pearson's χ^2 under $\mathcal{M}_{1:3}$	0.333	0.079	0.001	0.316	4.243	0.442
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.407	0.077	0.012	0.575	2.925	0.454
Bordes symmetry Triangular Kernel	0.272	0.119	0.773	0.947	$\mu_2 = -0.430$	sd(μ_2) =0.393
Bordes symmetry Gaussian Kernel	0.206	0.104	0.855	0.911	$\mu_2 = -0.308$	sd(μ_2) =0.350
Robin et al.	0.203	0.078	-0.109	0.947	—	—
EM-type Song et al.	0.494	0.035	-0.132	0.806	—	—
$\pi-$maximizing Song et al.	0.384	0.129	0.014	1.321	—	—
Stochastic EM	0.263	0.040	-0.062	0.646	—	—
Mixture 3 :$n = 300$ $\lambda^* = 0.2$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 3$, $\sigma_1^* = 1.5$(fixed)						
Pearson's χ^2 under $\mathcal{M}_{1:3}$	0.200	0.058	0.004	0.215	4.058	0.684
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.252	0.055	0.069	0.573	2.932	0.200
Bordes symmetry Triangular Kernel	0.439	0.108	-0.972	0.328	μ_2 =1.036	sd(μ_2) =0.496
Bordes symmetry Gaussian Kernel	0.414	0.096	-0.928	0.289	$\mu_2 = -1.125$	sd(μ_2) =0.470
Robin et al.	0.278	0.068	-0.062	1.253	—	—
EM-type Song et al.	0.461	0.023	0.162	1.128	—	—
$\pi-$maximizing Song et al.	0.362	0.020	0.025	1.224	—	—
Stochastic EM	0.292	0.057	0.118	1.027	—	—
Mixture 4 :$n = 10^5$ $\lambda^* = 0.2$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 2.5$, $\sigma_1^* = 2$(fixed)						
Pearson's χ^2 under $\mathcal{M}_{1:3}$	0.161	0.010	-0.002	0.019	3.874	0.661
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.203	0.004	-0.018	0.213	2.492	0.012
EM-type Song et al.	0.325	0.012	-0.061	1.469	—	—
$\pi-$maximizing Song et al.	0.251	0.002	-0.158	1.592	—	—
Mixture 5 :$n = 10^5$ $\lambda^* = 0.2$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 1.5$, $\sigma_1^* = 2$(fixed)						
Pearson's χ^2 under $\mathcal{M}_{1:3}$	0.015	0.030	0.203	2.381	2.150	0.138
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.213	0.013	-0.004	0.436	1.494	0.009
EM-type Song et al.	0.397	0.002	0.001	0.021	—	—
Mixture 6 :$n = 10^5$ $\lambda^* = 0.05$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 1.5$, $\sigma_1^* = 2$(fixed)						
Pearson's χ^2 under $\mathcal{M}_{1:3}$	0.005	0.033	-0.105	2.693	1.581	0.056
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.066	0.013	-0.036	0.857	1.493	0.008
EM-type Song et al.	0.304	0.014	-0.030	0.910	—	—
$\pi-$maximizing Song et al.	0.231	0.002	0.017	0.801	—	—
Mixture 7 :$n = 10^7$ $\lambda^* = 0.05$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 1.5$, $\sigma_1^* = 2$(fixed)						
Pearson's χ^2 under $\mathcal{M}_{1:3}$	0.006	0.010	0.024	0.197	1.500	0.019
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.051	0.001	0.002	0.259	1.500	0.001
Mixture 8 :$n = 10^7$ $\lambda^* = 0.01$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 1.5$, $\sigma_1^* = 2$(fixed)						
Pearson's χ^2 under $\mathcal{M}_{1:3}$	0.005	0.002	-0.011	0.162	1.509	0.004
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.011	0.001	-0.013	0.594	1.499	0.001

Table 3.4: The mean value with the standard deviation of estimates in a 100-run experiment on a two-component two-sided Weibull–Gaussian mixture.

3.5.4 Data generated from a bivariate Gaussian mixture and modeled by a semiparametric bivariate Gaussian mixture

We generate 1000 i.i.d. observations from a bivariate Gaussian mixture with proportion $\lambda = 0.7$ for the parametric component. The parametric component is a bivariate Gaussian with mean $(0, -1)$ and covariance matrix I_2. The unknown component is a bivariate Gaussian with mean $(3, 3)$ and covariance matrix:

$$\Sigma_2 = \begin{pmatrix} \sigma_2^{*2} & \rho^* \\ \rho^* & \sigma_2^{*2} \end{pmatrix}, \qquad \sigma_2^{*2} = 0.5, \quad \rho^* \in \{0, 0.25\}.$$

In a first experiment, we suppose that we know the whole parametric component, and that the unknown component belongs to the set $\mathcal{M}_1$

$$\mathcal{M}_1 = \left\{ \int_{\mathbb{R}^2} f_0(x,y)dxdy = 1, \quad \int_{\mathbb{R}^2} x f_0(x,y)dxdy = \int_{\mathbb{R}^2} y f_0(x,y)dydx = \theta, \quad \theta \in \mathbb{R} \right\}.$$

We suppose that the only unknown parameters are the center of the unknown cluster (θ, θ) and the proportion of the parametric component.
In a second experiment, we suppose that the center of the parametric component is unknown but given by $(\mu, \mu - 1)$ for some unknown $\mu \in \mathbb{R}$. The set of constraints is now replaced with $\mathcal{M}_2$ given by

$$\begin{aligned} \mathcal{M}_2 \quad = \quad \Bigg\{ & \int_{\mathbb{R}^2} f_0(x,y)dxdy = 1, \quad \int_{\mathbb{R}^2} x f_0(x,y)dxdy = \int_{\mathbb{R}^2} y f_0(x,y)dydx = \theta, \\ & \int_{\mathbb{R}^2} xy f_0(x,y)dxdy = \theta^2 + \rho^*, \theta \in \mathbb{R} \Bigg\}. \end{aligned}$$

The covariance between the two coordinates ρ^* in the unknown component is supposed to be known. We tested two values for $\rho^* = 0$ and $\rho^* = 0.25$, see figure (3.7).
Although existing methods were only proposed for univariate cases, we see no problem in using them in multivariate cases without any changes. The only method which cannot be used directly is the method of Bordes and Vandekerkhove [2010] because it is based on the symmetry of the density function, so it remained out of the competition.
For methods which use a kernel estimator, we used a kernel estimator for each coordinate of the random observations, i.e. $K_{w_x,w_y}(x,y) = K_{w_x}(x)K_{w_y}(y)$. The EM-type algorithm of Song et al. [2010] performs as good as our algorithm. The SEM algorithm of Bordes et al. [2007] gives also good results. The algorithm of Robin et al. [2007] and the $\pi-$maximizing algorithm of Song et al. [2010] failed to give satisfactory results.

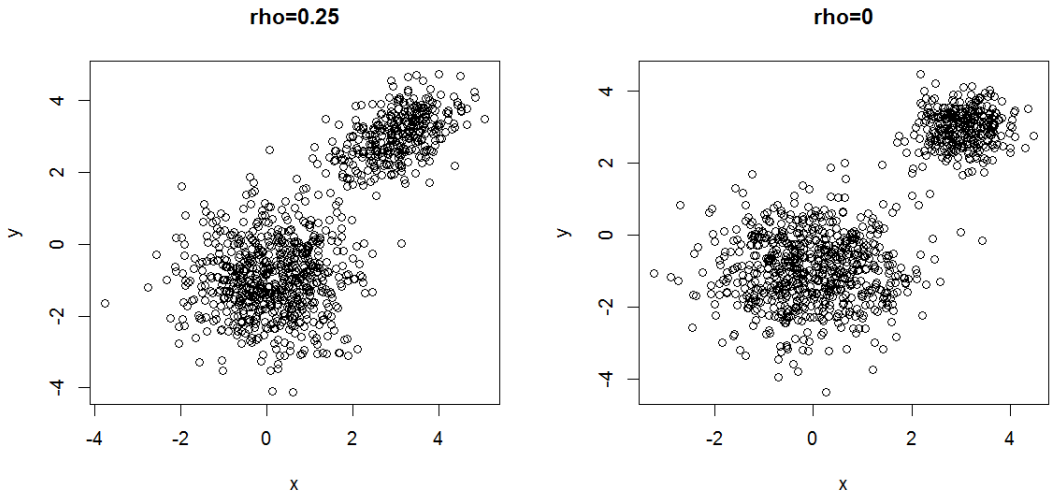

Figure 3.7: The two bivariate Gaussian mixtures.

Estimation method	λ	sd(λ)	μ	sd(μ)	θ	sd(θ)
Mixture 2 : $\rho^* = 0$ and $\mu_1 = (\mu, 1-\mu)$ is unknown						
Pearson's χ^2 under $\mathcal{M}_1$	0.680	0.027	—	—	2.854	0.233
Pearson's χ^2 under $\mathcal{M}_2$	0.694	0.019	0.016	0.035	3.034	0.045
SEM	0.724	0.015	0.090	0.043	$\mu_{1,2}$ = -0.880	sd($\mu_{1,2}$) =0.053
Robin	0.954	0.064	0.779	0.212	$\mu_{1,2}$ = -0.221	sd($\mu_{1,2}$) =0.218
Song EM	0.697	0.014	0.003	0.038	$\mu_{1,2}$ = -0.996	sd($\mu_{1,2}$) =0.039
Song $\pi-$maximizing	0.114	0.297	0.538	1.810	$\mu_{1,2}$ = -0.463	sd($\mu_{1,2}$) =1.810
Mixture 3 : $\rho^* = 0.25$ and $\mu_1 = (\mu, 1-\mu)$ is unknown						
Pearson's χ^2 under $\mathcal{M}_2$	0.704	0.026	0.033	0.060	3.071	0.101
SEM	0.730	0.016	0.083	0.052	$\mu_{1,2}$ =-0.878	sd($\mu_{1,2}$) =0.055
Robin	0.890	0.025	0.566	0.117	$\mu_{1,2}$ = -0.434	sd($\mu_{1,2}$) =0.117
Song EM	0.704	0.015	0.016	0.047	$\mu_{1,2}$ = -0.973	sd($\mu_{1,2}$) =0.040
Song $\pi-$maximizing	0.095	0.268	0.564	1.606	$\mu_{1,2}$ = -0.436	sd($\mu_{1,2}$) =1.606

Table 3.5: The mean value with the standard deviation of estimates in a 100-run experiment on a two-component bivariate normal mixture.

3.6 Conclusions

In this chapter, we proposed a structure for a two-component semiparametric mixture models where one component is parameteric with unknown parameter, and a component defined by linear constraints on its distribution function. These constraints may be moments constraints for example. We proposed also an algorithm which estimates the parameters of this model and showed how we can implement it efficiently even in multivariate contexts. The algorithm has a linear complexity when we use the Pearson's χ^2 divergence and the constraints are polynomials (thus moments constraints). We provided sufficient conditions in order to prove the consistency and the asymptotic normality of the resulting estimators.
Simulations show the gain we have by adding moments constraints in comparison to existing methods which do not consider any prior information. The method give clear good results even if the proportion of the parametric component is very low (equal to 0.01). In signal-noise applications, this can be interpreted otherwise. As long as we are able to estimate with relatively high precision the proportion of the signal (parametric component), we are proving the existence of the signal in a very heavy noise (99% of the data) even if the position of the signal is not accurately estimated. We showed in a simple example that our model can be applied in multivariate contexts. The new model shows encouraging properties and results, and should be tested further on real datasets.

3.7 Appendix: Proofs

3.7.1 Proof of Proposition 3.3.1

Proof. Based on equation (3.3.2), we may write the corresponding constraints equations, which are a fortiori equal:

$$\lambda \int g(x)dP_1(x|\theta) + (1-\lambda)m(\alpha) = \tilde{\lambda} \int g(x)dP_1(x|\tilde{\theta}) + (1-\tilde{\lambda})m(\tilde{\alpha}).$$

Define the following function:

$$G : \mathbb{R}^d \to \mathbb{R}^\ell : (\lambda, \theta, \alpha) \mapsto \lambda \int g(x)dP_1(x|\theta) + (1-\lambda)m(\alpha).$$

The solution to the previous system of equations is now equivalent to the fact that function G is one-to-one. This means that for a fixed m^*, we need that the nonlinear system of equations:

$$\frac{1}{1-\lambda}m^* - \frac{\lambda}{1-\lambda}m_1(\theta) = m_0(\alpha) \tag{3.7.1}$$

has a unique solution $(\lambda, \theta, \alpha)$. The value of m^* is given by $\int g(x)dP_T$ where P_T is the mixture we are considering. To conclude, suppose that the system (3.7.1) has a unique solution $(\lambda^*, \theta^*, \alpha^*)$ for each given m^*, then function G is one-to-one and the constraints equations imply that $\lambda = \tilde{\lambda}, \theta = \tilde{\theta}$ and $\alpha = \tilde{\alpha}$. Finally, using (3.3.2), we may deduce that $P_0 = \tilde{P}_0$. Thus, the semiparametric mixture model is identifiable as soon as the nonlinear system of equations (3.7.1) has a unique solution $(\lambda^*, \theta^*, \alpha^*)$. □

3.7.2 Proof of Proposition 3.3.2

Proof. Let P_0 be some signed measure which belongs to the intersection $\mathcal{N} \cap \mathcal{M}$. Since P_0 belongs to $\mathcal{N}$, there exists a couple (λ, θ) such that:

$$P_0 = \frac{1}{1-\lambda} P_T - \frac{\lambda}{1-\lambda} P_1(.|\theta). \tag{3.7.2}$$

This couple is unique by virtue of assumptions 3 and 4. Indeed, let (λ, θ) and $(\tilde{\lambda}, \tilde{\theta})$ be two couples such that:

$$\frac{1}{1-\lambda} P_T - \frac{\lambda}{1-\lambda} P_1(.|\theta) = \frac{1}{1-\tilde{\lambda}} P_T - \frac{\tilde{\lambda}}{1-\tilde{\lambda}} P_1(.|\tilde{\theta}) \quad dP_T - a.e. \tag{3.7.3}$$

This entails that:

$$\frac{1}{1-\lambda} - \frac{\lambda}{1-\lambda} \frac{dP_1(x|\theta)}{dP_T(x)} = \frac{1}{1-\tilde{\lambda}} - \frac{\tilde{\lambda}}{1-\tilde{\lambda}} \frac{dP_1(x|\tilde{\theta})}{dP_T(x)}.$$

Taking the limit as $\|x\|$ tends to ∞ results in:

$$\frac{1-c\lambda}{1-\lambda} = \frac{1-\tilde{c}\lambda}{1-\tilde{\lambda}}.$$

Note that function $z \mapsto (1-cz)/(1-z)$ is strictly monotone as long as $c \neq 1$. Hence, it is a one-to-one map. Thus $\lambda = \tilde{\lambda}$. Inserting this result in equation (3.7.3) entails that:

$$P_1(.|\theta) = P_1(.|\tilde{\theta}) \qquad dP_T - a.e.$$

Using the identifiability of P_1 (assumption 4), we get $\theta = \tilde{\theta}$ which proves the existence of a unique couple (λ, θ) in (3.7.2).
On the other hand, since P_0 belongs to $\mathcal{M}$, there exists a unique α such that $P_0 \in \mathcal{M}_\alpha$. Uniqueness comes from the fact that function $\alpha \mapsto m(\alpha)$ is one-to-one (assumption 2). Thus, P_0 verifies the constraints

$$\int dP_0(x) = 1, \qquad \int g_i(x) dP_0(x) = m_i(\alpha), \quad \forall i = 1, \cdots, \ell.$$

Combining this with (3.7.2), we get:

$$\int \left(\frac{1}{1-\lambda} dP_T - \frac{\lambda}{1-\lambda} dP_1(x|\theta) \right) = 1, \int g_i(x) \left(\frac{1}{1-\lambda} dP_T - \frac{\lambda}{1-\lambda} dP_1(x|\theta) \right) = m_i(\alpha), \tag{3.7.4}$$

for all $i = 1, \cdots, \ell$. This is a non linear system of equations with $\ell+1$ equations. The first one is verified for any couple (λ, θ) since both $P(.|\phi^*)$ and P_1 are probability measures. This reduces the system to ℓ nonlinear equations.
Now, let P_0 and $\tilde{P}_0$ be two elements in $\mathcal{N} \cap \mathcal{M}$, then there exist two couples (λ, θ) and $(\tilde{\lambda}, \tilde{\theta})$ with $\lambda \neq \tilde{\lambda}$ or $\theta \neq \tilde{\theta}$. Since $P_0 \in \mathcal{M}$, there exists α such that $P_0 \in \mathcal{M}_\alpha$. Similarly, there exists $\tilde{\alpha}$ possibly different from α. Now, $(\lambda, \theta, \alpha)$ and $(\tilde{\lambda}, \tilde{\theta}, \tilde{\alpha})$ are two solutions to the system of equations (3.7.4) which contradicts with assumption 1 of the present proposition.
We may now conclude that, if a signed measure P_0 belongs to the intersection $\mathcal{N} \cap \mathcal{M}$, then it has the representation (3.7.2) for a unique couple (λ, θ) and there exists a unique α such that the triplet $(\lambda, \theta, \alpha)$ is a solution to the non linear system (3.7.4). Conversely,

if there exists a triplet $(\lambda, \theta, \alpha)$ which solves the non linear system (3.7.4), then the signed measure P_0 defined by $P_0 = \frac{1}{1-\lambda}P(.|\phi^*) - \frac{\lambda}{1-\lambda}P_1(.|\theta)$ belongs to the intersection $\mathcal{N} \cap \mathcal{M}$. This is because on the one hand, it clearly belongs to $\mathcal{N}$ by its definition and on the other hand, it belongs to $\mathcal{M}_\alpha$ since it verifies the constraints and thus belongs to $\mathcal{M}$.
It is now reasonable to conclude that under assumptions 2-4, the intersection $\mathcal{N} \cap \mathcal{M}$ includes a *unique* signed measure P_0 if and only if the set of ℓ non linear equations (3.7.4) has a unique solution $(\lambda, \theta, \alpha)$. □

3.7.3 Proof of Lemma 3.4.1

Proof. The proof is based partially on the proof of Proposition 3.7 part (ii) in Keziou [2003].
We proceed by contradiction. Let $\varepsilon > 0$ be such that $\sup_\phi \|\xi_n(\phi) - \xi(\phi)\| > \varepsilon$. Then, there exists a sequence $a_k \in \Phi$ such that $\|\xi_n(a_k) - \xi(a_k)\| > \varepsilon$. By assumption A3, there exists $\eta > 0$ such that:

$$H(a_k, \xi(a_k)) - H(a_k, \xi_n(a_k)) > \eta.$$

Thus,

$$\mathbb{P}\left(\sup_\phi \|\xi_n(\phi) - \xi(\phi)\| > \varepsilon\right) \leq \mathbb{P}\left(H(a_k, \xi(a_k)) - H(a_k, \xi_n(a_k)) > \eta\right). \tag{3.7.5}$$

Let's prove that the right hand side tends to zero as n goes to infinity which is sufficient to accomplish our claim.
By definition of $\xi_n(a_k)$ and assumption A2, we can write:

$$\begin{aligned} H_n(a_k, \xi_n(a_k)) &\geq H_n(a_k, \xi(a_k)) \\ &\geq H(a_k, \xi(a_k)) - o_P(1) \end{aligned}$$

where $o_P(1)$ does not depend upon a_k by virtue of A2. Now we have:

$$\begin{aligned} H(a_k, , \xi(a_k)) - H(a_k, \xi_n(a_k)) &\leq H_n(a_k, \xi_n(a_k)) - H(a_k, , \xi_n(a_k)) + o_P(1) \\ &\leq \sup_{\xi,\phi} |H_n(\phi, \xi) - H(\phi, \xi)| + o_P(1). \end{aligned}$$

Last but not least, assumption A2 permits to conclude that the right hand side tends to zero in probability. Since the left hand side is already nonnegative by definition of $\xi(a_k)$, then by the previous result we conclude that $H(a_k, , \xi(a_k)) - H(a_k, \xi_n(a_k))$ tends to zero in probability. Employing this final result in inequality (3.7.5), we get that $\sup_\phi \|\xi_n(\phi) - \xi(\phi)\|$ tends to zero in probability. □

3.7.4 Proof of Theorem 3.4.1

Proof. We proceed by contradiction in a similar way to the proof of Lemma 3.4.1. Let $\kappa > 0$ be such that $\|\phi^* - \hat{\phi}\| > \kappa$, then by assumption A4, there exists $\eta > 0$ such that :

$$H(\hat{\phi}, \xi(\hat{\phi})) - H(\phi^*, \xi(\phi^*)) > \eta.$$

This can be rewritten as:

$$\mathbb{P}\left(\|\phi^* - \hat{\phi}\| > \kappa\right) \leq \mathbb{P}\left(H(\hat{\phi}, \xi(\hat{\phi})) - H(\phi^*, \xi(\phi^*)) > \eta\right). \tag{3.7.6}$$

We now demonstrate that the right hand side tends to zero as n goes to infinity. Let $\varepsilon > 0$ be such that for n sufficiently large, we have $\sup_{\xi,\phi} |H(\phi, \xi) - H_n(\phi, \xi)| < \varepsilon$. This

is possible by virtue of assumption A2. The definition of $\hat{\phi}$ together with assumption A2 will now imply:

$$\begin{aligned} H_n(\hat{\phi}, \xi_n(\hat{\phi})) &\leq H_n(\phi^*, \xi_n(\phi^*)) \\ &\leq H(\phi^*, \xi_n(\phi^*)) + \sup_{\xi,\phi} |H(\phi,\xi) - H_n(\phi,\xi)| \\ &\leq H(\phi^*, \xi_n(\phi^*)) + \varepsilon. \end{aligned} \tag{3.7.7}$$

We use now the continuity assumption A5 of function $\xi \mapsto H(\phi^*, \xi)$ at $\xi(\phi^*)$. For the ε chosen earlier, there exists $\delta(\phi^*, \varepsilon)$ such that if $\|\xi(\phi^*) - \xi_n(\phi^*)\| < \delta(\phi^*, \varepsilon)$, then:

$$|H(\phi^*, \xi_n(\phi^*)) - H(\phi^*, \xi(\phi^*))| < \varepsilon.$$

This is possible for sufficiently large n since $\sup_\phi \|\xi(\phi^*) - \xi_n(\phi^*)\|$ tends to zero in probability by Lemma 3.4.1. Inserting this result in (3.7.7) gives:

$$H_n(\hat{\phi}, \xi_n(\hat{\phi})) \leq H(\phi^*, \xi(\phi^*)) + 2\varepsilon.$$

We now have:

$$\begin{aligned} H(\hat{\phi}, \xi(\hat{\phi})) - H(\phi^*, \xi(\phi^*)) &\leq H(\hat{\phi}, \xi(\hat{\phi})) - H_n(\hat{\phi}, \xi_n(\hat{\phi})) + 2\varepsilon \\ &\leq H(\hat{\phi}, \xi(\hat{\phi})) - H(\hat{\phi}, \xi_n(\hat{\phi})) + H(\hat{\phi}, \xi_n(\hat{\phi})) - H_n(\hat{\phi}, \xi_n(\hat{\phi})) + 2\varepsilon. \end{aligned}$$

Continuity assumption of H implies that for $\varepsilon > 0$, there exists $\delta(\hat{\phi}, \varepsilon) > 0$ such that if $\|\xi(\hat{\phi}) - \xi_n(\hat{\phi})\| < \delta(\hat{\phi}, \varepsilon)$, then:

$$\left| H(\hat{\phi}, \xi(\hat{\phi})) - H(\hat{\phi}, \xi_n(\hat{\phi})) \right| \leq \varepsilon.$$

This is again possible for sufficiently large n since $\sup_\phi \|\xi(\phi^*) - \xi_n(\phi^*)\|$ tends to zero in probability by Lemma 3.4.1. This entails that:

$$\begin{aligned} H(\hat{\phi}, \xi(\hat{\phi})) - H(\phi^*, \xi(\phi^*)) &\leq H(\hat{\phi}, \xi_n(\hat{\phi})) - H_n(\hat{\phi}, \xi_n(\hat{\phi})) + 3\varepsilon \\ &\leq \sup_{\xi,\phi} |H(\phi,\xi) - H_n(\phi,\xi)| + 3\varepsilon \\ &\leq 4\varepsilon \end{aligned}$$

We conclude that the right hand side in (3.7.6) goes to zero and the proof is completed. □

3.7.5 Proof of Theorem 3.4.2

Proof. We will use Theorem 3.4.1. We need to verify assumptions A2 and A3. Since the class of functions $\{(\phi,\xi) \mapsto h(\phi,\xi,.)\}$ is a Glivenko-Cantelli class of functions, then assumption A2 is fulfilled by the Glivenko-Cantelli theorem. Finally, assumption A3 can be checked by strict concavity of function $\xi \mapsto H(\phi,\xi)$. Indeed, for any $\eta \in (0,1)$ and any ξ_1, ξ_2, we have by strict convexity of ψ :

$$\psi\left(\eta\xi_1^t g(x) + (1-\eta)\xi_2^t g(x)\right) < \eta\psi\left(\xi_1^t g(x)\right) + (1-\eta)\psi\left(\xi_2^t g(x)\right).$$

If the measure $dP/(1-\lambda) - \lambda dP_1(.|\theta)/(1-\lambda)$ is positive[10], we may write:

$$\int \psi\left(\eta\xi_1^t g(x) + (1-\eta)\xi_2^t g(x)\right)\left(\frac{1}{1-\lambda}dP(x) - \frac{\lambda}{1-\lambda}dP_1(x|\theta)\right) <$$
$$\eta\int \psi\left(\xi_1^t g(x)\right)\left(\frac{1}{1-\lambda}dP(x) - \frac{\lambda}{1-\lambda}dP_1(x|\theta)\right) + (1-\eta)\int \psi\left(\xi_2^t g(x)\right)\left(\frac{1}{1-\lambda}dP(x) - \frac{\lambda}{1-\lambda}dP_1(x|\theta)\right),$$

[10] This measure can never be zero since it integrates to one, thus we do not need to suppose that it is nonnegative.

which entails that

$$H(\phi, \eta\xi_1 + (1-\eta)\xi_2) > \eta H(\phi, \xi_1) + (1-\eta)H(\phi, \xi_2),$$

and function $\xi \mapsto H(\phi, \xi)$ becomes strictly concave. However, the measure $dP/(1-\lambda) - \lambda dP_1(.|\theta)/(1-\lambda)$ is in general a signed measure and the previous implication does not hold. This is not dramatic because function $\xi \mapsto H(\phi, \xi)$ has only two choices; it is either strictly convex or strictly concave. In case function $\xi \mapsto H(\phi, \xi)$ is strictly convex, then its supremum is infinity and the corresponding vector ϕ does not count in the calculus of the infimum after all. This means that the only vectors $\phi \in \Phi$ which interest us are those for which function $\xi \mapsto H(\phi, \xi)$ is strictly concave. In other words, the infimum in (3.3.10) can be calculated over the set:

$$\Phi^+ = \Phi \cap \{\phi : \quad \xi \mapsto H(\phi, \xi) \text{ is strictly concave}\}$$

instead of over Φ. All assumptions of Theorem 3.4.1 are now fulfilled and $\hat{\phi}$ converges in probability to ϕ^*. □

3.7.6 Proof of Proposition 3.4.1

Proof. We already have:

$$\frac{1}{1-\lambda^*} P_T - \frac{\lambda^*}{1-\lambda^*} P_1(.|\theta^*) = P_0^*,$$

and since P_0^* is supposed to be a probability measure, the matrix $J_{H(\phi^*,.)}$ is definite negative. Thus $\phi^* \in \Phi^+$. Since the set of negative definite matrices is an open set (see for example page 36 in Lange [2013]), there exists a ball $\mathcal{U}$ of negative definite matrices centered at $J_{H(\phi^*,.)}$. Continuity of $\phi \mapsto J_{H(\phi,.)}$ permits[11] to find a ball $B(\phi^*, \tilde{r})$ such that the subset $\{J_{H(\phi,.)} : \phi \in B(\phi^*, \tilde{r})\}$ is inside $\mathcal{U}$. Now the neighborhood we are looking at is the ball $B(\phi^*, \tilde{r})$.
For the second part of the proposition, the existence and finiteness of $\xi(\phi)$ for $\phi \in \mathcal{V} = B(\phi^*, \tilde{r})$ is immediate since function $\xi \mapsto H(\phi, \xi)$ is strictly concave. Besides the the differentiability of the function $\phi \mapsto \xi(\phi)$ is a direct result of the implicit function theorem applied on the equation $\xi \mapsto \nabla H(\phi, .)$. Notice that the Hessian matrix of $H(\phi, .)$ is invertible since it is symmetric definite negative. □

3.7.7 Proof of Theorem 3.4.3

Proof. We follow the steps of Theorem 3.2 in Newey and Smith [2004]. The idea behind the proof is a mean value expansion with Lagrange remainder of the estimating equations. We need at first to verify if $\hat{\phi}$ belongs to the interior of Φ^+ in order to be able to differentiate $\phi \mapsto H_n(\phi, \xi)$. This can be done similarly to Proposition 3.4.1. We also can prove (by replacing H by H_n and $\xi(\phi)$ by $\xi_n(\phi)$) that $\phi \mapsto \xi_n(\phi)$ is continuously differentiable in a neighborhood of ϕ^*.
We may now proceed to the mean value expansion. By the very definition of $\xi_n(\phi)$, we have:

$$\frac{\partial H_n}{\partial \xi}(\phi, \xi_n(\phi)) = 0 \qquad \forall \phi \in \text{int}(\Phi^+),$$

[11] To see this, consider Sylvester's rule which is based on a test using the determinant of the sub-matrices of J_H. Each determinant needs to be negative. The continuity of the determinant function together with the continuity of $\phi \mapsto J_{H(\phi,.)}$ will imply that we may move around $J_(H(\phi^*, .))$ in a small neighborhood in a way that the determinants of the sub-matrices stay negative.

which also holds for $\phi = \hat{\phi}$, i.e.

$$\frac{\partial H_n}{\partial \xi}(\hat{\phi}, \xi_n(\hat{\phi})) = 0.$$

On the other hand, the definition of $\hat{\phi}$ implies that:

$$\frac{\partial}{\partial \phi} H_n(\phi, \xi_n(\phi))\Big|_{\phi=\hat{\phi}} = 0.$$

Since function $\phi \mapsto \xi_n(\phi)$ is continuously differentiable. A simple chain rule implies

$$\begin{aligned}\frac{\partial}{\partial \phi}\left(H_n(\phi, \xi_n(\phi))\right)\Big|_{\phi=\hat{\phi}} &= \frac{\partial}{\partial \phi} H_n(\hat{\phi}, \xi_n(\hat{\phi})) + \frac{\partial}{\partial \xi} H_n(\hat{\phi}, \xi_n(\hat{\phi}))\frac{\partial \xi_n}{\partial \phi}(\hat{\phi}) \\ &= \frac{\partial}{\partial \phi} H_n(\hat{\phi}, \xi_n(\hat{\phi})).\end{aligned}$$

The second line comes from the definition of $\xi_n(\phi)$ as the argument of the supremum of function $\xi \mapsto H_n(\phi, \xi)$. Now, the estimating equations are given simply by:

$$\begin{aligned}\frac{\partial H_n}{\partial \xi}(\hat{\phi}, \xi_n(\hat{\phi})) &= 0; \\ \frac{\partial H_n}{\partial \phi}(\hat{\phi}, \xi_n(\hat{\phi})) &= 0.\end{aligned}$$

We need to calculate these partial derivatives. We start by the derivative with respect to ξ:

$$\frac{\partial H_n}{\partial \xi}(\phi, \xi) = m(\alpha) - \frac{1}{1-\lambda}\frac{1}{n}\sum_{i=1}^{n}\psi'\left(\xi^t g(x_i)\right) g(x_i) + \frac{\lambda}{1-\lambda}\int \psi'\left(\xi^t g(x)\right) g(x) p_1(x|\theta) dx \tag{3.7.8}$$

We calculate the partial derivatives with respect to α, λ and θ:

$$\frac{\partial H_n}{\partial \alpha} = \xi^t \nabla m(\alpha) \tag{3.7.9}$$

$$\frac{\partial H_n}{\partial \lambda} = -\frac{1}{(1-\lambda)^2}\frac{1}{n}\sum_{i=1}^{n}\psi\left(\xi^t g(x_i)\right) + \frac{1}{(1-\lambda)^2}\int \psi\left(\xi^t g(x)\right) p_1(x|\theta) dx \tag{3.7.10}$$

$$\frac{\partial H_n}{\partial \theta} = \frac{\lambda}{1-\lambda}\int \psi\left(\xi^t g(x)\right) \nabla_\theta p_1(x|\theta) dx \tag{3.7.11}$$

Notice that by Lemma 3.4.1, the continuity of $\phi \mapsto \xi(\phi)$ and the consistency of $\hat{\phi}$ towards ϕ^*, we have $\xi_n(\hat{\phi}) \to \xi(\phi^*) = 0$ in probability. A mean value expansion of the estimating equation between $(\hat{\phi}, \xi_n(\hat{\phi}))$ and $(\phi^*, 0)$ implies that there exists $(\bar{\phi}, \bar{\xi})$ on the line between these two points such that:

$$\begin{pmatrix} \frac{\partial H_n}{\partial \phi}(\hat{\phi}, \xi_n(\hat{\phi})) \\ \frac{\partial H_n}{\partial \xi}(\hat{\phi}, \xi_n(\hat{\phi})) \end{pmatrix} = \begin{pmatrix} \frac{\partial H_n}{\partial \phi}(\phi^*, 0) \\ \frac{\partial H_n}{\partial \xi}(\phi^*, 0) \end{pmatrix} + J_{H_n}(\bar{\phi}, \bar{\xi}) \begin{pmatrix} \hat{\phi} - \phi^* \\ \xi_n(\hat{\phi}) \end{pmatrix}, \tag{3.7.12}$$

where $J_{H_n}(\bar{\phi}, \bar{\xi})$ is the matrix of second derivatives of H_n calculated at the mid point $(\bar{\phi}, \bar{\xi})$. The left hand side is zero, so we need to calculate the first vector in the right hand side. We have by simple substitution in formula (3.7.8):

$$\frac{\partial H_n}{\partial \xi}(\phi^*, 0) = m(\alpha^*) - \frac{1}{1-\lambda^*}\frac{1}{n}\sum_{i=1}^{n} g(x_i) + \frac{\lambda^*}{1-\lambda^*}\int g(x) p_1(x|\theta^*) dx.$$

Using the assumption that the model (3.3.5) verify the set of constraints defining $\mathcal{M}_\alpha$ together with the CLT, we write:

$$\sqrt{n}\frac{\partial H_n}{\partial \xi}(\phi^*, 0) \underset{\mathcal{L}}{\to} \mathcal{N}\left(0, \frac{1}{(1-\lambda^*)^2}\mathrm{Var}_{P_T}(g(X))\right). \tag{3.7.13}$$

Using formulas (3.7.9), (3.7.10) and (3.7.11), we may write:

$$\begin{aligned}
\frac{\partial H_n}{\partial \alpha}(\phi^*, 0) &= 0;\\
\frac{\partial H_n}{\partial \lambda}(\phi^*, 0) &= -\frac{1}{(1-\lambda^*)^2} + \frac{1}{(1-\lambda^*)^2} = 0;\\
\frac{\partial H_n}{\partial \theta}(\phi^*, 0) &= \frac{\lambda^*}{1-\lambda^*}\int \nabla_\theta p_1(x|\theta^*)dx = \frac{\lambda^*}{1-\lambda^*}\nabla_\theta \int p_1(x|\theta^*)dx = 0.
\end{aligned}$$

The final line holds since by Lebesgue's differentiability theorem using assumption 5 for $\xi = 0$, we can change between the sign of integration and derivation. Combine this with the fact that $p_1(x|\theta^*)$ is a probability density function which integrates to 1, gives the result in the last line.
We need now to write an explicit form for the matrix $J_{H_n}(\bar{\phi}, \bar{\xi})$ and study its limit in probability. It contains the second order partial derivatives of function H_n with respect to its parameters. We start by the double derivatives. Using formulas (3.7.8), (3.7.9), (3.7.10) and (3.7.11), we write:

$$\begin{aligned}
\frac{\partial^2 H_n}{\partial \xi^2} &= -\frac{1}{1-\lambda}\frac{1}{n}\sum_{i=1}^n \psi''\left(\xi^t g(x_i)\right) g(x_i)g(x_i)^t + \frac{\lambda}{1-\lambda}\int \psi''\left(\xi^t g(x)\right) g(x)g(x)^t p_1(x|\theta)dx;\\
\frac{\partial^2 H_n}{\partial \alpha^2} &= \xi^t J_m(\alpha);\\
\frac{\partial^2 H_n}{\partial \lambda^2} &= -\frac{2}{(1-\lambda)^3}\frac{1}{n}\sum_{i=1}^n \psi\left(\xi^t g(x_i)\right) + \frac{2}{(1-\lambda)^3}\int \psi\left(\xi^t g(x)\right) p_1(x|\theta)dx;\\
\frac{\partial^2 H_n}{\partial \theta^2} &= \frac{\lambda}{1-\lambda}\int \psi\left(\xi^t g(x)\right) J_{p_1(x|\theta)}dx;\\
\frac{\partial^2 H_n}{\partial \xi \partial \alpha} &= \nabla m(\alpha);\\
\frac{\partial^2 H_n}{\partial \xi \partial \lambda} &= -\frac{1}{(1-\lambda)^2}\frac{1}{n}\sum_{i=1}^n \psi'\left(\xi^t g(x_i)\right) g(x_i) + \frac{1}{(1-\lambda)^2}\int \psi'\left(\xi^t g(x)\right) g(x)p_1(x|\theta)dx;\\
\frac{\partial^2 H_n}{\partial \xi \partial \theta} &= \frac{\lambda}{1-\lambda}\int \psi'\left(\xi^t g(x)\right) g(x)\nabla_\theta p_1(x|\theta)^t dx;\\
\frac{\partial^2 H_n}{\partial \alpha \partial \lambda} &= 0;\\
\frac{\partial^2 H_n}{\partial \alpha \partial \theta} &= 0;\\
\frac{\partial^2 H_n}{\partial \lambda \partial \theta} &= \frac{1}{(1-\lambda)^2}\int \psi\left(\xi^t g(x)\right) \nabla_\theta p_1(x|\theta)dx.
\end{aligned}$$

As n goes to infinity, we have $\bar{\xi} \to 0$ and $\bar{\phi} \to \phi^*$. Then, under regularity assumptions of the present theorem, we can calculate the limit in probability of the matrix $J_{H_n}(\bar{\phi}, \bar{\xi})$. The blocks limits are given by:

$$\frac{\partial^2 H_n}{\partial \xi^2} \overset{\mathbb{P}}{\to} -\mathbb{E}_{P_0^*}\left[g(X)g(X)^t\right], \quad \frac{\partial^2 H_n}{\partial \alpha^2} \overset{\mathbb{P}}{\to} 0, \quad \frac{\partial^2 H_n}{\partial \lambda^2} \overset{\mathbb{P}}{\to} 0, \quad \frac{\partial^2 H_n}{\partial \theta^2} \overset{\mathbb{P}}{\to} 0, \quad \frac{\partial^2 H_n}{\partial \xi \partial \alpha} \overset{\mathbb{P}}{\to} \nabla m(\alpha^*)$$

$$\frac{\partial^2 H_n}{\partial\xi\partial\lambda} \xrightarrow{\mathbb{P}} -\frac{1}{(1-\lambda^*)^2}\mathbb{E}_{P_T}\left[g(X)\right] + \frac{1}{(1-\lambda^*)^2}\int g(x)p_1(x|\theta^*)dx$$

$$\frac{\partial^2 H_n}{\partial\xi\partial\theta} \xrightarrow{\mathbb{P}} \frac{\lambda^*}{1-\lambda^*}\int g(x)\nabla_\theta p_1(x|\theta^*)dx, \qquad \frac{\partial^2 H_n}{\partial\alpha\partial\lambda} \xrightarrow{\mathbb{P}} 0, \qquad \frac{\partial^2 H_n}{\partial\alpha\partial\theta} \xrightarrow{\mathbb{P}} 0, \qquad \frac{\partial^2 H_n}{\partial\lambda\partial\theta} \xrightarrow{\mathbb{P}} 0,$$

taking into account that $\psi(0) = 0, \psi'(0) = 1$ and $\psi''(0) = 1$. The limit in probability of the matrix $J_{H_n}(\bar{\phi}, \bar{\xi})$ can be written in the form:

$$J_H = \begin{bmatrix} 0 & J^t_{\phi^*,\xi^*} \\ J_{\phi^*,\xi^*} & J_{\xi^*,\xi^*} \end{bmatrix},$$

where J_{ϕ^*,ξ^*} and J_{ξ^*,ξ^*} are given by (3.4.3) and (3.4.4). The inverse of matrix J_H has the form:

$$J_H^{-1} = \begin{pmatrix} -\Sigma & H \\ H^t & W \end{pmatrix},$$

where

$$\Sigma = \left(J^t_{\phi^*,\xi^*}J_{\xi^*,\xi^*}J_{\phi^*,\xi^*}\right)^{-1}, \quad H = \Sigma J^t_{\phi^*,\xi^*}J^{-1}_{\xi^*,\xi^*}, \quad W = J^{-1}_{\xi^*,\xi^*} - J^{-1}_{\xi^*,\xi^*}J_{\phi^*,\xi^*}\Sigma J^t_{\phi^*,\xi^*}J^{-1}_{\xi^*,\xi^*}.$$

Going back to (3.7.12), we have:

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{\partial H_n}{\partial\xi}(\phi^*, 0) \end{pmatrix} + J_{H_n}(\bar{\phi}, \bar{\xi}) \begin{pmatrix} \hat{\phi} - \phi^* \\ \xi_n(\hat{\phi}) \end{pmatrix}.$$

Solving this equation in ϕ and ξ gives:

$$\begin{pmatrix} \sqrt{n}\left(\hat{\phi} - \phi^*\right) \\ \sqrt{n}\xi_n(\hat{\phi}) \end{pmatrix} = J_H^{-1} \begin{pmatrix} 0 \\ \sqrt{n}\frac{\partial H_n}{\partial\xi}(\phi^*, 0) \end{pmatrix} + o_P(1).$$

Finally, using (3.7.13), we get that:

$$\begin{pmatrix} \sqrt{n}\left(\hat{\phi} - \phi^*\right) \\ \sqrt{n}\xi_n(\hat{\phi}) \end{pmatrix} \xrightarrow[\mathcal{L}]{} \mathcal{N}\left(0, S\right)$$

where

$$S = \frac{1}{(1-\lambda^*)^2}\begin{pmatrix} H \\ W \end{pmatrix} \mathrm{Var}_{P_T}(g(X)) \begin{pmatrix} H^t & W^t \end{pmatrix}.$$

This ends the proof. □

Chapter 4

Semiparametric two-component mixture models where one component is defined through L-moments constraints

Recall that a semiparametric two-component mixutre model is defined by:

$$f(x) = \lambda f_1(x|\theta) + (1-\lambda) f_0(x), \qquad \text{for } x \in \mathbb{R} \tag{4.0.1}$$

for $\lambda \in (0,1)$ and $\theta \in \mathbb{R}^d$ to be estimated and the density f_0 is considered to be unknown. We have proposed in Chapter 3 a method which incorporates moment-type constraints on the unknown component. The method outperforms other semiparametric methods which do not use prior information encouraging the use of a suitable prior information. Moment-type constraints are not suitable for positive-support mixtures especially when the density does not decrease fast enough. The method needs more observations to be able to estimate such mixtures.
We thus propose here to use L-moments constraints. L-moments have become classical tools alternative to central moments for the description of dispersion, skewness and kurtosis of a univariate heavy-tailed distribution. Distributions such as the Lognormal, the Pareto and the Weibull distributions are standard examples of such distributions. The use of L-moments is evolving since their introduction by Hosking [1990]. One of the main interests of L-moments is that they can be defined as soon as the expectation of the random variable exists. Broniatowski and Decurninge [2016] has proposed a structure and an estimation procedure for semiparametric models defined through L-moments conditions. The resulting estimators performe well under the model, and they outperforme existing methods in misspecification contexts.
Similarly to the case of moment constraints seen in the previous chapter, the incorporation of L-moments constraints cannot be done directly in existing methods for semiparametric mixtures (see paragraph 3.1) because the optimization will be carried over a (possibly) infinite dimensional space on the one hand, and on the other hand, existing methods use either the distribution function or the probability density function and cannot adapt an approach based on the quantile function. Our approach introduced in the previous chapter cannot be used either because L-moments are not linear functions of the distribution function as we will see in paragraph 4.1.1. We thus need a new tool. Convex analysis offer away using Fenchel-Legendre duality to transform an optimization problem over an

infinite dimensional space to the space of Lagrangian parameters (finite dimensional one). $\varphi-$divergences, by their convexity properties, are suitable tools in order to use the duality result. Chap 1 in the PhD thesis of Decurninge [2015] introduced a method based on $\varphi-$divergences to estimate a semiparametric model defined subject to L-moments constraints. We will exploit his methodology to build a new estimation procedure which takes into account L-moments constraints over the unknown component's distribution.

4.1 Semiparametric models defined through L-moments constraints

In this section, we present a definition of a semiparametric model subject to L-moments constraints in a similar way to semiparametric models defined through moments constraints. An essential part to begin with is the definition of L-moments. We will keep this part brief and one can consult Decurninge [2015] Chap. 1 or Hosking [1990] for more details.
We recall two important notions; the quantile function and the quantile measure. Let $X_1, \dots X_n$ be n i.i.d. copies of a random variable X taking values in $\mathbb{R}$ with unknown cumulative distribution function (cdf) $\mathbb{F}$. Denote by $\mathbb{F}^{-1}(u)$ for $u \in (0,1)$ the associated quantile function of the cdf $\mathbb{F}$ defined by

$$\mathbb{F}^{-1}(u) = \inf \{x \in \mathbb{R}, \quad s.t. \ \mathbb{F}(x) \geq u\}, \quad u \in (0,1).$$

We can associate to $\mathbb{F}^{-1}$ a measure $\mathbf{F}^{-1}$ on $\mathcal{B}([0,1])$ given by

$$\mathbf{F}^{-1}(B) = \int_0^1 \mathbb{1}_{x \in B} d\mathbb{F}^{-1}(x) \in \mathbb{R} \cup \{-\infty, +\infty\}.$$

The integral here is a Riemann-Stieltjes one. $\mathbf{F}^{-1}$ is a $\sigma-$finite measure since $\mathbb{F}^{-1}$ has bounded variations on every subinterval $[a,b]$ from $(0,1)$.
In this section, we suppose that $\mathbb{E}|X| < \infty$ and $\int |x| dF(x) < \infty$. We adapt the standard notation for the cumulative distribution function (cdf) and measures, i.e. a measure P has a cdf $\mathbb{F}$, a density p with respect to the Lebesgue measure and a quantile measure $\mathbf{F}^{-1}$, and a measure Q has a cdf $\mathbb{Q}$, a density q with respect to the Lebesgue measure and a quantile measure $\mathbf{Q}$.

4.1.1 L-moments: Definition and first properties

Let $X_{1:n} < \ldots < X_{n:n}$ be the order statistics associated to the sample $X_1, \cdots, X_n$.

Definition 4.1.1. *The L-moment of order r, denoted λ_r, $r = 1, 2, \dots$ is defined as a linear combination of the expectation of order statistics:*

$$\lambda_r = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \mathbb{E}\left(X_{r-k:r}\right).$$

If $\mathbb{F}$ is continuous, then the expectation of the j-th order statistic is given by

$$\mathbb{E}\left[X_{j:r}\right] = \frac{r!}{(j-1)!(r-j)!} \int_{\mathbb{R}} x \mathbb{F}(x)^{j-1} \left[1 - \mathbb{F}(x)\right]^{r-j} d\mathbb{F}(x). \tag{4.1.1}$$

In particular, the first three L-moments are

$$\begin{aligned}\lambda_1 &= \mathbb{E}[X];\\ \lambda_2 &= \left(\mathbb{E}\left[X_{2:2}\right] - \mathbb{E}\left[X_{1:2}\right]\right)/2;\\ \lambda_3 &= \left(\mathbb{E}\left[X_{3:3}\right] - 2\mathbb{E}\left[X_{2:3}\right] + \mathbb{E}\left[X_{1:3}\right]\right)/3.\end{aligned}$$

Using formula (4.1.1), L-moments can be expressed using the quantile function $\mathbb{F}^{-1}$ (see Proposition 1.1. from Decurninge [2015]) as follows:

$$\lambda_r = \int_0^1 \mathbb{F}^{-1}(u) L_{r-1}(u) du \qquad \forall r \geq 1,$$

where L_r is the shifted Legendre polynomial of order r and is given by:

$$L_r(u) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} u^k.$$

Moreover, for $r \geq 2$:

$$\lambda_r = -\int_{\mathbb{R}} K_r(t) d\mathbb{F}^{-1}(t), \tag{4.1.2}$$

where

$$K_r(t) = \int_0^t L_{r-1}(u) du = \sum_{k=0}^{r-1} \frac{(-1)^{r-k}}{k+1} \binom{r}{k} \binom{r+k}{k} t^{k+1} \tag{4.1.3}$$

is the integrated shifted Legendre polynomial (see Proposition 1.2 in Decurninge [2015]). Notice that L-moments are polynomials in the cdf and linear in the quantile measure.

4.1.2 Semiparametric Linear Quantile Models (SPLQ)

SPLQ models were introduced by Broniatowski and Decurninge [2016]. The definition passes by the quantile measures instead of the distribution function. It is possible to define semiparametric models subject to L-moments constraints using the distribution function. However, their estimation would be very difficult because the constraints are not linear in the distribution function. They are instead linear in the quantile measure. This will become clearer as we go further in this subject. Denote M^{-1} the set of all $\sigma-$finite measures.

Definition 4.1.2. *A semiparametric linear quantile model related to some quantile measure $\boldsymbol{F}_T^{-1}$ is a collection of quantile measures absolutely continuous with respect to $\boldsymbol{F}_T^{-1}$ sharing the same form of L-moments, i.e.*

$$\mathcal{M} = \bigcup_{\alpha \in \mathcal{A}} \left\{ \boldsymbol{F}^{-1} \ll \boldsymbol{F}_T^{-1},\ s.t.\ \int_0^1 K_r(t) \boldsymbol{F}^{-1}(dt) = m(\alpha) \right\},$$

where $m(\alpha) = (-\lambda_2, \cdots, -\lambda_\ell)$ and $\alpha \in \mathcal{A} \subset \mathbb{R}^s$.

Example 4.1.1 (Decurninge [2015])**.** Consider the model which is the family of all the distributions of a r.v. X whose second, third and fourth L-moments satisfy:

$$\begin{aligned}\lambda_2 &= \sigma\left(1 - 2^{-1/\nu}\right)\Gamma\left(1 + \frac{1}{\nu}\right)\\ \lambda_3 &= \lambda_2\left[3 - 2\frac{1 - 3^{1/\nu}}{1 - 2^{-1/\nu}}\right]\\ \lambda_4 &= \lambda_2\left[6 + \frac{5(1 - 4^{-1/\nu}) - 10(1 - 3^{-1/\nu})}{1 - 2^{-1/\nu}}\right],\end{aligned}$$

for $\sigma > 0$, $\nu > 0$. These distributions share their first L-moments of order 2, 3 and 4 with those of a Weibull distribution with scale and shape parameter; σ, ν.

In SPLQ models, the objective is to estimate the value of α^* for which the true quantile measure $\mathbf{F}_T^{-1}$ of the data belongs to $\mathcal{M}_{\alpha^*}$ on the basis of a sample $X_1, \cdots, X_n$. The estimation procedure is generally done by either solving the set of equations defining the constraints or by minimizing a suitable distance-like function between the set $\mathcal{M}$ and some estimator of $\mathbf{F}_T^{-1}$ based on an observed sample. In other words, we search for the "projection" of $\mathbf{F}_T^{-1}$ on $\mathcal{M}$.
We have seen in the previous chapter that $\varphi-$divergences offer a way to calculate a "projection" of a finite signed measure on a set of finite signed measures, see definitions 3.2.2 and 3.2.3. $\varphi-$divergences can still be used to identify some distance between a $\sigma-$finite measure and a set of $\sigma-$finite measures. We may write:

$$\begin{aligned} \alpha^* &= \arg\inf_{\alpha\in\mathcal{A}} D_\varphi\left(\mathcal{M}_\alpha, \mathbf{F}_T^{-1}\right) \\ &= \arg\inf_{\alpha\in\mathcal{A}} \inf_{\mathbf{F}^{-1}\in\mathcal{M}_\alpha} D_\varphi\left(\mathbf{F}^{-1}, \mathbf{F}_T^{-1}\right). \end{aligned} \tag{4.1.4}$$

Of course, if $\mathbf{F}_T^{-1} \in \mathcal{M}_{\alpha^*}$ for some $\alpha^* \in \mathcal{A}$, then $D_\varphi\left(\cup_\alpha \mathcal{M}_\alpha, \mathbf{F}_T^{-1}\right) = 0$. Otherwise, α^* corresponds to the parameter of the closest set $\mathcal{M}_\alpha$ from the $\varphi-$divergence point of view to the quantile measure $\mathbf{F}_T^{-1}$.

4.1.3 Estimation using the duality technique

The estimation procedure (4.1.4) is not feasible because it concerns the minimization over a subset of possibly infinite dimension. The duality technique presented in paragraph 3.2.2 can be applied here too in order to transform the calculus of the projection from an optimization problem over a possibly infinite dimensional space into an optimization problem over $\mathbb{R}^{\ell-1}$, where $\ell - 1$ is the number of constraints defining the set $\mathcal{M}_\alpha$. We recall briefly this techniques by applying it directly in the context of quantile measures. Corollary 1.1 from Decurninge [2015] states the following. If there exists some $\mathbf{F}^{-1}$ in $\mathcal{M}_\alpha$ such that $a_\varphi < d\mathbf{F}^{-1}/d\mathbf{F}_T^{-1} < b_\varphi$ $\mathbf{F}_T^{-1}$-a.s. where $\text{dom}\varphi = (a_\varphi, b_\varphi)$ then,

$$\inf_{\mathbf{F}^{-1}\in\mathcal{M}_\alpha} \int_0^1 \varphi\left(\frac{d\mathbf{F}^{-1}}{d\mathbf{F}_T^{-1}}\right)(u)\mathbf{F}_T^{-1}(du) = \sup_{\xi\in\mathbb{R}^{\ell-1}} \xi^t m(\alpha) - \int_0^1 \psi\left(\xi^t K(u)\right)\mathbf{F}_T^{-1}(du). \tag{4.1.5}$$

This formula permits to build a plug-in estimate for α by considering a sample $X_1, \cdots, X_n$, see Remark 1.15 in Decurninge [2015].

$$\hat{\alpha} = \arg\inf_{\alpha} \sup_{\xi\in\mathbb{R}^{\ell-1}} \xi^t m(\alpha) - \sum_{i=1}^{n-1} \psi\left(\xi^t K\left(\frac{i}{n}\right)\right)(X_{i+1:n} - X_{i:n}). \tag{4.1.6}$$

This plug-in estimate is very interesting in its own, because it does not need any numerical integration. Besides, if we take φ to be the χ^2 generator, i.e. $\varphi(t) = (t-1)^2/2$ whose convex conjugate is $\psi(t) = t^2/2 + t$, the optimization over ξ can be solved directly in a similar way to Example 3.3.2, see also Example 1.12 in Decurninge [2015]. We will get back to this interesting case study later on.

4.2 Semiparametric two-component mixture models when one component is defined through L-moments constraints

4.2.1 Definition and identifiability

Definition 4.2.1. *Let X be a random variable taking values in $\mathbb{R}$ distributed from a probability measure P whose cdf is $\mathbb{F}$. We say that $P(.|\phi)$ with $\phi = (\lambda, \theta, \alpha)$ is a two-component semiparametric mixture model subject to L-moments constraints if it can be written as follows:*

$$\begin{aligned} P(.|\phi) &= \lambda P_1(.|\theta) + (1-\lambda)P_0 \quad s.t. \\ \boldsymbol{F}_0^{-1} \in \mathcal{M}_\alpha &= \left\{ \boldsymbol{Q}^{-1} \in M^{-1}, \boldsymbol{Q}^{-1} \ll \boldsymbol{F}_0^{-1} \ s.t. \ \int_0^1 K(u)\boldsymbol{Q}^{-1}(du) = m(\alpha) \right\} \end{aligned} \tag{4.2.1}$$

for $\lambda \in (0,1)$ the proportion of the parametric component, $\theta \in \Theta \subset \mathbb{R}^d$ a set of parameters defining the parametric component, $\alpha \in \mathcal{A} \subset \mathbb{R}^s$ is the constraints parameter, $K = (K_2, ..., K_\ell)$ is defined through formula (4.1.3) and finally $m(\alpha) = (m_2(\alpha), \cdots, m_\ell(\alpha))$ is a vector-valued function determining the values of the L-moments.

Notice that $m(\alpha)$ must contain the negative values of the L-moments by equation (4.1.2), i.e $m_r(\alpha) = -\lambda_r$. In this definition, it may appear that we have mixed quantiles with probabilities. This is however necessary in order to show the structure of the mixture model which generates the data. This structure is uniquely defined through the distribution function and does not have a "proper" writing using the quantile measure. In general, there is no formula which gives the quantile of a mixture model, and in practice, statisticians use approximations to calculate the quantile of a mixture model. Thus, working with the quantiles will make us lose the linearity property relating the two components with the mixture's distribution. In the previous chapter, this linearity played an essential role in the estimation procedure and simplified the calculus of the estimator on several levels. We will get back to this idea later on, and a "partial" solution will be proposed in order to get back to the mixture distribution instead of its quantile.
It is important to recall that the use of quantiles in the definition of semiparametric models subject to L-moments constraints stems from the fact that the constraints are linear functionals in the quantiles. Thus, an estimation procedure which employs the quantiles instead of the distribution function can be solved using the Fenchel-Legendre duality in a similar way to paragraph 4.1.3.

The identifiability of the model was not questioned in the context of SPLQ models because it suffices that the sets $\mathcal{M}_\alpha$ are disjoint (the function $m(\alpha)$ is one-to-one). However, in the context of this semiparametric mixture model, identifiability cannot be achieved only by supposing that the sets $\mathcal{M}_\alpha$ are disjoint.

Definition 4.2.2. *We say that the two-component semiparametric mixture model subject to L-moments constraints is identifiable if it verifies the following assertion. If*

$$\lambda P_1(.|\theta) + (1-\lambda)P_0 = \tilde{\lambda} P_1(.|\tilde{\theta}) + (1-\tilde{\lambda})\tilde{P}_0, \quad \text{with } \boldsymbol{F}_0^{-1} \in \mathcal{M}_\alpha, \tilde{\boldsymbol{F}}_0^{-1} \in \mathcal{M}_{\tilde{\alpha}}, \tag{4.2.2}$$

then $\lambda = \tilde{\lambda}, \theta = \tilde{\theta}$ and $P_0 = \tilde{P}_0$ (and hence $\alpha = \tilde{\alpha}$).

This is the same identifiability concept considered in Definition 3.3.2 (and by Bordes et al. [2006]) except that the unknown component's quantile belongs to the set $\mathcal{M}_\alpha$.

Proposition 4.2.1. *For a given mixture distribution $P_T = P(.|\phi^*)$ whose cdf is $\mathbb{F}_T$, suppose that the system of equations:*

$$\int_0^1 K(u)\left(\frac{1}{1-\lambda}\boldsymbol{F}_T - \frac{\lambda}{1-\lambda}\boldsymbol{F}_1(.|\theta)\right)^{-1}(du) = m(\alpha) \tag{4.2.3}$$

has a unique solution $(\lambda^*, \theta^*, \alpha^*)$. Then, equation (4.2.2) has a unique solution, i.e. $\lambda = \tilde{\lambda}, \theta = \tilde{\theta}$ and $P_0 = \tilde{P}_0$, and the semiparametric mixture model $P_T = P(.|\phi^*)$ is identifiable.

The proof is differed to Appendix 4.5.1. Note that the proof although has a close idea to the proof of Proposition 3.3.1 is different with more technical difficulties.

Example 4.2.1 (Two-component exponential mixture)**.** We propose to look at an exponential mixture defined by:

$$f(x|\lambda^*, a_1^*) = \lambda^* a_1^* e^{-a_1^* x} + (1-\lambda^*)a_0^* e^{-a_0^* x}$$

where $a_1^* = 1.5, a_0^* = 0.5$ and $\lambda^* \in \{0.3, 0.5, 0.7, 0.85\}$. This is considered to be the distribution generating the observed data. Suppose that the second component $f_0^*(x) = a_0^* e^{-a_0^* x}$ is unknown during the estimation. Furthermore, suppose that we hold an information about f_0^* that its quantile $\mathbf{F}_0^{*-1}$ belongs to the following class of functions:

$$\mathcal{M} = \left\{\mathbf{F}^{-1} \ll \mathbf{F}_0^{*-1}, \quad \int_0^1 u(1-u)\mathbf{F}^{-1}(du) = \frac{1}{2a_0^*}\right\}.$$

This set contains all probability distributions whose second L-moment has the value $\frac{1}{2a_0^*}$. We would like to check the identifiability of the semiparametric mixture model subject to the second L-moment constraint of the exponential distribution $\mathcal{E}(a_0^*)$. The system of equations (4.2.3) is given by:

$$\int_0^1 u(1-u)\left(\frac{1}{1-\lambda}\mathbf{F}_T - \frac{\lambda}{1-\lambda}\mathbf{F}_1(.|a_1)\right)^{-1}(du) = \frac{1}{2a_0^*}.$$

In order to calculate the left hand side, we use the alternative definition of the second L-moment $\lambda_2 = \left(\mathbb{E}[X_{2:2}] - \mathbb{E}[X_{1:2}]\right)/2$ and exploit formula (4.1.1). We have

$$\int_0^1 u(1-u)\left(\frac{1}{1-\lambda}\mathbf{F}_T - \frac{\lambda}{1-\lambda}\mathbf{F}_1(.|a_1)\right)^{-1}(du) = \int_{\mathbb{R}_+} x\left[2\left(\frac{1}{1-\lambda}\mathbb{F}_T(x) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(x|a_1)\right) - 1\right]\left(\frac{1}{1-\lambda}p_T(x) - \frac{\lambda}{1-\lambda}p_1(x|a_1)\right)dx$$

A direct calculus of the right hand side shows:

$$\int_{\mathbb{R}_+} x\left[2\left(\frac{1}{1-\lambda}\mathbb{F}_T(x) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(x|a_1)\right) - 1\right]\left(\frac{1}{1-\lambda}p_T(x) - \frac{\lambda}{1-\lambda}p_1(x|a_1)\right)dx = \frac{2C_1 - (\lambda+1)C_2}{(1-\lambda)^2} + \frac{\lambda^2 - 2\lambda}{2a_1(1-\lambda)^2} + \frac{2\lambda^*\lambda}{(1-\lambda)^2(a_1+a_1^*)} + \frac{2\lambda(1-\lambda^*)}{(1-\lambda)^2(a_1+a_0^*)}$$

where

$$\begin{aligned} C2 &= \frac{\lambda^*}{a_1^*} + \frac{1-\lambda^*}{a_0^*} \\ C1 &= \frac{\lambda^*}{a_1^*} + \frac{1-\lambda^*}{a_0^*} - \frac{\lambda^{*2}}{4a_1^*} - \frac{(1-\lambda^*)^2}{4a_0^*} - \frac{\lambda^*(1-\lambda^*)}{a_1^*+a_0^*}. \end{aligned}$$

In figure (4.1), we show the set of solutions of the following equation:

$$\frac{2C_1-(\lambda+1)C_2}{(1-\lambda)^2}+\frac{\lambda^2-2\lambda}{2a_1(1-\lambda)^2}+\frac{2\lambda^*\lambda}{(1-\lambda)^2(a_1+a_1^*)}+\frac{2\lambda(1-\lambda^*)}{(1-\lambda)^2(a_1+a_0^*)}=\frac{1}{2a_0^*}, \quad (4.2.4)$$

for several values of λ^* in the figure to the left. The figure to the right shows the intersection between the set of solutions and the set $\Phi^+ = \{(\lambda, a), \text{ s.t. } \frac{1}{1-\lambda}\mathbb{F}_T(x) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(x|a_1) \text{ is a cdf}\}$. It is clear that the nonlinear system of equations (4.2.3) has an infinite number of solutions. In order to reduce the number of solutions into one, we need to consider another L-moment constraint. We do not pursue this here because the calculus is already complicated even in this simple model.
Note that the set of solutions is shrinking as the proportion of the unknown component f_0 becomes smaller (the value of λ^* increases). This gives rise to a difficult and an important question; what happens if we have a number of constraints inferior to the number of parameters. This question is not pursued here.

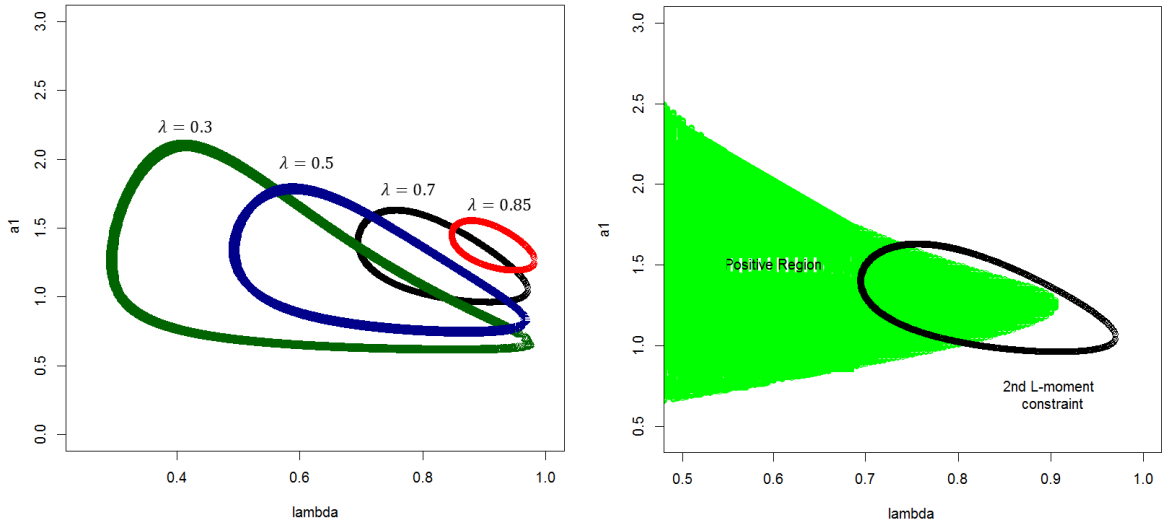

Figure 4.1: The set of solutions under a constraint over the second L-moment. Each closed trajectory corresponds to a value of the proportion of the parametric part indicated above of it. The figure to the left represents the whole set of solutions of the equation (4.2.4) for different values of the true proportion λ^*. The figure to the right represents the intersection between the set of solutions of equation (4.2.4) for $\lambda^* = 0.7$ with the set Φ^+.

4.2.2 An algorithm for the estimation of the semiparametric mixture model

In the context of our semiparametric mixture model, we want to estimate the parameters $(\lambda, \theta, \alpha)$ on the basis of two pieces of information; an i.i.d. sample $X_1, \cdots, X_n$ drawn from P_T and the fact that $\mathbf{F}_0^{*-1}$ belongs to the set $\mathcal{M}$. For SPLQ models, we have seen that using $\varphi-$divergences, we were able to construct an estimation procedure by minimizing some distance between the set of constraints and the distribution generating the data. The resulting estimation procedure is an optimizing problem over an infinite dimensional space. We exploited the linearity of the constraints and transformed the estimation procedure into a feasible optimization problem over $\mathbb{R}^{\ell-1}$ using the Fenchel-Legendre duality.

In order to use the Fenchel-Legendre duality, the constraints need to apply over the whole mixture. In our semiparametric mixture model, the constraints apply over the quantile of only one component; $\mathbf{F}_0^{-1}$. We thus propose to define another "model" based on $\mathbf{F}_0^{-1}$. We have:

$$\mathbb{F}_0^{*-1} = \left(\frac{1}{1-\lambda^*} \mathbb{F}_T(.|\phi^*) - \frac{\lambda^*}{1-\lambda^*} \mathbb{F}_1(|\theta^*) \right)^{-1}.$$

Denote the associated quantile measure

$$\mathbf{F}_0^{*-1} = \left(\frac{1}{1-\lambda^*} \mathbf{F}_T(.|\phi^*) - \frac{\lambda^*}{1-\lambda^*} \mathbf{F}_1(|\theta^*) \right)^{-1}.$$

Define the set $\mathcal{N}^{-1}$ by:

$$\mathcal{N}^{-1} = \left\{ \mathbf{Q}^{-1} \in M^{-1} \ : \ \exists (\lambda, \theta) \in (0,1) \times \Theta \text{ s.t. } \mathbf{Q}^{-1} = \left(\frac{1}{1-\lambda} \mathbf{F}_T - \frac{\lambda}{1-\lambda} \mathbf{F}_1(.|\theta) \right)^{-1} \right\}.$$

Notice that the set $\mathcal{N}^{-1}$ here is different from the set $\mathcal{N}$ defined in the previous chapter by (3.3.6). Here, not all the couples (λ, θ) in $(0,1) \times \Theta$ are accepted, because function $\frac{1}{1-\lambda}\mathbb{F}_T - \frac{\lambda}{1-\lambda}\mathbb{F}_1(|\theta)$ may not be a cdf for for these couples. Define the set of effective parameters Φ^+ by:

$$\Phi^+ = \left\{ (\lambda, \theta) \in (0,1) \times \Theta : \frac{1}{1-\lambda} \mathbb{F}_T - \frac{\lambda}{1-\lambda} \mathbb{F}_1(.|\theta) \text{ is a cdf} \right\}. \tag{4.2.5}$$

Now, the set $\mathcal{N}^{-1}$ can be characterized using Φ^+ by:

$$\mathcal{N}^{-1} = \left\{ \left(\frac{1}{1-\lambda} \mathbf{F}_T - \frac{\lambda}{1-\lambda} \mathbf{F}_1(.|\theta) \right)^{-1}, \text{ for } (\lambda, \theta) \in \Phi^+ \right\}.$$

The introduction of the set Φ^+ is only temporary, and we will not need it at the end of this section in order to build our estimation procedure. Notice now that $\mathbf{F}_0^{*-1}$ is a member of $\mathcal{N}^{-1}$ for $(\lambda, \theta) = (\lambda^*, \theta^*)$. On the other hand, and by definition of the semiparametric mixture model, $\mathbf{F}_0^{*-1} \in \mathcal{M}_{\alpha^*}$. We may write:

$$\mathbf{F}_0^{*-1} \in \mathcal{N}^{-1} \bigcap \cup_\alpha \mathcal{M}_\alpha. \tag{4.2.6}$$

If we suppose that the intersection (which is not void) contains only one element which will be $\mathbf{F}_0^{*-1}$ (see paragraph 4.2.5 for a discussion on the uniqueness), then it becomes reasonable to consider an estimation procedure by calculating a "distance" between the two sets $\cup_\alpha \mathcal{M}_\alpha$ and $\mathcal{N}^{-1}$. Using definition 3.2.2, we may write:

$$\begin{aligned} D_\varphi \left(\cup_\alpha \mathcal{M}_\alpha, \mathcal{N}^{-1} \right) &= \inf_{\mathbf{Q}^{-1} \in \mathcal{N}^{-1}} \inf_{\mathbf{F}_0^{-1} \in \cup_\alpha \mathcal{M}_\alpha} D_\varphi \left(\mathbf{F}_0^{-1}, \mathbf{Q}^{-1} \right) \\ &= \inf_{(\lambda,\theta) \in \Phi^+, \alpha \in \mathcal{A}} \inf_{\mathbf{F}_0^{-1} \in \mathcal{M}_\alpha} D_\varphi \left(\mathbf{F}_0^{-1}, \left(\frac{1}{1-\lambda} \mathbf{F}_T - \frac{\lambda}{1-\lambda} \mathbf{F}_1(.|\theta) \right)^{-1} \right). \end{aligned} \tag{4.2.7}$$

Now by virture of (4.2.6), it holds that

$$(\lambda^*, \theta^*, \alpha^*) \in \underset{(\lambda,\theta,\alpha) \in \Phi^+}{\arg\inf} \inf_{\mathbf{F}_0^{-1} \in \mathcal{M}_\alpha} D_\varphi \left(\mathbf{F}_0^{-1}, \left(\frac{1}{1-\lambda} \mathbf{F}_T - \frac{\lambda}{1-\lambda} \mathbf{F}_1(.|\theta) \right)^{-1} \right). \tag{4.2.8}$$

Next, we will treat this estimation procedure using the Fenchel duality in order to write a feasible optimization procedure, and then proceed to build upon a plug-in estimator based on an observed dataset $X_1, \cdots, X_n$.

4.2.3 Estimation using the duality technique

Applying the duality result (4.1.5) on the estimation procedure (4.2.7) gives:

$$D_\varphi\left(\cup_\alpha \mathcal{M}_\alpha, \mathcal{N}^{-1}\right) = \inf_{(\lambda,\theta,\alpha)\in\Phi^+} \sup_{\xi\in\mathbb{R}^{\ell-1}} \xi^t m(\alpha) - \int_0^1 \psi\left(\xi^t K(u)\right)\left(\frac{1}{1-\lambda}\mathbf{F}_T - \frac{\lambda}{1-\lambda}\mathbf{F}_1(.|\theta)\right)^{-1}(du). \tag{4.2.9}$$

In order to keep formulas clearer, we adapt the following notation:

$$\mathbb{F}_0(y|\phi) = \frac{1}{1-\lambda}\mathbb{F}_T(y) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(y|\theta)$$

Note that we must ensure the integrability condition

$$\int \|K\left(\mathbb{F}_0(y|\phi)\right)\| dx < \infty,$$

in order to be able to use the duality technique. This is ensured by the definition of the polynomial vector K. Indeed, there exists a constant c such that:

$$\|K\left(\mathbb{F}_0(y|\phi)\right)\| \leq c\left(\mathbb{F}_0(y|\phi)\right)\left(1-\mathbb{F}_0(y|\phi)\right).$$

Since $\mathbb{F}_0(y|\phi) = \left(\frac{1}{1-\lambda}\mathbb{F}_T(y) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(y|\theta)\right)$ is supposed here to be a cdf because $(\lambda,\theta,\alpha)\in\Phi^+$, it suffices then that $\mathbb{F}_0(y|\phi)$ has a finite expectation so that the previous integral becomes finite.
This formulation is only useful when one has a sample of i.i.d. observations of the distribution $\frac{1}{1-\lambda}P_T - \frac{\lambda}{1-\lambda}P_1(.|\theta)$ for every λ and θ, because the integral can be approximated directly using the order statistics as in formula (4.1.6). We need, however, a formula which shows directly the cdf because it would permit to approximate directly the objective function and avoid the calculus of the inverse of $\frac{1}{1-\lambda}\mathbb{F}_T - \frac{\lambda}{1-\lambda}\mathbb{F}_1(.|\theta)$. Besides, the replacement of the true cdf by the empirical one does not guarantee that the difference $\frac{1}{1-\lambda}\mathbb{F}_T - \frac{\lambda}{1-\lambda}\mathbb{F}_1(.|\theta)$ remains a cdf and more complications would appear in the proof of the consistency.
Using Lemma 1.2 from Decurninge [2015] we may make the change of variable desired.

$$\int_0^1 \psi\left(\xi^t K(u)\right)\left(\frac{1}{1-\lambda}\mathbf{F}_T - \frac{\lambda}{1-\lambda}\mathbf{F}_1(.|\theta)\right)^{-1}(du) = \int_{\mathbb{R}} \psi\left(\xi^t K\left(\frac{1}{1-\lambda}\mathbb{F}_T(x) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(x|\theta)\right)\right)dx. \tag{4.2.10}$$

Employing (4.2.10) and (4.2.9) in (4.2.8), we may write:

$$(\lambda^*,\theta^*,\alpha^*) \in \arg\inf_{(\lambda,\theta,\alpha)\in\Phi^+} \sup_{\xi\in\mathbb{R}^{\ell-1}} \xi^t m(\alpha) - \int_{\mathbb{R}} \psi\left(\xi^t K\left(\frac{1}{1-\lambda}\mathbb{F}_T(x) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(x|\theta)\right)\right)dx. \tag{4.2.11}$$

We may now construct an estimator of ϕ^* by replacing $\mathbb{F}_T$ by the empirical cdf calculated on the basis of an i.i.d. sample $X_1,\cdots,X_n$. The resulting estimation procedure is still very complicated. This is mainly because we need to characterize the set Φ^+. It is possible but is very expensive. For example, we may think about checking if the derivative with respect to x $\frac{1}{1-\lambda}p_T(x) - \frac{\lambda}{1-\lambda}p_1(x|\theta)$ is non negative at a large randomly selected set of points. On the other hand, the set Φ^+ can take *fearful* forms for some mixtures. In Figure (4.2), we have two examples of Φ^+. In the exponential mixture (the figure to the left), Φ^+ has a "good" form in the sense that it is convex and contains $(\lambda^*,\theta^*) = (0.7, 1.5)$ with a sufficiently large neighborhood around it. Thus, optimization procedures should not face

any problem finding the optimum. However, in the Weibull-Lognormal mixture (the figure to the right) with $(\lambda^*, \mu^*) = (0.7, 3)$, the set Φ^* is not even connected. Besides, there is not a sufficient neighborhood around (λ^*, μ^*) which permits an optimization algorithm to move around. During my simulations on data generated from a Weibull-Lognormal mixture distribution, the optimization algorithms could not reach such optimum and were always stuck at the initial point. A solution will be proposed in the next paragraph where we introduce the final step in the sequel of this estimation procedure.

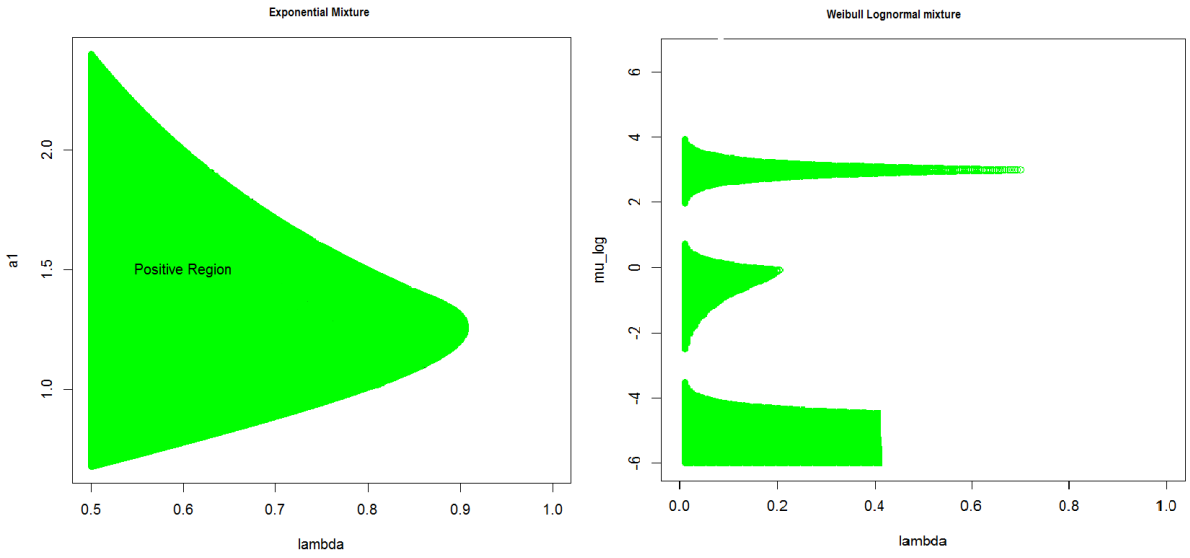

Figure 4.2: Different forms of the set Φ^+. For the Weibull-Lognormal mixture, the Weibull is the semiparametric component.

4.2.4 The algorithm in practice and a plug-in estimate

The problem with the estimation procedure (4.2.11) is that the optimization is over the set Φ^+ which may take "non-practical forms" as explained in the previous paragraph. The problem can be reread otherwise. The difficulty comes mainly from the fact that function $\frac{1}{1-\lambda}\mathbb{F}_T - \frac{\lambda}{1-\lambda}\mathbb{F}_1(.|\theta)$ may not be a cdf and the quantile would not exist. Thus, the estimation procedure in formula (4.2.8) cannot be used. We have, however, made disappear the quantiles in formula (4.2.11) using a change of variable. Besides, there is no problem in calculating the optimized function in formula (4.2.11) for any triplet $(\lambda, \theta, \alpha) \in \Phi$ even if the parameters do not define a proper cdf for function $\frac{1}{1-\lambda}\mathbb{F}_T - \frac{\lambda}{1-\lambda}\mathbb{F}_1(.|\theta)$. Besides, and more importantly, ϕ^* is a global infimum of the objective function $H(\phi, \xi(\phi))$ over the whole set Φ (and not only overy Φ^+) where:

$$H(\phi, \xi) = \xi^t m(\alpha) - \int \psi \left[\xi^t K \left(\frac{1}{1-\lambda}\mathbb{F}_T(y) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(y|\theta) \right) \right] dy$$

and $\xi(\phi) = \arg\sup_{\xi \in \mathbb{R}^{\ell-1}} H(\phi, \xi)$. Indeed, for any $\phi \in \Phi$, we have:

$$H(\phi, \xi(\phi)) \geq H(\phi, 0) = 0.$$

Besides, using the duality attainment at $\phi = \phi^*$, we may write

$$\begin{aligned} H(\phi^*, \xi(\phi^*)) &= \inf_{\mathbf{F}_0^{-1} \in \mathcal{M}_{\alpha^*}} D_\varphi \left(\mathbf{F}_0^{-1}, \left(\frac{1}{1-\lambda^*} \mathbf{F}_T - \frac{\lambda^*}{1-\lambda^*} \mathbf{F}_1(.|\theta^*) \right)^{-1} \right) \\ &= \inf_{\mathbf{F}_0^{-1} \in \mathcal{M}_{\alpha^*}} D_\varphi \left(\mathbf{F}_0^{-1}, \mathbf{F}_0^{*-1} \right) \\ &= 0. \end{aligned}$$

Thus, if function $H(\phi, \xi(\phi))$ does not have several global infima inside Φ, $(\lambda^*, \theta^*, \alpha^*)$ will hold as the only global minimum of it. In other words

$$\phi^* = \underset{(\alpha,\theta,\lambda)\in\Phi}{\arg\inf} \sup_{\xi \in \mathbb{R}^{\ell-1}} \xi^t m(\alpha) - \int \psi \left[\xi^t K \left(\frac{1}{1-\lambda} \mathbb{F}_T(x) - \frac{\lambda}{1-\lambda} \mathbb{F}_1(x|\theta) \right) \right] dx. \quad (4.2.12)$$

Provided an i.i.d. sample $X_1, \cdots, X_n$ distributed from P_T, the cdf $\mathbb{F}_T$ can be approximated by its empirical version $\frac{1}{n} \sum \mathbb{1}_{X_i \leq x}$. Hence, ϕ^* can be estimated by:

$$\hat{\phi} = \underset{(\alpha,\theta,\lambda)\in\Phi}{\arg\inf} \sup_{\xi \in \mathbb{R}^{\ell-1}} \xi^t m(\alpha) - \int \psi \left[\xi^t K \left(\frac{1}{1-\lambda} \mathbb{F}_n(x) - \frac{\lambda}{1-\lambda} \mathbb{F}_1(x|\theta) \right) \right] dx. \quad (4.2.13)$$

Remark 4.2.1. Notice that the dual attainment no longer holds on the complementary set $\Phi \setminus \Phi^+$ since we are working with "signed cumulative functions". Our idea is to offer the optimization algorithm a larger neighborhood around the optimum in order to be able to find it. The important fact in the extended procedure is that ϕ^* is a *global* infimum of the objective function. Our simulation study shows that the extension to Φ does not affect the results in several examples, and the estimator $\hat{\phi}$ is not biased and has an acceptable variance, see Section 4.4 for more details.

4.2.5 Uniqueness of the solution "under the model"

By a unique solution we mean that only one quantile measure, which can be written in the form of $\left(\frac{1}{1-\lambda} \mathbf{F}_T - \frac{\lambda}{1-\lambda} \mathbf{F}_1(.|\theta) \right)^{-1}$ for $(\lambda, \theta) \in \Phi^+$, verifies the L-moments constraints with a unique triplet $(\lambda^*, \theta^*, \alpha^*)$. The existence of a unique solution is essential in order to ensure that the procedure (4.2.11) is a reasonable estimation method. We provide next a result ensuring the uniqueness of the solution. The proof is differed to Appendix 4.5.2. The proof does not provide sufficient conditions for the existence of a unique solution over Φ because in the proof we only study the intersection $\mathcal{N}^{-1} \cap \mathcal{M}$ and characterize it without using the Fenchel duality.

Proposition 4.2.2. *Assume that* $\boldsymbol{F}_0^{*-1} \in \mathcal{M} = \cup_\alpha \mathcal{M}_\alpha$. *Suppose also that:*

1. *the system of equations:*

$$\int_0^1 K(u) \left(\frac{1}{1-\lambda} \boldsymbol{F}_T - \frac{\lambda}{1-\lambda} \boldsymbol{F}_1(.|\theta) \right)^{-1} (du) = m(\alpha) \quad (4.2.14)$$

   has a unique solution $(\lambda^*, \theta^*, \alpha^*)$*;*

2. *the function* $\alpha \mapsto m(\alpha)$ *is one-to-one;*

3. *for any* $\theta \in \Theta$ *we have :*

$$\lim_{x \to \infty} \frac{p_1(x|\theta)}{p_T(x)} = c, \qquad \text{with } c \in [0, \infty) \setminus \{1\};$$

4. *the parametric component is identifiable, i.e. if* $p_1(.|\theta) = p_1(.|\theta')$ $dP_T-a.e.$ *then* $\theta = \theta'$,

then, the intersection $\mathcal{N}^{-1} \cap \mathcal{M}$ *contains a unique measure* $\boldsymbol{F}_0^{*-1}$*, and there exists a unique vector* $(\lambda^*, \theta^*, \alpha^*)$ *such that* $P_T = \lambda^* P_1(.|\theta^*) + (1-\lambda) P_0^*$ *where* P_0^* *is given by (3.3.5) and belongs to* $\mathcal{M}_{\alpha^*}$*. Moreover, provided assumptions 2-4, the conclusion holds if and only if assumption 1 is fulfilled.*

There is no general result for a non linear system of equations to have a unique solution; still, it is necessary to ensure that we impose a number of constraints at least equal to the number of unknown variables, otherwise there would be an infinite number of $\sigma-$finite measures in the intersection $\mathcal{N}^{-1} \bigcap \cup_{\alpha \in \mathcal{A}} \mathcal{M}_\alpha$.

Remark 4.2.2. Assumptions 3 and 4 of Proposition 3.3.2 are used to prove the identifiability of the "model" $\left(\frac{1}{1-\lambda} P_T - \frac{\lambda}{1-\lambda} P_1(.|\theta)\right)_{\lambda,\theta}$. These conditions may be rewritten using the cdf. Furthermore, according to the considered situation we may find simpler ones for particular cases (or even for the general case). Our assumptions remain sufficient but not necessary for the proof. Note also that similar assumption to 3 can be found in the literature on semiparametric mixture models, see Proposition 3 in Bordes et al. [2006].

4.3 Asymptotic properties

We study the asymptotic properties of the estimator $\hat{\phi}$ defined by (4.2.13). For the consistency, we will assume that function $H(\phi, \xi(\phi))$ has a unique infimum on Φ. This infimum is a fortiori ϕ^*. On the other hand, the limiting law would not change if the infimum is truly ϕ^* or any other point. $\hat{\phi}$ will be centered at the infimum with a multivariate Gaussian limit law. It would not, however, be interesting unless it is centered around ϕ^*.

4.3.1 Consistency

We will use Theorem 3.4.1 since we are in the same context of double optimization. This time, function H_n does not have the form of $P_n h$. Let's start by precising the functions H and H_n.

$$\begin{aligned} H(\phi, \xi) &= \xi^t m(\alpha) - \int \psi \left[\xi^t K \left(\frac{1}{1-\lambda} \mathbb{F}_T(y) - \frac{\lambda}{1-\lambda} \mathbb{F}_1(y|\theta) \right) \right] dy; \\ H_n(\phi, \xi) &= \xi^t m(\alpha) - \int \psi \left[\xi^t K \left(\frac{1}{1-\lambda} \mathbb{F}_n(y) - \frac{\lambda}{1-\lambda} \mathbb{F}_1(y|\theta) \right) \right] dy, \end{aligned}$$

and recall the notations:

$$\begin{aligned} \mathbb{F}_0(y|\phi) &= \frac{1}{1-\lambda} \mathbb{F}_T(y) - \frac{\lambda}{1-\lambda} \mathbb{F}_1(y|\theta); \\ \hat{\mathbb{F}}_0(y|\phi) &= \frac{1}{1-\lambda} \mathbb{F}_n(y) - \frac{\lambda}{1-\lambda} \mathbb{F}_1(y|\theta); \\ \xi(\phi) &= \underset{\xi \in \mathbb{R}^{\ell-1}}{\arg\sup}\ H(\phi, \xi); \\ \xi_n(\phi) &= \underset{\xi \in \mathbb{R}^{\ell-1}}{\arg\sup}\ H_n(\phi, \xi). \end{aligned}$$

We start by calculating the difference $H(\phi,\psi) - H_n(\phi,\psi)$.

$$H(\phi,\xi) - H_n(\phi,\xi) = \int \psi\left[\xi^t K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right] - \psi\left[\xi^t K\left(\mathbb{F}_0(y|\phi)\right)\right] dy. \tag{4.3.1}$$

The following lemma is essential for the proof of the consistency. We need to transform the optimization over ξ onto a compact set. Thus, *important* values of ξ which are necessary for the calculus of the supremum are bounded. The proof is differed to Appendix 4.5.3.

Lemma 4.3.1. *Suppose that function $\xi \mapsto H(\phi,\xi)$ is of class $\mathcal{C}^2(\mathbb{R}^{\ell-1})$. Then, functions $\phi \mapsto \xi(\phi)$ and $\phi \mapsto \xi_n(\phi)$ are well defined and $\mathcal{C}^1$ on the interior of the whole set Φ. Moreover, if Φ is compact, then $\hat{\phi}$ and ϕ^* exist and the sets $Im(\xi(.))$ and $Im(\xi_n(.))$ are compact.*

Differentiability of function H with respect to ξ can be checked in general using Lebesgue theorems, but it would not have been wise to impose an assumption over the integrand since ψ' is increasing and ξ is a priori in $\mathbb{R}^{\ell-1}$. For the class of functions of Cressie-Read (1.1.3), we have $\psi(t) = \frac{1}{\gamma}(\gamma t - t + 1)^{\gamma/(\gamma-1)} - \frac{1}{\gamma}$. Thus, for $\gamma > 1$, $\psi(t) \to \infty$ as $t \to \infty$. Therefore, it is important to study each special case alone. For example, $\psi(y) = y^2/2 + y$ is the dual of the Chi square generator $\varphi(t) = (t-1)^2/2$, then $H(\xi,\phi)$ is a polynomial of degree 2 in ξ and hence differentiable up to second order, see Example 4.3.1 below for more details.
We state the consistency of the estimator $\hat{\phi}$ defined by (4.2.13). The proof is based on Theorem 3.4.1 from the previous chapter and is differed to Appendix 4.5.4.

Theorem 4.3.1. *Suppose that*

C1. Φ is a compact subset of $\mathbb{R}^d$;

C2. function ψ is continuously differentiable;

C3. the infimum of $\phi \mapsto H(\phi,\xi(\phi))$ is unique and isolated, i.e. $\forall \varepsilon > 0, \forall \phi$ such that $\|\phi - \phi^*\| > \varepsilon$, there exists $\eta > 0$ such that $H(\phi,\xi(\phi)) - H(\phi^*,\xi(\phi^*)) > \eta$;

C4. function $\alpha \mapsto m(\alpha)$ is continuous;

C5. function $\xi \mapsto H(\phi,\xi)$ is of class $\mathcal{C}^2(\mathbb{R}^{\ell-1})$;

C6. the integral $\int \sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}dy$ is finite,

then the estimator $\hat{\phi}$ defined by (4.2.13) converges in probability to ϕ^*.

Remark 4.3.1. If we use $\tilde{\phi}$ defined by (4.2.11), only assumption C3 should be changed. We need to suppose that the infimum exists and is unique inside Φ^+ instead of the whole parameter space Φ. This is less restrictive than assumption C3 since we are working inside a subset of Φ.

Remark 4.3.2. Assumption C5 is used (together with assumption C1) in order to apply Lemma 4.3.1. As discussed earlier after Lemma 4.3.1, differentiability of function $\xi \mapsto H(\phi,\xi)$ may be very difficult to check using Lebesgue theorems. When $\psi(t) = t^2/2 + t$, function $H(\phi,\xi)$ is twice differentiably continuous as a function of ξ, because it is a polynomial of order 2 in ξ. Assumption C6 will be needed again in the proof of the asymptotic normality. Sufficient conditions are discussed in Remark 4.3.4 hereafter.

Example 4.3.1 (χ^2 case)**.** The case of the χ^2 divergence is very interesting similarly to the case of moment-type constraints (see Example 3.4.1) simply because the optimization over ξ can be calculated. Write function $H(\phi, \xi)$ for $\psi(t) = t^2/2 + t$.

$$H(\phi, \xi) = \xi^t m(\alpha) - \int \frac{1}{2} \left(\xi^t K\left(\mathbb{F}_0(y|\phi)\right)\right)^2 + \xi^t K\left(\mathbb{F}_0(y|\phi)\right) dy.$$

This is a polynomial of order 2 in ξ and thus $H(\phi, \xi)$ is of class $\mathcal{C}^2(\mathbb{R}^{\ell-1})$ as soon as the integrals exist. Indeed, for any $r \leq \ell$, there exists c_r such that:

$$\begin{aligned} \left|K_r\left(\mathbb{F}_0(y,|\phi)\right)\right| &\leq c_r \left|\mathbb{F}_0(y,|\phi)\left(1 - \mathbb{F}_0(y,|\phi)\right)\right| && (4.3.2) \\ &\leq \frac{c_r}{(1-\lambda)^2}\left[\mathbb{F}_T(y)\left(1 - \mathbb{F}_T(y)\right) + \lambda \mathbb{F}_T(y)(1 - \mathbb{F}_1(y)) + \lambda \mathbb{F}_1(1 - \mathbb{F}_T(y)) + \right. \\ &\quad \left. \lambda^2 \mathbb{F}_1(y)(1 - \mathbb{F}_1(y))\right]. && (4.3.3) \end{aligned}$$

For example, if the distributions $\mathbb{F}_T$ and $\mathbb{F}_1$ are defined on $\mathbb{R}_+$, then the right hand side is integrable as soon as the expectations of $\mathbb{F}_T$ and $\mathbb{F}_1$ are finite.
A simple calculus of the derivative of function $\xi \mapsto H(\phi, \xi)$ gives

$$\frac{\partial H}{\partial \xi}(\xi, \phi) \quad = \quad m(\alpha) \quad - \quad \int K\left(\mathbb{F}_0(y,|\phi)\right) \xi^t K\left(\mathbb{F}_0(y,|\phi)\right) dy \quad + \quad \int K\left(\mathbb{F}_0(y,|\phi)\right) dy.$$

The optimum is attained for:

$$\xi(\phi) = \Omega^{-1}\left(m(\alpha) - \int K(\mathbb{F}_0(y|\phi))dy\right),$$

where

$$\Omega = \int K\left(\mathbb{F}_0(y|\phi)\right) K\left(\mathbb{F}_0(y|\phi)\right)^t dy.$$

Furthermore, the Hessian matrix is equal to $-\Omega$, so it is symmetric definite negative whatever the value of the vector ϕ. Thus, $\xi(\phi)$ is a global maximum of function $\xi \mapsto H(\xi, \phi)$ for any $\phi \in \Phi$. This was not the case for moment-type constraints since the Hessian matrix might be definite positive for some values of the vector ϕ. The empirical version of this calculus is obtained similarly by replacing $\mathbb{F}_0(y|\phi)$ by $\hat{\mathbb{F}}_0(y|\phi)$.
Conditions of the consistency theorem can be verified. Assumption C1 is very natural in practice since in general, we have in mind a range of values for the parameters. Assumption C2 is fulfilled since $\psi(t)$ is polynomial of degree 2. Assumption 3 is not simple in general and depends on the model. Assumption C4 follows the problem we have. In Example 4.1.1, $m(\alpha) = (-\lambda_2, -\lambda_3, -\lambda_4)$ is continuous on $(0, \infty) \times (0, \infty)$, and assumption C4 becomes verified. We have verified assumption C5 at the beginning of the example. Assumption C6 is not restrictive. It is verified for example in an exponential mixture. The idea is to control the tail behavior of the distribution, see remark (4.3.4) for a general approach.

4.3.2 Asymptotic normality

The convergence in law of the estimator $\hat{\phi}$ defined by (4.2.13) is not simply deduced in the same way we obtained it in the moment-constraints case. A Taylor expansion of the gradient of function H would not show directly the empirical distribution which combined with the CLT gives the asymptotic normality. The expansion results in the term $\int K(\hat{\mathbb{F}}_0(x))dx$ which is a functional of the empirical distribution, that is

$$\begin{pmatrix} \sqrt{n}\left(\hat{\phi} - \phi^*\right) \\ \sqrt{n}\xi_n(\hat{\phi}) \end{pmatrix} = J_H^{-1} \begin{pmatrix} 0 \\ \sqrt{n}\left[m(\alpha^*) - \int K(\hat{\mathbb{F}}_0(y|\phi^*))dy\right] \end{pmatrix} + o_P(1).$$

In the case of simply one component (no parametric component) defined through L-moment constraints, Decurninge [2015] used a result based on Theorem 6 from Stigler [1974] in order to establish the limit law of $\sqrt{n}\left[m(\alpha^*) - \int K(\hat{\mathbb{F}}_0(y|\phi^*))dy\right]$. This result is based on sums of order statistics which *cannot* be adapted to our context since $\hat{\mathbb{F}}_0$ is an estimator of $\mathbb{F}_0$ different from the corresponding empirical distribution. We present a new result adapted to our context where the proof still shares a part of the idea of the proof of the result of Stigler [1974]. The proof is differed to Appendix 4.5.5.

Proposition 4.3.1. *Suppose that* $\mathbb{E}|X_i| < \infty$. *Suppose also that*

$$\int \sqrt{\mathbb{F}_T(y)\left(1-\mathbb{F}_T(y)\right)}dy \quad < \quad \infty, \tag{4.3.4}$$

$$\int\int \mathbb{F}_T(\min(x,y)) - \mathbb{F}_T(x)\mathbb{F}_T(y)dxdy \quad < \quad \infty. \tag{4.3.5}$$

For any vector $\phi = (\lambda, \theta, \alpha) \in \Phi$, *we then have*

$$\sqrt{n}\left[\int K\left(\frac{1}{1-\lambda}\mathbb{F}_n(y) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(y|\theta)\right)dy - \int K\left(\frac{1}{1-\lambda}\mathbb{F}_T(y) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(y|\theta)\right)dy\right] \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\Sigma\right),$$

where the covariance matrix Σ *is given by*

$$\Sigma_{r_1,r_2} = \int\int \left(\mathbb{F}_T\left(\min(x,y)\right) - \mathbb{F}_T(x)\mathbb{F}_T(y)\right)\sum_{k=0}^{r_1-1} c_{r_1,k}\mathbb{F}_0(x|\phi)^k \sum_{k=0}^{r_2-1} c_{r_2,k}\mathbb{F}_0(y|\phi)^k dydx, \tag{4.3.6}$$

and $c_{r,k} = (-1)^{r-k-1}\binom{r-1}{k}\binom{r+k-1}{k}$ *for* $r, r_1, r_2 \in \{2, \cdots, \ell\}$.

Remark 4.3.3. It was not possible to use a functional delta method (see van der Vaart [1998] Chap. 20, Theorem 20.8) in a similar way to Theorem 3.2 in Bordes and Vandekerkhove [2010] in order to prove the limiting law here because the functional $G \mapsto \int K(G)$ is not Hadamard differentiable.

Remark 4.3.4. Integrability conditions (4.3.4) and (4.3.5) over the distribution function can be reformulated by imposing directly conditions over the distribution function using the notion of regular variations and the Lemma page 280 in Feller [1971]. Regular variations transform the problem into conditions over the tails of the distribution functions. Suppose that there exists a constant $\rho_+ < -2$ and a function $L_+(x)$ such that:

$$1 - \mathbb{F}_T(x) = x^{\rho_+}L_+(x), \text{ with } \quad \frac{L_+(tx)}{L_+(t)} \xrightarrow[t\to\infty]{} 1, \forall x > 0. \tag{4.3.7}$$

Then, the integral $\int_y^\infty \sqrt{1-\mathbb{F}_T(x)}dx$ converges and there exists a function $M_+(y)$ such that $M_+(ty)/M_+(t) \to 1, \forall y$ and $\int_y^\infty \left[1-\mathbb{F}_T(x)\right]dx = y^{\rho_+ +1}M_+(y)$. For the neighborhood of $-\infty$, we make similar assumptions over $\mathbb{F}_T(x)$. Suppose that there exists a constant $\rho_- < -2$ and a function $L_-(x)$ such that:

$$\mathbb{F}_T(x) = x^{\rho_-}L_-(x), \text{ with } \quad \frac{L_-(-tx)}{L_-(t)} \xrightarrow[t\to-\infty]{} 1, \forall x < 0. \tag{4.3.8}$$

Then, the integral $\int_{-\infty}^{y}\sqrt{\mathbb{F}_T(x)}dx$ converges and there exists a function $M_-(y)$ such that $M_-(ty)/M_-(t)\to 1, \forall y$ and $\int_y^{\infty}\mathbb{F}_T(x)dx = y^{\rho_-+1}M_-(y)$.
These two assertions permit to conclude that condition (4.3.4) is verified since $\sqrt{\mathbb{F}_T(x)(1-\mathbb{F}_T(x))}\leq \sqrt{\mathbb{F}_T(x)}\mathbb{1}_{x\in(-\infty,0)}+\sqrt{1-\mathbb{F}_T(x)}\mathbb{1}_{x\in(0,\infty)}$. Moreover, condition (4.3.5) can also be check. Let's discuss what happens when y is at a neighborhood of either $+\infty$ or $-\infty$. For any $y>0$, one may write:

$$\begin{aligned}\int_y^{+\infty}[\mathbb{F}_T(\min(x,y))-\mathbb{F}_T(y)\mathbb{F}_T(x)]dx &= \mathbb{F}_T(y)\int_y^{+\infty}[1-\mathbb{F}_T(x)]dx\\ &= \mathbb{F}_T(y)y^{\rho_++1}M_+(y)\end{aligned}$$

which is integrable in a neighborhood of $+\infty$ with respect to y by (4.3.7). On the other hand, for any $y<0$, one may write

$$\begin{aligned}\int_{-\infty}^{y}[\mathbb{F}_T(\min(x,y))-\mathbb{F}_T(y)\mathbb{F}_T(x)]dx &= [1-\mathbb{F}_T(y)]\int_{-\infty}^{y}\mathbb{F}_T(x)dx\\ &= [1-\mathbb{F}_T(y)]\,y^{\rho_-+1}M_-(y)\end{aligned}$$

which is integrable in a neighborhood of $-\infty$ with respect to y by (4.3.8). Thus, condition (4.3.5) is ensured under assumptions (4.3.7,4.3.8).

We move on now to show the asymptotic normality of the estimator $\hat{\phi}$. Define the following matrices:

$$J_{\phi^*,\xi^*} = \begin{pmatrix} -\int\left[\frac{1}{(1-\lambda^*)^2}\mathbb{F}_T(y)-\frac{1}{(1-\lambda^*)^2}\mathbb{F}_1(y|\theta^*)\right]K'(\mathbb{F}_0(y|\phi^*))dy \\ \frac{\lambda^*}{1-\lambda^*}\int\nabla_\theta\mathbb{F}_1(y|\theta^*)K'(\mathbb{F}_0(y|\phi^*))^t dy \\ \nabla m(\alpha^*)\end{pmatrix}^t; \quad (4.3.9)$$

$$J_{\xi^*,\xi^*} = \int K(\mathbb{F}_0(y|\phi^*))K(\mathbb{F}_0(y|\phi^*))^t dy; \quad (4.3.10)$$

$$\tilde{\Sigma} = \left(J^t_{\phi^*,\xi^*}J_{\xi^*,\xi^*}J_{\phi^*,\xi^*}\right)^{-1};$$

$$H = \tilde{\Sigma}J^t_{\phi^*,\xi^*}J^{-1}_{\xi^*,\xi^*}; \quad (4.3.11)$$

$$P = J^{-1}_{\xi^*,\xi^*}-J^{-1}_{\xi^*,\xi^*}J_{\phi^*,\xi^*}\tilde{\Sigma}J^t_{\phi^*,\xi^*}J^{-1}_{\xi^*,\xi^*}. \quad (4.3.12)$$

We use the same notations considered at the beginning of this section for $\mathbb{F}_0(x|\phi), \hat{\mathbb{F}}_0(x|\phi), \xi(\phi)$ and $\xi_n(\phi)$.

Theorem 4.3.2. *Suppose that assumptions of Proposition 4.3.1 are fulfilled. Suppose also that*

1. $(\hat{\phi}, \xi_n(\hat{\phi}))$ *tends to* $(\phi^*, 0)$ *in probability;*
2. $\phi^* \in$ *int*(Φ)*;*
3. $\alpha \mapsto m(\alpha)$ *is of class* $\mathcal{C}^2$*;*
4. *there exists an integrable function* B_1 *such that* $\|\nabla_\theta\mathbb{F}_1(y|\theta)\| \leq B_1(y)$ *for* θ *in a neighborhood of* θ^**;*
5. *there exist integrable functions* $B_{2,1}$ *and* $B_{2,2}$ *such that* $\|\nabla_\theta\mathbb{F}_1(y|\theta)\nabla_\theta\mathbb{F}_1(y|\theta)^t\| \leq B_{2,2}(y)$ *and* $\|J_{\mathbb{F}_1(y|\theta)}\| \leq B_{2,1}(y)$ *for* θ *in a neighborhood of* θ^**;*

6. *the integral* $\int [\mathbb{F}_T(y) - \mathbb{F}_1(y)]dy$ *exists and is finite;*

7. *the matrices* J_{ξ^*,ξ^*} *and* $J^t_{\phi^*,\xi^*} J_{\xi^*,\xi^*} J_{\phi^*,\xi^*}$ *are invertible.*

Then,

$$\begin{pmatrix} \sqrt{n}\left(\hat{\phi} - \phi^*\right) \\ \sqrt{n}\xi_n(\hat{\phi}) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \begin{pmatrix} H \\ P \end{pmatrix} \Sigma \begin{pmatrix} H^t & P^t \end{pmatrix}\right),$$

where H, P *and* Σ *are given respectively by formulas (4.3.11), (4.3.12) and (4.3.6).*

The proof of this theorem is differed to Appendix 4.5.6. In assumption 1, we could only demand the consistency of $\hat{\phi}$, since the consistency of $\xi_n(\hat{\phi})$ can be deduced from it using the continuity of $\phi \mapsto \xi(\phi)$ and the uniform convergence of $\xi_n(.)$ towards $\xi(.)$, see Lemma 3.4.1. Assumptions 4-6 are used in the proof to ensure the differentiability up to second order with respect to ξ and ϕ of $H_n(\phi, \xi)$ for any n.

4.4 Simulation study

We perform several simulations and show how a prior information about the distribution of the semiparametric component P_0 can help us better estimate the set of parameters $(\lambda^*, \theta^*, \alpha^*)$ in regular examples, i.e. the components of the mixture can be clearly distinguished when we plot the probability density function. We also show how our approach permits to estimate even in difficult situations when the proportion of the parametric component is very low; such cases could *not* be estimated using existing methods. We show also the advantage of using L-moments constraints over moment constraints using the approach developed in the previous chapter.
In our experiments, the datasets were generated by the following mixtures:

- A two-component Weibull mixture;
- A two-component Weibull – Lognormal mixture;
- A two-component Gaussian – Two-sided Weibull mixture;

We have chosen a variety of values for the parameters especially the proportion. Programming tools are the same as in the case of the moment-type constraints. We only used the χ^2 divergence, because the optimization over ξ can be calculated without numerical methods, see Examples 4.3.1 and 3.4.1. Since the objective function $\phi \mapsto H_n(\phi, \xi_n(\phi))$ as a function of ϕ is not ensured to be strictly convex, we used 6 fixed initial points which we specify for each example separately. We then ran the Nelder-Mead algorithm and chose the vector of parameters for which the objective function has the lowest value. We applied a similar procedure on the algorithm of Bordes and Vandekerkhove [2010] in order to ensure a *fair* comparison.
All numerical integrations were calculated using function `integral` of package `pracma`. It was the only function that converged on all the calculus, see Section 1.7 for more details about other numerical integration functions.
We did not use any function error criterion here because the compared methods do not provide the same set of parameters. For example, the method of Bordes and Vandekerkhove [2010] estimates a mean value for the unknown component whereas our approach estimates a shape parameter. Other existing methods do not estimate any information about the parameters of the unknown component.

4.4.1 Data generated from a two-component Weibull mixture modeled by a semiparametric Weibull mixture

We consider a mixture of two Weibull components with scales $\sigma_1 = 0.5, \sigma_2 = 1$ and shapes $\nu_1 = 2, \nu_2 = 1$ in order to generate the dataset. In the semiparametric mixture model, the parametric component will be "the one to the right", i.e. the component whose true set of parameters is $(\nu_1 = 2, \sigma_1 = 0.5)$.
We impose on the unknown component three L-moments constraints; the second, the third and the fourth Weibull L-moments. They are given in Example 4.1.1. We thus have

$$m(\alpha = \nu) = \begin{pmatrix} -\lambda_2 = -\sigma\left(1 - 2^{-1/\nu}\right)\Gamma(1 + 1/\nu) \\ -\lambda_3 = -\lambda_2 \times \left(3 - 2\frac{1-3^{-1/\nu}}{1-2^{-1/\nu}}\right) \\ -\lambda_4 = -\lambda_2 \times \left(6 + \frac{5(1-4^{-1/\nu})-10*(1-3^{-1/\nu})}{1-2^{-1/\nu}}\right) \end{pmatrix}$$

and $K(t) = (t(t-1), t(t-1)(2t-1), t(t-1)(1+5(t-1)+5(t-1)^2))^t$. This mixture was not easily estimated by either our estimation procedure or the semiparametric methods from the literature. Our estimator, although has a higher variance, is still not biased in the same way estimates of other methods are. The L-moment constraints gave an estimator with less variance than the estimator based on moments constraints, but with slightly higher bias on the proportion.

Nb of observations	λ	sd(λ)	ν_1	sd(ν_1)	ν_2	sd(ν_2)
Mixture 1 : $n = 10^4$ $\lambda^* = 0.3$, $\nu_1^* = 2$, $\sigma_1^* = 0.5$(fixed), $\nu_2^* = 1$, $\sigma_2^* = 1$(fixed)						
Pearson's χ^2 3 moments	0.304	0.016	2.191	0.887	0.998	0.013
Pearson's χ^2 3 L-moments	0.348	0.062	1.828	0.648	0.984	0.021
Robin	0.604	0.029	1.256	0.037	—	—
Song EM-type	0.806	0.005	1.185	0.018	—	—
Song $\pi-$maximizing	0.624	0.007	1.312	0.013	—	—

Table 4.1: The mean value with the standard deviation of estimates in a 100-run experiment on a two-component Weibull mixture.

4.4.2 Data generated from a two-component Weibull-LogNormal mixture modeled by a semiparametric Weibull-LogNormal mixture

We consider a dataset generated from a mixture of a Weibull and a Lognormal distributions. The Weibull component has a scale $\sigma_1^* = 1$ and a shape $\nu_1^* \in \{1.5, 1, 0.4\}$ in order to illustrate several scenarios; a distribution whose pdf explodes to infinity at zero, a distribution whose pdf has finite value at zero and a distribution whose pdf goes back to zero at zero. The Lognormal component has a scale $\sigma_2^* = 0.5$ and a mean parameter $\mu^* = 3$. The Lognormal distribution has a heavy tail which is inherited in the mixture distribution.
In a first part, we perform a comparison of convergence speed between the method under moments constraints and the method under L-moments constraints as we increase the number of observations n. Details about the simulations under moments constraints can be found in paragraph 3.5.2. The Weibull component is considered as the unknown component during estimation, and impose three L-moments constraints. The first 4 L-moments of the Weibull distribution are given in Example 4.1.1.
In a second part, we perform an estimation of a semiparametric mixture model where the

Lognormal component is considered unknown and defined through 3 L-moments conditions; the second, the third and the fourth L-moment. The L-moments of the Lognormal distribution do not have a close formula and are calculated numerically using function `lmrln3` of package `lmom` written by Hosking.
Results in table (4.2) show that L-moments are more informative and we need less data in order to get good estimates in comparison to moments constraints. In order to calculate the estimate $\hat{\phi}$, we considered 6 initial points; namely the set

$$\phi^{(0)} \in \{(0.8, 2, 1), (0.5, 2, 1), (0.8, 1, 1), (0.7, 3, 1.5), (0.7, 2, 2), (0.5, 4, 2), (0.5, 1.5, 2)\}\,.$$

The vector $\hat{\phi}$ was taken as the one which corresponds to the lowest value among the infima produced by the optimization algorithm.
In table (4.3) the Lognormal component is the unknown component during estimation. Initialization of the optimization algorithm, for example in mixture 2, was taken from the set $\{(0.1, 0.5, 1), (0.15, 0.5, 0.7), (0.05, 1.5, 2.5), (0.1, 1, 3)\}$.
It is clear that the moments constraints gave better results than L-moments constraints in mixture 1 for the estimation of the scale of the Weibull component. For the second mixture, both types of constraints give similar results. The two methods have the same bias in the estimation of the scale of Weibull component; the moments constraints produced a positive bias whereas the L-moments constraints produced a negative bias. The L-moments produced a smaller variance. In the third mixture, the L-moments constraints gave clear better results. The last mixture is the most difficult one in the sense that the proportion of the parametric component is very low.

nb of observations	Estimation method	λ	sd(λ)	μ	sd(μ)	ν	sd(ν)
True Parameters : $\lambda^* = 0.7$, $\mu^* = 3$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 1.5$, $\sigma_1^* = 1$(fixed)							
$n = 10^2$	Pearson's χ^2 L-moments	0.685	0.069	2.798	0.413	0.436	0.074
	Pearson's χ^2 Moments	0.384	0.117	2.654	0.153	0.488	0.018
$n = 10^3$	Pearson's χ^2 L-moments	0.677	0.017	3.014	0.028	0.726	0.272
	Pearson's χ^2 Moments	0.518	0.068	2.806	0.099	0.473	0.014
$n = 10^4$	Pearson's χ^2 L-moments	0.697	0.009	3.003	0.010	1.343	0.185
	Pearson's χ^2 Moments	0.605	0.044	2.903	0.069	0.531	0.326

Table 4.2: The mean value with the standard deviation of estimates in a 100-run experiment on a two-component Weibull-log normal mixture. The parametric component is the log-normal with unknown mean parameter μ. The semiparametric component is the Weibull component which is defined by its first three L-moments (moments resp.) with unknown shape ν.

4.4.3 Data generated from a two-sided Weibull Gaussian mixture modeled by a semiparametric two-sided Weibull Gaussian mixture

We have already presented this model in paragraph 3.5.3. The 2nd, 3rd and 4th L-moments of the two-sided Weibull distribution are given by:

$$\begin{aligned}
\lambda_2 &= \left[1 - \frac{1}{2^{1+1/\nu}}\right] \sigma_2 \Gamma\left(1 + \frac{1}{\nu}\right);\\
\lambda_3 &= 0;\\
\lambda_4 &= \left[1 - \frac{6}{2^{1+1/\nu}} + \frac{15}{2 \times 3^{1+1/\nu}} - \frac{5}{2 \times 4^{1+1/\nu}}\right] \sigma_2 \Gamma\left(1 + \frac{1}{\nu}\right).
\end{aligned}$$

Nb of observations	λ	sd(λ)	ν	sd(ν)	μ	sd(μ)
Mixture 1 : $n = 10^3$, $\lambda^* = 0.3$, $\nu^* = 1.5$, $\sigma_1^* = 1$(fixed), $\mu^* = 3$, $\sigma_2^* = 0.5$(fixed)						
Pearson's χ^2 L-moments	0.313	0.019	1.027	0.541	2.992	0.050
Pearson's χ^2 Moments	0.308	0.017	1.484	0.624	3.002	0.026
Mixture 2 : $n = 10^4$, $\lambda^* = 0.1$, $\nu^* = 1$, $\sigma_1^* = 1$(fixed), $\mu^* = 3$, $\sigma_2^* = 0.5$(fixed)						
Pearson's χ^2 L-moments	0.104	0.006	0.795	0.379	2.994	0.015
Pearson's χ^2 Moments	0.103	0.006	1.284	0.677	3.001	0.007
Mixture 3 : $n = 5 \times 10^4$, $\lambda^* = 0.05$, $\nu^* = 0.4$, $\sigma_1^* = 1$(fixed), $\mu^* = 3$, $\sigma_2^* = 0.5$(fixed)						
Pearson's χ^2 L-Moments	0.049	0.002	0.448	0.129	3.000	0.006
Pearson's χ^2 Moments	0.049	0.002	0.629	0.438	3.001	0.004

Table 4.3: The mean value with the standard deviation of estimates in a 100-run experiment on a two-component Weibull-log normal mixture. The parametric component is the Weibull with unknown shape ν. The semiparametric component is the lognormal component which is defined by its first three L-moments (moments resp.) with unknown mean parameter μ.

Results are presented in table (4.4). The L-moments constraints produce clear better results than the moments constraints in all the mixtures. The estimation based on L-moments constraints produced clear lower variance. Besides, and once again, the L-moments constraints seem to be more informative and we need less number of observations than moments constraints in order to produce good estimates.
In this example we presented a challenge to our estimation method by simulating mixtures with very low proportion of the parametric part; mixture 3 with $\lambda^* = 0.05$ and mixture 4 with $\lambda^* = 0.01$. Using signal-noise terms, in mixture 4, only one percent of the data comes from the signal whereas 99% of the data is pure noise. The location of the signal is then estimated around zero with standard deviation of 0.3 with the L-moments constraints. It is not well localized however using moments constraints with 10^5 observations, and we need at least 10^8 observations to reach a similar precision to the result obtained with L-moments constraints. It is still important to notice that using moments or L-moments constraints, we were able to confirm the existence of a signal component (the parametric component).
In what concerns the initialization of the algorithm under L-moments constraints, we used:

$$\begin{aligned}
\text{Mix } 1 &: \{(0.8,1,1),(0.5,-1,2.5),(0.8,0.5,2),(0.7,0,3),(0.7,1,4),(0.5,2,3.5)\}\\
\text{Mix } 2 &: \{(0.2,1,1),(0.5,-1,2.5),(0.2,0.5,2),(0.3,0,3),(0.3,1,4)\}\\
\text{Mix } 3 &: \{(0.1,1,1),(0.05,-1,2.5),(0.03,0.5,2),(0.01,0,1.5),(0.005,1,0.7)\}\\
\text{Mix } 4 &: \{(0.1,1,1),(0.005,1,0.7)\}
\end{aligned}$$

For the last mixture, we have found no changes in using more initial points than the two given points. Besides, execution time was very long (about 5 samples per day), so we preferred to use only two starting points.

Estimation method	λ	sd(λ)	μ	sd(μ)	ν	sd(ν)
Mixture 1 : $n = 100$, $\lambda^* = 0.7$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 3$, $\sigma_1^* = 1.5$(fixed)						
Pearson's χ^2 – L-Moments	0.758	0.067	-2.28×10^{-3}	0.098	3.040	0.639
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.764	0.067	-0.012	0.342	2.893	0.731
Mixture 2 : $n = 100$, $\lambda^* = 0.3$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 3$, $\sigma_1^* = 1.5$(fixed)						
Pearson's χ^2 – L-Moments	0.364	0.082	-0.016	0.246	3.058	0.418
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.407	0.077	0.012	0.575	2.925	0.454
Mixture 3 : $n = 5000$, $\lambda^* = 0.05$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 1.5$, $\sigma_1^* = 2$(fixed)						
Pearson's χ^2 – L-Moments	0.050	0.013	0.026	0.365	1.496	0.020
Pearson's χ^2 under $\mathcal{M}_{2:4}$ 0.066	0.013	-0.036	0.857	1.493	0.008	
Mixture 4 : $n = 10^5$, $\lambda^* = 0.01$, $\mu^* = 0$, $\sigma_2^* = 0.5$(fixed), $\nu^* = 1.5$, $\sigma_1^* = 2$(fixed)						
Pearson's χ^2 – L-Moments	0.011	0.003	0.023	0.377	1.500	0.005
Pearson's χ^2 under $\mathcal{M}_{2:4}$	0.025	0.010	- 0.047	1.356	1.495	0.006

Table 4.4: The mean value with the standard deviation of estimates in a 100-run experiment on a two-component two-sided Weibull–Gaussian mixture under L-moment constraints.

4.4.4 Conclusions

In this chapter, we introduced another structure for semiparametric mixture models with unknown component by imposing L-moments constraints on it. The method was proved to be consistent and asymptotic normal under standard assumptions. The estimation method under L-moments constraints presented several advantages in comparison to the estimation method under moments constraints. We were able to estimate over the whole parameter space and no need to check if the optimized function $\xi \mapsto H(\phi, \xi)$ is strictly concave for every ϕ. Although the estimation method under L-moments constraints need numerical integrations (which is not the case of moments-type constraints procedure), the resulting estimator seems to have lower variance. Moreover, L-moments are demonstrated through simulations to be more informative than moments constraints, and we need less number of observations in order to obtain good estimates.

4.5 Appendix: Proofs

4.5.1 Proof of Proposition 4.2.1

Proof. Denote M^1 the set of all probability measures. Based on equation (4.2.2), we have:

$$\begin{aligned} P_0 &= \frac{1}{1-\lambda} P_T - \frac{\lambda}{1-\lambda} P_1(.|\theta) \\ \tilde{P}_0 &= \frac{1}{1-\tilde{\lambda}} P_T - \frac{\tilde{\lambda}}{1-\tilde{\lambda}} P_1(.|\tilde{\theta}) \end{aligned}$$

Define the following function:

$$G : \mathbb{R}^{d-s} \times M^+ \to \mathrm{Im}(G) \subset M^1 : (\lambda, \theta, P_0) \mapsto \lambda P_1(.|\theta) + (1-\lambda) P_0.$$

where

$$M^+ = \{P_0 \in M^1 \text{ s.t. } \mathbf{F}_0^{-1} \in \mathcal{M}\}.$$

Identifiability is now equivalent to the fact that function G is one-to-one. This means that for a given mixture distribution $P_T \in \mathrm{Im}(G)$, we need that there exists a unique triplet (λ, θ, P_0) such that

$$P_T = \lambda P_1(.|\theta) + (1-\lambda) P_0$$

In other words:

$$P_0 = \frac{1}{1-\lambda} P_T - \frac{\lambda}{1-\lambda} P_1(.|\theta)$$

The equality of measures imply the equality of the quantiles. Thus, we may write:

$$\int_0^1 K(u) \mathbf{F}_0^{-1}(du) = m(\alpha) = \int_0^1 K(u) \left(\frac{1}{1-\lambda} \mathbf{F}_T - \frac{\lambda}{1-\lambda} \mathbf{F}_1(.|\theta) \right)^{-1} (du). \tag{4.5.1}$$

The assumption of the present proposition imposes the existence of unique solution $(\lambda^*, \theta^*, \alpha^*)$ to the previous nonlinear system of equations. Let's go back to function G. For a given mixture distribution $P_T \in \mathrm{Im}(G)$, take $\lambda = \lambda^*, \theta = \theta^*$ to be the solution to the nonlinear system (4.5.1), and define P_0^* by:

$$P_0^* = \frac{1}{1-\lambda^*} P_T - \frac{\lambda^*}{1-\lambda^*} P_1(.|\theta^*).$$

Notice that $P_0^* \in \mathcal{M}_{\alpha^*}$. Suppose that P_T can be written in two manners. In other words, suppose that there exists another triplet $(\tilde{\lambda}, \tilde{\theta}, \tilde{P}_0)$ with $\tilde{P}_0 \in \mathcal{M}_{\tilde{\alpha}}$ such that:

$$P_T = \tilde{\lambda} P_1(.|\tilde{\theta}) + (1 - \tilde{\lambda})\tilde{P}_0.$$

We then have:

$$\tilde{P}_0 = \frac{1}{1-\tilde{\lambda}} P_T - \frac{\tilde{\lambda}}{1-\tilde{\lambda}} P_1(.|\tilde{\theta}),$$

and consequently,

$$m(\tilde{\alpha}) = \int_0^1 K(u) \left(\frac{1}{1-\tilde{\lambda}} \mathbf{F}_T - \frac{\tilde{\lambda}}{1-\tilde{\lambda}} \mathbf{F}_1(.|\tilde{\theta}) \right)^{-1} (du).$$

Thus, $(\tilde{\lambda}, \tilde{\theta}, \tilde{\alpha})$ is a second solution to the system (4.5.1). Nevertheless, the system of equations (4.5.1) has a unique solution by assumption of the present proposition. Hence, a contradiction is reached and the triplet $(\lambda^*, \theta^*, P_0^*)$ is unique. We conclude that function G is one-to-one and the semiparametric mixture model subject to L-moments constraints is identifiable. □

4.5.2 Proof of Proposition 4.2.2

Proof. Let $\mathbf{F}_0^{-1}$ be some quantile measure which belongs to the intersection $\mathcal{N}^{-1} \cap \mathcal{M}$. Since $\mathbf{F}_0^{-1}$ belongs to $\mathcal{N}^{-1}$, there exists a couple $(\lambda, \theta) \in \Phi^+$ such that:

$$\mathbf{F}_0^{-1} = \left(\frac{1}{1-\lambda} \mathbf{F}_T - \frac{\lambda}{1-\lambda} \mathbf{F}_1(.|\theta) \right)^{-1}. \tag{4.5.2}$$

This couple is unique by virtue of assumptions 3 and 4. Indeed, let (λ, θ) and $(\tilde{\lambda}, \tilde{\theta})$ be two couples such that:

$$\left(\frac{1}{1-\lambda} \mathbf{F}_T - \frac{\lambda}{1-\lambda} \mathbf{F}_1(.|\theta) \right)^{-1} = \left(\frac{1}{1-\tilde{\lambda}} \mathbf{F}_T - \frac{\tilde{\lambda}}{1-\tilde{\lambda}} \mathbf{F}_1(.|\tilde{\theta}) \right)^{-1}$$

This entails that:

$$\frac{1}{1-\lambda} \mathbb{F}_T(x) - \frac{\lambda}{1-\lambda} \mathbb{F}_1(x|\theta) = \frac{1}{1-\tilde{\lambda}} \mathbb{F}_T(x) - \frac{\tilde{\lambda}}{1-\tilde{\lambda}} \mathbb{F}_1(x|\tilde{\theta}). \tag{4.5.3}$$

By derivation of both sides, we get an identity in the densities:

$$\frac{1}{1-\lambda} - \frac{\lambda}{1-\lambda} \frac{p_1(x|\theta)}{p_T(x)} = \frac{1}{1-\tilde{\lambda}} - \frac{\tilde{\lambda}}{1-\tilde{\lambda}} \frac{p_1(x|\tilde{\theta})}{p_T(x)}.$$

Taking the limit as x tends to ∞ results in:

$$\frac{1 - c\lambda}{1-\lambda} = \frac{1 - c\tilde{\lambda}}{1-\tilde{\lambda}}.$$

Note that function $z \mapsto (1 - cz)/(1 - z)$ is strictly monotone as long as $c \neq 1$. Hence, it is a one-to-one map. Thus $\lambda = \tilde{\lambda}$. Inserting this result in equation (4.5.3) entails that:

$$\mathbb{F}_1(.|\theta) = \mathbb{F}_1(.|\tilde{\theta}).$$

Using the identifiability of P_1 (assumption 4), we get $\theta = \tilde{\theta}$ which proves the existence of a unique couple (λ, θ) in (4.5.2).
On the other hand, since $\mathbf{F}_0^{-1}$ belongs to $\mathcal{M}$, there exists a unique α such that $\mathbf{F}_0^{-1} \in \mathcal{M}_\alpha$. Uniqueness comes from the fact that the function $\alpha \mapsto m(\alpha)$ is one-to-one (assumption 2). Thus, $\mathbf{F}_0^{-1}$ verifies the constraints

$$\int_0^1 K(u)\mathbf{F}_0^{-1}(du) = m(\alpha).$$

Combining this with (4.5.2), we get:

$$\int_0^1 K(u)\left(\frac{1}{1-\lambda}\mathbf{F}_T - \frac{\lambda}{1-\lambda}\mathbf{F}_1(.|\theta)\right)^{-1}(du) = m(\alpha). \tag{4.5.4}$$

This is a non linear system of equations with ℓ equations. Now, let $\mathbf{F}_0^{-1}$ and $\tilde{\mathbf{F}}_0^{-1}$ be two elements in $\mathcal{N}^{-1} \cap \mathcal{M}$, then there exist two couples (λ, θ) and $(\tilde{\lambda}, \tilde{\theta})$ with $\lambda \neq \tilde{\lambda}$ or $\theta \neq \tilde{\theta}$ such that $\mathbf{F}_0^{-1}$ and $\tilde{\mathbf{F}}_0^{-1}$ can be written in the form of (4.5.2) with respectively (λ, θ) and $(\tilde{\lambda}, \tilde{\theta})$. Since $\mathbf{F}_0^{-1} \in \mathcal{M}$, there exists α such that $\mathbf{F}_0^{-1} \in \mathcal{M}_\alpha$. Similarly, there exists $\tilde{\alpha}$ possibly different from α such that $\tilde{\mathbf{F}}_0^{-1} \in \mathcal{M}_{\tilde{\alpha}}$. Now, $(\lambda, \theta, \alpha)$ and $(\tilde{\lambda}, \tilde{\theta}, \tilde{\alpha})$ are two solutions to the system of equations (4.5.4) which contradicts with assumption 1 of the present proposition.
We may now conclude that, if a quantile measure $\mathbf{F}_0^{-1}$ belongs to the intersection $\mathcal{N}^{-1} \cap \mathcal{M}$, then it has the representation (4.5.2) for a unique couple (λ, θ) and there exists a unique α such that the triplet $(\lambda, \theta, \alpha)$ is a solution to the non linear system (4.5.4). Conversely, if there exists a triplet $(\lambda, \theta, \alpha)$ which solves the non linear system (4.5.4), then the quantile measure $\mathbf{F}_0^{-1}$ defined by $\mathbf{F}_0^{-1} = \left(\frac{1}{1-\lambda}\mathbf{F}_T - \frac{\lambda}{1-\lambda}\mathbf{F}_1(.|\theta)\right)^{-1}$ belongs to the intersection $\mathcal{N}^{-1} \cap \mathcal{M}$. This is because on the one hand, it clearly belongs to $\mathcal{N}^{-1}$ by its definition and on the other hand, it belongs to $\mathcal{M}_\alpha$ since it verifies the constraints and thus belongs to $\mathcal{M}$.
It is now reasonable to conclude that under assumptions 2-4, the intersection $\mathcal{N}^{-1} \cap \mathcal{M}$ includes a *unique* quantile measure $\mathbf{F}_0^{-1}$ if and only if the set of ℓ non linear equations (3.7.4) has a unique solution $(\lambda, \theta, \alpha)$. □

4.5.3 Proof of Lemma 4.3.1

Proof. The same arguments hold for both functions $\xi(\phi)$ and $\xi_n(\phi)$. We therefore, proceed with $\xi(\phi)$. Function $\xi \mapsto H(\phi, \xi)$ is strictly concave since[1] it is $\mathcal{C}^2$ and have the following Hessian matrix:

$$J_{H(\phi,.)} = -\int K\left(\mathbb{F}_0(y,|\phi)\right) K\left(\mathbb{F}_0(y,|\phi)\right)^t \psi''\left(\xi^t K\left(\mathbb{F}_0(y,|\phi)\right)\right) dy.$$

Since ψ is strictly convex, then $\psi''(z) > 0$ for any z. Thus the matrix $J_{H(\phi,.)}$ is definite negative and $\xi \mapsto H(\phi, \xi)$ is strictly concave. By the implicit function theorem, function $\phi \mapsto \xi(\phi)$ is uniquely defined and $\mathcal{C}^1$ over $\text{int}(\Phi)$. Notice here that even if $\frac{1}{1-\lambda}\mathbb{F}_T(y) - \frac{\lambda}{1-\lambda}\mathbb{F}_1(y|\theta)$ is negative, the matrix $J_{H(\phi,.)}$ can still be definite negative unlike the case of moment constraints.
The second part of the proposition is a direct consequence from the continuity of function $\phi \mapsto \xi(\phi)$. □

[1] One can prove the strict concavity simply by calculating $H(\phi, u\xi_1 + (1-u)\xi_2)$.

4.5.4 Proof of Theorem 4.3.1

Proof. We will use Theorem 3.4.1. We start with assumption A2. We prove, first, that the supremum over ξ can only be calculated over a compact subset of $\mathbb{R}^l$. This is a direct result from Lemma 4.3.1. One can redefine the estimator by maximizing over ξ on the subset $\Xi = \text{Im}(\xi(.)) \subset \mathbb{R}^l$ independently of ϕ. We thus have:

$$\begin{aligned} D_\varphi(\mathcal{M}_\alpha, \mathbb{F}_0(.|\phi)) &= \sup_{\xi \in \Xi} H(\phi, \xi) \\ \phi^* &= \arg\inf_{\phi} \sup_{\xi \in \Xi} H(\phi, \xi). \end{aligned}$$

We redefine now the estimation procedure (4.2.13) as follows:

$$\hat{\phi} = \arg\inf_{\alpha,\theta,\lambda} \sup_{\xi \in \Xi} \xi^t m(\alpha) - \int \psi \left[\xi^t K \left(\frac{1}{1-\lambda} \mathbb{F}_T(y) - \frac{\lambda}{1-\lambda} \mathbb{F}_1(y|\theta) \right) \right] dy$$

Using the mean value theorem, there exists $\eta(y) \in (0,1)$ such that[2]:

$$\begin{aligned} \psi\left(\xi^t K\left(\mathbb{F}_0(y|\phi)\right)\right) - \psi\left(\xi^t K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right) &= \xi^t \left(K\left(\mathbb{F}_0(y|\phi)\right) - K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right) \\ &\times \psi'\left(\eta(y)\xi^t K\left(\mathbb{F}_0(y|\phi)\right) + (1-\eta(y))\xi^t K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right) \end{aligned} \quad (4.5.5)$$

An exact formula of function $\eta(y)$ will not be needed. We will only use the fact that its image is included in $(0,1)$. By the central limit theorem, one can write:

$$\sqrt{n} \frac{\mathbb{F}_n(y) - \mathbb{F}_T(y)}{\sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}} \to \mathcal{N}(0,1).$$

Since $\hat{\mathbb{F}}_0(y|\phi) - \mathbb{F}_0(y|\phi) = \mathbb{F}_n(y) - \mathbb{F}_T(y)$, we write

$$\sqrt{n} \frac{\hat{\mathbb{F}}_0(y|\phi) - \mathbb{F}_0(y|\phi)}{\sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}} \to \mathcal{N}(0,1),$$

which entails by the delta method that:

$$\sqrt{n} \frac{K\left(\hat{\mathbb{F}}_0(y|\phi)\right) - K\left(\mathbb{F}_0(y|\phi)\right)}{\sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}} \to \mathcal{N}\left(0, \nabla K\left(\mathbb{F}_0(y|\phi)\right) \nabla K\left(\mathbb{F}_0(y|\phi)\right)^t\right). \quad (4.5.6)$$

Since function K is a vector of polynomials, its gradient is a matrix of polynomials. Besides, the distribution function $\mathbb{F}_0(y|\phi)$ takes its values in $[0,1]$, thus the variance of the limiting law in (4.5.6) is of order $\frac{1}{n}$ independently of y and ϕ. We may now write:

$$\frac{K\left(\hat{\mathbb{F}}_0(y|\phi)\right) - K\left(\mathbb{F}_0(y|\phi)\right)}{\sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}} = o_P(1) \quad (4.5.7)$$

[2] In the case of the Chi square, $\lambda(y) = \frac{1}{2}$

Going back to equation (4.3.1), we use equations (4.5.5) and (4.5.7) to write:

$$\begin{aligned} H(\phi,\xi) - H_n(\phi,\xi) &= \int \xi^t \left(K\left(\mathbb{F}_0(y|\phi)\right) - K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right) \psi'\left[\eta(y)\xi^t K\left(\mathbb{F}_0(y|\phi)\right)\right. \\ &\quad \left. +(1-\eta(y))\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right] dy \\ &= \int \sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}\xi^t \frac{\left(K\left(\mathbb{F}_0(y|\phi)\right) - K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right)}{\sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}} \\ &\quad \times\psi'\left(\eta(y)\xi^t K\left(\mathbb{F}_0(y|\phi)\right) + (1-\eta(y))\xi^t K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right) dy \\ &= \xi^t o_p(1) \int \sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}\psi'\left[\eta(y)\xi^t K\left(\mathbb{F}_0(y|\phi)\right)\right. \\ &\quad \left. +(1-\eta(y))\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right] dy. \end{aligned}$$

The finale line can also be justified by the Chebyshev's inequality, see Remark 4.5.1, or even using the calculus in the proof of Proposition 4.3.1 below.
It suffices now to prove that the integral in the previous display is finite. Here, ξ (resp. ϕ) is inside the compact set Ξ (resp. Φ), and functions $\eta(y), \mathbb{F}_0(y|\phi)$ and $\hat{\mathbb{F}}_0(y|\phi)$ all take values inside the compact interval $[0,1]$. Thus, continuity of ψ' suffices to conclude that there exists a constant M independent of y, ϕ and ξ such that:

$$\left|\psi'\left(\eta(y)\xi^t K\left(\mathbb{F}_0(y|\phi)\right) + (1-\eta(y))\xi^t K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right)\right| \leq M. \tag{4.5.8}$$

This entails using assumption C6 that:

$$\begin{aligned} \int \sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}\left|\psi'\left(\eta(y)\xi^t K\left(\mathbb{F}_0(y|\phi)\right) + (1-\eta(y))\xi^t K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right)\right| dy &\leq \\ M\int \sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}dy & \\ < +\infty. & \end{aligned}$$

Finally, the integral is finite and the compactness of Ξ implies that $\|\xi\|$ is bounded. Therefore, we have:

$$H(\phi,\xi) - H_n(\phi,\xi) = o_P(1),$$

independently of ξ and ϕ. We may deduce now that:

$$\sup_{\phi,\xi} |H(\phi,\xi) - H_n(\phi,\xi)| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

This proves assumption A2.
Assumption A3 is immediately verified since function $\xi \mapsto H(\phi,\xi)$ is strictly concave. Assumption A4 is what we have assumed in assumption C3. Finally, continuity assumption A5 is a direct result from assumptions C4 and C5 using Lebesgue's continuity theorem. All assumptions of Theorem 3.4.1 are fulfilled and the consistency of $\hat{\phi}$ follows as a consequence. □

Remark 4.5.1. We can prove assumption A2 in the previous proof without the use of the "small o" notation. We first have:

$$\frac{K\left(\hat{\mathbb{F}}_0(y|\phi)\right) - K\left(\mathbb{F}_0(y|\phi)\right)}{\sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}} \xrightarrow{\mathbb{P}} 0.$$

This is translated into the following limit:

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left| \frac{K\left(\hat{\mathbb{F}}_0(y|\phi)\right) - K\left(\mathbb{F}_0(y|\phi)\right)}{\sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}} \right| < \varepsilon \right) \xrightarrow[n\to\infty]{} 1$$

Thus, there exists a sequence of positive numbers $(a_n)_n$ independent of[3] y which goes to zero at infinity such that:

$$\mathbb{P}\left(\left| \frac{K\left(\hat{\mathbb{F}}_0(y|\phi)\right) - K\left(\mathbb{F}_0(y|\phi)\right)}{\sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}} \right| < \frac{\varepsilon}{\tilde{M}} \right) \geq 1 - a_n$$

where $\tilde{M} = M \sup_{\Xi} \|\xi\| \int \mathbb{F}_T(y)(1-\mathbb{F}_T(y))dy$ and M is defined through inequality (4.5.8). On the other hand, the event:

$$\left\| \frac{K\left(\hat{\mathbb{F}}_0(y|\phi)\right) - K\left(\mathbb{F}_0(y|\phi)\right)}{\sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}} \right\| < \frac{\varepsilon}{\tilde{M}}$$

implies the event:

$$\begin{aligned}
&\int \sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}\|\xi\| \left\| \frac{\left(K\left(\mathbb{F}_0(y|\phi)\right) - K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right)}{\sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}} \right\| \psi'\left(\eta(y)\xi^t K\left(\mathbb{F}_0(y|\phi)\right)\right. \\
&\left. +(1-\eta(y))\xi^t K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right) dy \\
&< \frac{\varepsilon}{\tilde{M}} \int \sqrt{\mathbb{F}_T(y)(1-\mathbb{F}_T(y))}\|\xi\|\psi'\left(\eta(y)\xi^t K\left(\mathbb{F}_0(y|\phi)\right) + (1-\eta(y))\xi^t K\left(\hat{\mathbb{F}}_0(y|\phi)\right)\right) dy \\
&< \varepsilon.
\end{aligned}$$

This entails that

$$|H(\phi,\xi) - H_n(\phi,\xi)| < \varepsilon.$$

The final line does not depend on (ϕ,ξ), and we may deduce that:

$$\mathbb{P}\left(\sup_{\phi,\xi} |H(\phi,\xi) - H_n(\phi,\xi)| < \varepsilon \right) \geq 1 - a_n.$$

4.5.5 Proof of Proposition 4.3.1

Proof. We would like to calculate the difference $\int K(\hat{\mathbb{F}}_0(y|\phi))dy - \int K(\mathbb{F}_0(y|\phi))dy$ as a functional of the difference $\hat{\mathbb{F}}_0(y|\phi) - \mathbb{F}_0(y|\phi)$. For two reals a and b, we have:

$$K_r(a) - K_r(b) = \sum_{k=0}^{r-1} \frac{c_{r,k}}{k+1}\left(a^{k+1} - b^{k+1}\right),$$

where $c_{r,k} = (-1)^{r-k-1}\binom{r-1}{k}\binom{r+k-1}{k}$. Using the identity $a^{k+1} - b^{k+1} = (a-b)\sum_{j=0}^{k} a^j b^{k-j}$, we can write:

$$K_r(a) - K_r(b) = (a-b)\sum_{k=0}^{r-1}\sum_{j=0}^{k} \frac{c_{r,k}}{k+1} a^j b^{k-j}.$$

[3] This is possible using Chebyshev's inequality and using the fact that $K\left(\mathbb{F}_0(y|\phi)\right)$ can be bounded independently of y and ϕ.

Applying this formula on $a = \hat{\mathbb{F}}_0(y|\phi)$ and $b = \mathbb{F}_0(y|\phi)$ yields

$$\begin{aligned} K_r\left(\hat{\mathbb{F}}_0(y|\phi)\right) - K_r\left(\mathbb{F}_0(y|\phi)\right) &= \left(\hat{\mathbb{F}}_0(y|\phi) - \mathbb{F}_0(y|\phi)\right) \sum_{k=0}^{r-1}\sum_{j=0}^{k} \frac{c_{r,k}}{k+1} \hat{\mathbb{F}}_0(y|\phi)^j \mathbb{F}_0(y|\phi)^{k-j} \\ &= \frac{1}{1-\lambda}\left(\mathbb{F}_n(y) - \mathbb{F}_T(y)\right) \sum_{k=0}^{r-1}\sum_{j=0}^{k} \frac{c_{r,k}}{k+1} \hat{\mathbb{F}}_0(y|\phi)^j \mathbb{F}_0(y|\phi)^{k-j}. \end{aligned} \tag{4.5.9}$$

We will show that the sum term can be rewritten using only $\mathbb{F}_0(y|\phi)$. By the Kolmogorov-Smirnov theorem, we have:

$$\sup_y \left|\hat{\mathbb{F}}_0(y|\phi) - \mathbb{F}_0(y|\phi)\right| = \sup_y \left|\mathbb{F}_n(y) - \mathbb{F}_T(y)\right| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

This permits us to simply write that

$$\hat{\mathbb{F}}_0(y|\phi) = \mathbb{F}_0(y|\phi) + O_P\left(\frac{1}{\sqrt{n}}\right),$$

with $O_P\left(\frac{1}{\sqrt{n}}\right)$ tends to zero in probability as n goes to infinity independently of y. Thus formula (4.5.9) can be rewritten as:

$$\begin{aligned} K_r\left(\hat{\mathbb{F}}_0(y|\phi)\right) - K_r\left(\mathbb{F}_0(y|\phi)\right) &= \frac{1}{1-\lambda}\left(\mathbb{F}_n(y) - \mathbb{F}_T(y)\right) \sum_{k=0}^{r-1}\sum_{j=0}^{k} \frac{c_{r,k}}{k+1} \left(\mathbb{F}_0(y|\phi)^j + O_P\left(\frac{1}{\sqrt{n}}\right)\right) \\ &\quad \times \mathbb{F}_0(y|\phi)^{k-j} \\ &= \frac{1}{1-\lambda}\left(\mathbb{F}_n(y) - \mathbb{F}_T(y)\right) \sum_{k=0}^{r-1} c_{r,k} \mathbb{F}_0(y|\phi)^k \\ &\quad + O_P\left(\frac{1}{\sqrt{n}}\right) \frac{1}{1-\lambda}\left(\mathbb{F}_n(y) - \mathbb{F}_T(y)\right) \sum_{k=0}^{r-1}\sum_{j=0}^{k} \frac{c_{r,k}}{k+1} \mathbb{F}_0(y|\phi)^{k-j}. \end{aligned}$$

Integrating the two sides of the previous equation and multiplying by $\sqrt{n}$ gives:

$$\begin{aligned} \sqrt{n}\int \left[K_r\left(\hat{\mathbb{F}}_0(y|\phi)\right) - K_r\left(\mathbb{F}_0(y|\phi)\right)\right] dy &= \frac{1}{1-\lambda}\int \sqrt{n}\left(\mathbb{F}_n(y) - \mathbb{F}_T(y)\right) \sum_{k=0}^{r-1} c_{r,k} \mathbb{F}_0(y|\phi)^k dy \\ + O_P\left(\frac{1}{\sqrt{n}}\right) \frac{1}{1-\lambda}\int \sqrt{n}\left(\mathbb{F}_n(y) - \mathbb{F}_T(y)\right) &\sum_{k=0}^{r-1}\sum_{j=0}^{k} \frac{c_{r,k}}{k+1} \mathbb{F}_0(y|\phi)^{k-j} dy. \end{aligned} \tag{4.5.10}$$

The first integral in the right hand side is the part which will produce the Gaussian distribution of the limit law using the CLT. It remains to prove that the second integral in the right hand side tends to zero in probability. Using the law of iterated logarithm, we can write:

$$\limsup_{n\to\infty} \sqrt{\frac{n}{\log\log n}} \frac{\mathbb{F}_n(y) - \mathbb{F}_T(y)}{\sqrt{\mathbb{F}_T(y)\left(1 - \mathbb{F}_T(y)\right)}} = \sqrt{2}. \tag{4.5.11}$$

We now may write the integral in the second term as follows:

$$O_P\left(\frac{1}{\sqrt{n}}\right)\int\sqrt{n}\left(\mathbb{F}_n(y)-\mathbb{F}_T(y)\right)\sum_{k=0}^{r-1}\sum_{j=0}^{k}\frac{c_{r,k}}{k+1}\mathbb{F}_0(y|\phi)^{k-j}dy =$$

$$O_P\left(\sqrt{\frac{\log\log n}{n}}\right)\int\sqrt{\frac{n}{\log\log n}}\frac{\mathbb{F}_n(y)-\mathbb{F}_T(y)}{\sqrt{\mathbb{F}_T(y)\left(1-\mathbb{F}_T(y)\right)}}\sqrt{\mathbb{F}_T(y)\left(1-\mathbb{F}_T(y)\right)}\sum_{k=0}^{r-1}\sum_{j=0}^{k}\frac{c_{r,k}}{k+1}\mathbb{F}_0(y|\phi)^{k-j}dy.$$

The sum term inside the integral is bounded uniformly on y. Combine this with the limit in (4.5.11), we may deduce that for n sufficiently large, there exists a constant M such that:

$$\begin{aligned}\int\sqrt{\frac{n}{\log\log n}}\frac{|\mathbb{F}_n(y)-\mathbb{F}_T(y)|}{\sqrt{\mathbb{F}_T(y)\left(1-\mathbb{F}_T(y)\right)}}\sqrt{\mathbb{F}_T(y)\left(1-\mathbb{F}_T(y)\right)}\sum_{k=0}^{r-1}\sum_{j=0}^{k}\frac{|c_{r,k}|}{k+1}\mathbb{F}_0(y|\phi)^{k-j}dy &\leq\\ M\int\sqrt{\mathbb{F}_T(y)\left(1-\mathbb{F}_T(y)\right)}dy &\\ <\infty &\end{aligned}$$

Thus, the integral exists and is finite for sufficiently large n. This entails that:

$$O_P\left(\frac{1}{\sqrt{n}}\right)\int\sqrt{n}\left(\mathbb{F}_n(y)-\mathbb{F}_T(y)\right)\sum_{k=0}^{r-1}\sum_{j=0}^{k}\frac{c_{r,k}}{k+1}\mathbb{F}_0(y|\phi)^{k-j}dy\xrightarrow[\mathbb{P}]{n\to\infty}0,\quad\text{in probability.}\tag{4.5.12}$$

Going back to equation (4.5.10), the second term in the right hand side tends to zero in probability. We need now to treat the first term.

$$\int\sqrt{n}\left(\mathbb{F}_n(y)-\mathbb{F}_T(y)\right)\sum_{k=0}^{r-1}c_{r,k}\mathbb{F}_0(y|\phi)^k=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int\left(\mathbb{1}_{X_i\leq y}-\mathbb{F}_T(y)\right)\sum_{k=0}^{r-1}c_{r,k}\mathbb{F}_0(y|\phi)^k dy.\tag{4.5.13}$$

This is a sum of i.i.d. random variables. Before proceeding any further, it is necessary to prove that such random variables are well defined (the integrals exist) and have a finite variance. First of all, we have:

$$\begin{aligned}\int_{-\infty}^{\infty}|\mathbb{1}_{X_i\leq y}-\mathbb{F}_T(y)|\sum_{k=0}^{r-1}c_{r,k}\mathbb{F}_0(y|\phi)^k dy=&\int_{-\infty}^{X_i}\mathbb{F}_T(y)\sum_{k=0}^{r-1}c_{r,k}\mathbb{F}_0(y|\phi)^k dy\\ &+\int_{X_i}^{\infty}\left(1-\mathbb{F}_T(y)\right)\sum_{k=0}^{r-1}c_{r,k}\mathbb{F}_0(y|\phi)^k dy\end{aligned}\tag{4.5.14}$$

On the other hand, since the $|X_i|$'s have finite expectation, then X_i is finite almost surely and we have:

$$\begin{aligned}\mathbb{E}|X_i| &= \int_{t=0}^{\infty}\mathbb{P}\left(|X_i|>t\right)dt\\ &= \int_0^{\infty}\left(1-\mathbb{F}_T(t)\right)dt+\int_{-\infty}^{0}\mathbb{F}_T(t)dt.\end{aligned}$$

Thus, $\mathbb{F}_T(t)$ is integrable in the neighborhood of $-\infty$, and $1-\mathbb{F}_T(t)$ is integrable in the neighborhood of $+\infty$. This proves that the integral in equation (4.5.14) exists and is finite.

Now the random variables in (4.5.13) are well defined. The expectation is zero using the Fubini's theorem:

$$\mathbb{E}\left[\int \left(\mathbb{1}_{X_i\leq y}-\mathbb{F}_T(y)\right)\sum_{k=0}^{r-1}c_{r,k}\mathbb{F}_0(y|\phi)^k dy\right]=\int \mathbb{E}\left(\mathbb{1}_{X_i\leq y}-\mathbb{F}_T(y)\right)\sum_{k=0}^{r-1}c_{r,k}\mathbb{F}_0(y|\phi)^k dy=0.$$

The final part of the proof is to calculate the covariance matrix. Let r_1 and r_2 be two positive natural numbers such that $r_1\leq \ell$ and $r_2\leq \ell$. The Fubini's theorem yields:

$$\begin{aligned}\mathbb{E}\int \left(\mathbb{1}_{X_i\leq y}-\mathbb{F}_T(y)\right)\sum_{k=0}^{r_1-1}c_{r_1,k}\mathbb{F}_0(y|\phi)^k dy\int \left(\mathbb{1}_{X_i\leq x}-\mathbb{F}_T(x)\right)\sum_{k=0}^{r_2-1}c_{r_2,k}\mathbb{F}_0(x|\phi)^k dx=\\ \int\int \mathbb{E}\left(\mathbb{1}_{X_i\leq x}-\mathbb{F}_T(x)\right)\left(\mathbb{1}_{X_i\leq y}-\mathbb{F}_T(y)\right)\sum_{k=0}^{r_1-1}c_{r_1,k}\mathbb{F}_0(x|\phi)^k\sum_{k=0}^{r_2-1}c_{r_2,k}\mathbb{F}_0(y|\phi)^k dydx\end{aligned}$$

Denoting Σ the covariance matrix, we may write:

$$\Sigma_{r_1,r_2}=\int\int \left(\mathbb{F}_T\left(\min(x,y)\right)-\mathbb{F}_T(x)\mathbb{F}_T(y)\right)\sum_{k=0}^{r_1-1}c_{r_1,k}\mathbb{F}_0(x|\phi)^k\sum_{k=0}^{r_2-1}c_{r_2,k}\mathbb{F}_0(y|\phi)^k dydx.$$

The sum of i.i.d. variables in (4.5.13) are now well defined and the CLT applies and gives:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int \left(\mathbb{1}_{X_i\leq y}-\mathbb{F}_T(y)\right)\sum_{k=0}^{r-1}c_{r,k}\mathbb{F}_0(y|\phi)^k dy\xrightarrow{\mathcal{D}}\mathcal{N}(0,\Sigma).$$

This result together with (4.5.12) and (4.5.10) complete the proof. □

4.5.6 Proof of Theorem 4.3.2

Proof. The proof is based on a mean value expansion between $(\hat{\phi},\xi_n(\hat{\phi}))$ and $(\phi^*,0)$ similarly to the case of moment constraints Theorem 3.4.3. We therefore, need to calculate the first and second order derivatives.
First order derivatives are given by:

$$\begin{aligned}\frac{\partial H_n}{\partial \xi}(\phi,\xi) &= m(\alpha)-\int K(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy\\ \frac{\partial H_n}{\partial \alpha}(\phi,\xi) &= \xi^t\nabla m(\alpha)\\ \frac{\partial H_n}{\partial \lambda}(\phi,\xi) &= -\int\left[\frac{1}{(1-\lambda)^2}\mathbb{F}_n(y)-\frac{1}{(1-\lambda)^2}\mathbb{F}_1(y|\theta)\right]\xi^t K'(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy\\ \frac{\partial H_n}{\partial \theta}(\phi,\xi) &= \frac{\lambda}{1-\lambda}\int \nabla_\theta\mathbb{F}_1(y|\theta)\xi^t K'(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy.\end{aligned}$$

Second order derivatives are given by:

$$\begin{aligned}
\frac{\partial^2 H_n}{\partial \xi^2}(\phi,\xi) &= \int K(\hat{\mathbb{F}}_0(y|\phi))K(\hat{\mathbb{F}}_0(y|\phi))^t\psi''\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
\frac{\partial^2 H_n}{\partial \alpha^2}(\phi,\xi) &= \xi^t J_{m(\alpha)} \\
\frac{\partial^2 H_n}{\partial^2 \lambda}(\phi,\xi) &= -\int\left[\frac{2}{(1-\lambda)^3}\mathbb{F}_n(y)-\frac{2}{(1-\lambda)^3}\mathbb{F}_1(y|\theta)\right]\xi^t K'(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
&\quad -\int\left[\frac{1}{(1-\lambda)^2}\mathbb{F}_n(y)-\frac{1}{(1-\lambda)^2}\mathbb{F}_1(y|\theta)\right]^2\xi^t K''(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
&\quad -\int\left[\frac{1}{(1-\lambda)^2}\mathbb{F}_n(y)-\frac{1}{(1-\lambda)^2}\mathbb{F}_1(y|\theta)\right]^2\left[\left(\xi^t K'(\hat{\mathbb{F}}_0(y|\phi))\right)\right]^2\psi''\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
\frac{\partial^2 H_n}{\partial \theta^2}(\phi,\xi) &= \frac{\lambda}{1-\lambda}\int J_{\mathbb{F}_1(.|\theta)}\xi^t K'(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
&\quad -\frac{\lambda^2}{(1-\lambda)^2}\int \nabla_\theta \mathbb{F}_1(y|\theta)\nabla_\theta \mathbb{F}_1(y|\theta)^t\xi^t K''(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
&\quad -\frac{\lambda^2}{(1-\lambda)^2}\int \nabla_\theta \mathbb{F}_1(y|\theta)\nabla_\theta \mathbb{F}_1(y|\theta)^t\left[\xi^t K'(\hat{\mathbb{F}}_0(y|\phi))\right]^2\psi''\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy
\end{aligned}$$

Crossed derivatives:

$$\begin{aligned}
\frac{\partial^2 H_n}{\partial \xi \partial \alpha}(\phi,\xi) &= \nabla m(\alpha) \\
\frac{\partial^2 H_n}{\partial \xi \partial \lambda}(\phi,\xi) &= -\int\left[\frac{1}{(1-\lambda)^2}\mathbb{F}_n(y)-\frac{1}{(1-\lambda)^2}\mathbb{F}_1(y|\theta)\right]K'(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy - \\
&\quad \int K(\hat{\mathbb{F}}_0(y|\phi))\left[\frac{1}{(1-\lambda)^2}\mathbb{F}_n(y)-\frac{1}{(1-\lambda)^2}\mathbb{F}_1(y|\theta)\right]\xi^t K'(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
\frac{\partial^2 H_n}{\partial \xi \partial \theta}(\phi,\xi) &= \frac{\lambda}{1-\lambda}\int \nabla_\theta \mathbb{F}_1(y|\theta)K'(\hat{\mathbb{F}}_0(y|\phi))^t\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
&\quad +\frac{\lambda}{1-\lambda}\int K(\hat{\mathbb{F}}_0(y|\phi))\nabla_\theta \mathbb{F}_1(y|\theta)^t\xi^t K'(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
\frac{\partial^2 H_n}{\partial \alpha \partial \lambda}(\phi,\xi) &= 0 \\
\frac{\partial^2 H_n}{\partial \alpha \partial \theta}(\phi,\xi) &= 0 \\
\frac{\partial^2 H_n}{\lambda \partial \theta}(\phi,\xi) &= \frac{1}{(1-\lambda)^2}\int \nabla \mathbb{F}_1(y|\theta)\xi^t K'(\hat{\mathbb{F}}_0(y|\phi))\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
&\quad +\frac{\lambda}{1-\lambda}\int \nabla_\theta \mathbb{F}_1(y|\theta)\left[\frac{1}{(1-\lambda)^2}\mathbb{F}_n(y)-\frac{1}{(1-\lambda)^2}\mathbb{F}_1(y|\theta)\right]\xi^t K''(\hat{\mathbb{F}}_0(y|\phi)) \\
&\qquad \times\psi'\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy \\
&\quad +\frac{\lambda}{1-\lambda}\int \nabla_\theta \mathbb{F}_1(y|\theta)\left[\frac{1}{(1-\lambda)^2}\mathbb{F}_n(y)-\frac{1}{(1-\lambda)^2}\mathbb{F}_1(y|\theta)\right]\left[\xi^t K'(\hat{\mathbb{F}}_0(y|\phi))\right]^2 \\
&\qquad \times\psi''\left(\xi^t K(\hat{\mathbb{F}}_0(y|\phi))\right)dy
\end{aligned}$$

Notice that by assumption 1, interesting values of ξ are only in a neighborhood of the vector 0 which can be taken to be the ball $B(0,\varepsilon)$ for some $\varepsilon > 0$. Besides, the derivatives

given here above are well defined using Lebesgue theorems. Indeed, all integrands are controlled by either $K(\hat{\mathbb{F}}(y|\phi))$ or $\mathbb{F}_n(y) - \mathbb{F}_1(.|\theta)$ which are both integrable independently of ϕ as soon as $\mathbb{F}_1(.|\theta)$ has a finite expectation. Similar discussion for the former was given in Example 4.3.1, and for the later in the proof of Proposition 4.3.1 but for $\mathbb{F}_n(y) - \mathbb{F}_T(.|\theta)$ instead. Other derivatives are controlled by assumtions 4-6 of the present theorem.
A mean value expansion of the gradient of H_n between $(\hat{\phi}, \xi_n(\hat{\phi}))$ with Lagrange remainder gives that there exists $(\bar{\phi}, \bar{\xi})$ on the line between these two points such that:

$$\begin{pmatrix} \frac{\partial H_n}{\partial \phi}(\hat{\phi}, \xi(\hat{\phi})) \\ \frac{\partial H_n}{\partial \xi}(\hat{\phi}, \xi_n(\hat{\phi})) \end{pmatrix} = \begin{pmatrix} \frac{\partial H_n}{\partial \phi}(\phi^*, 0) \\ \frac{\partial H_n}{\partial \xi}(\phi^*, 0) \end{pmatrix} + J_{H_n}(\bar{\phi}, \bar{\xi}) \begin{pmatrix} \hat{\phi} - \phi^* \\ \xi_n(\hat{\phi}) \end{pmatrix}, \tag{4.5.15}$$

where $J_{H_n}(\bar{\phi}, \bar{\xi})$ is the matrix of second derivatives of H_n calculated at the mid point $(\bar{\phi}, \bar{\xi})$. First order optimality condition at $(\hat{\phi}, \xi_n(\hat{\phi}))$ is translated by:

$$\begin{aligned} \frac{\partial}{\partial \xi} H_n(\hat{\phi}, \xi_n(\hat{\phi})) &= 0 \\ \frac{\partial}{\partial \phi}\left(H_n(\phi, \xi_n(\phi))\right)\Big|_{\phi=\hat{\phi}} &= 0. \end{aligned}$$

The chain rule permits us to calculate the second line simply as a derivative with respect to ϕ calculated at the optimal point $(\hat{\phi}, \xi_n(\hat{\phi}))$, i.e.

$$\begin{aligned} \frac{\partial}{\partial \phi}\left(H_n(\phi, \xi_n(\phi))\right)\Big|_{\phi=\hat{\phi}} &= \frac{\partial}{\partial \phi} H_n(\hat{\phi}, \xi_n(\hat{\phi})) + \frac{\partial}{\partial \xi} H_n(\hat{\phi}, \xi_n(\hat{\phi})) \frac{\partial \xi_n}{\partial \phi}(\hat{\phi}) \\ &= \frac{\partial}{\partial \phi} H_n(\hat{\phi}, \xi_n(\hat{\phi})). \end{aligned}$$

Thus, optimality conditions at $(\hat{\phi}, \xi_n(\hat{\phi}))$ are given by:

$$\frac{\partial H_n}{\partial \xi}(\hat{\phi}, \xi_n(\hat{\phi})) = 0, \quad \frac{\partial H_n}{\partial \alpha}(\hat{\phi}, \xi_n(\hat{\phi})) = 0, \quad \frac{\partial H_n}{\partial \lambda}(\hat{\phi}, \xi_n(\hat{\phi})) = 0, \quad \frac{\partial H_n}{\partial \theta}(\hat{\phi}, \xi_n(\hat{\phi})) = 0.$$

On the other hand, we have at $(\phi^*, 0)$:

$$\frac{\partial H_n}{\partial \xi}(\phi^*, 0) = m(\alpha^*) - \int K(\hat{\mathbb{F}}_0(y|\phi^*))dy, \frac{\partial H_n}{\partial \alpha}(\phi^*, 0) = 0, \frac{\partial H_n}{\partial \lambda}(\phi^*, 0) = 0, \frac{\partial H_n}{\partial \theta}(\phi^*, 0) = 0.$$

By proposition 4.3.1, since $m(\alpha^*) = \int K(\mathbb{F}_0(y|\phi^*))$,

$$\sqrt{n}\left[m(\alpha^*) - \int K(\hat{\mathbb{F}}_0(y|\phi^*))dy\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma) \tag{4.5.16}$$

with Σ is the matrix of covariance defined by formula (4.3.6). It remains now to calculate the limit in probability of the matrix $J_{H_n}(\bar{\phi}, \bar{\xi})$. Recall first that as n goes to infinity $\bar{\phi} \to \phi^*$ and $\bar{\xi} \to 0$. Moreover, by the Slutsky theorem and the law of large numbers, we have:

$$\hat{\mathbb{F}}_0(y, |\bar{\phi}) = \frac{1}{1-\bar{\lambda}}\mathbb{F}_n(y) - \frac{\bar{\lambda}}{1-\bar{\lambda}}\mathbb{F}_1(y|\bar{\theta}) \xrightarrow{n\to\infty} \frac{1}{1-\lambda^*}\mathbb{F}_T(y) - \frac{\lambda^*}{1-\lambda^*}\mathbb{F}_1(y|\theta^*) = \mathbb{F}_0(y|\phi^*).$$

We may now give the limit of the blocs of the matrix $J_{H_n}(\bar{\phi}, \bar{\xi})$:

$$\begin{aligned} \frac{\partial^2 H_n}{\partial \xi^2}(\phi^*, 0) = \int K(\mathbb{F}_0(y|\phi^*))K(\mathbb{F}_0(y|\phi^*))^t dy, \qquad & \frac{\partial^2 H_n}{\partial \alpha^2}(\phi^*, 0) = 0, \\ \frac{\partial^2 H_n}{\partial^2 \lambda}(\phi^*, 0) = 0, \qquad & \frac{\partial^2 H_n}{\partial \theta^2}(\phi^*, 0) = 0. \end{aligned}$$

Crossed derivatives:

$$\frac{\partial^2 H_n}{\partial\xi\partial\alpha}(\phi^*,0)=\nabla m(\alpha^*),\quad \frac{\partial^2 H_n}{\partial\alpha\partial\lambda}(\phi^*,0)=0,\quad \frac{\partial^2 H_n}{\partial\alpha\partial\theta}(\phi^*,0)=0,\quad \frac{\partial^2 H_n}{\lambda\partial\theta}(\phi^*,0)=0,$$

$$\frac{\partial^2 H_n}{\partial\xi\partial\lambda}(\phi^*,0)=-\int\left[\frac{1}{(1-\lambda^*)^2}\mathbb{F}_T(y)-\frac{1}{(1-\lambda^*)^2}\mathbb{F}_1(y|\theta^*)\right]K'(\mathbb{F}_0(y|\phi^*))dy$$

$$\frac{\partial^2 H_n}{\partial\xi\partial\theta}(\phi^*,0)=\frac{\lambda^*}{1-\lambda^*}\int\nabla_\theta\mathbb{F}_1(y|\theta^*)K'(\mathbb{F}_0(y|\phi^*))^t dy.$$

The limit in probability of the matrix $J_{H_n}(\bar{\phi},\bar{\xi})$ can be written in the form:

$$J_H=\begin{bmatrix}0 & J_{\phi^*,\xi^*}^t\\ J_{\phi^*,\xi^*} & J_{\xi^*,\xi^*}\end{bmatrix}$$

where J_{ϕ^*,ξ^*} and J_{ξ^*,ξ^*} are given by (4.3.9) and (4.3.10). The inverse of matrix J_H has the form:

$$J_H^{-1}=\begin{pmatrix}-\tilde{\Sigma} & H\\ H^t & P\end{pmatrix},$$

where

$$\tilde{\Sigma}=\left(J_{\phi^*,\xi^*}^t J_{\xi^*,\xi^*}J_{\phi^*,\xi^*}\right)^{-1},\quad H=\tilde{\Sigma}J_{\phi^*,\xi^*}^t J_{\xi^*,\xi^*}^{-1},\quad P=J_{\xi^*,\xi^*}^{-1}-J_{\xi^*,\xi^*}^{-1}J_{\phi^*,\xi^*}\tilde{\Sigma}J_{\phi^*,\xi^*}^t J_{\xi^*,\xi^*}^{-1}$$

Going back to (4.5.15), we have:

$$\begin{pmatrix}0\\0\end{pmatrix}=\begin{pmatrix}0\\ \frac{\partial H_n}{\partial\xi}(\phi^*,0)\end{pmatrix}+J_{H_n}(\bar{\phi},\bar{\xi})\begin{pmatrix}\hat{\phi}-\phi^*\\ \xi_n(\hat{\phi})\end{pmatrix}.$$

Solving this equation in ϕ and ξ gives:

$$\begin{pmatrix}\sqrt{n}\left(\hat{\phi}-\phi^*\right)\\ \sqrt{n}\xi_n(\hat{\phi})\end{pmatrix}=J_H^{-1}\begin{pmatrix}0\\ \sqrt{n}\frac{\partial H_n}{\partial\xi}(\phi^*,0)\end{pmatrix}+o_P(1).$$

Finally, using (4.5.16), we get that:

$$\begin{pmatrix}\sqrt{n}\left(\hat{\phi}-\phi^*\right)\\ \sqrt{n}\xi_n(\hat{\phi})\end{pmatrix}\xrightarrow[\mathcal{L}]{}\mathcal{N}(0,S)$$

where

$$S=\begin{pmatrix}H\\ P\end{pmatrix}\Sigma\begin{pmatrix}H^t & P^t\end{pmatrix}.$$

This ends the proof. □

Conclusions and Perspectives

We summarize some of the most important contributions achieved in this work, and give some future perspectives and research directions concerning the different subjects presented in this manuscript.

- We studied in the first chapter the dual formula of $\varphi-$divergences and showed the limitations of the estimators built using it. We emphasized on the lack of robustness of the so-called MDφDE and explained the reason behind this problem. This permitted us to introduce a new robust estimator which we called the kernel-based MDφDE. The new estimator is proved to be consistent, asymptotically Gaussian and robust under standard conditions.

- The detailed simulation study presented at the end of Chapter 1 opened several questions. When we work with symmetric kernels, the choice of the window influences on the estimation result. Automatic methods did not give the best results and a suitably-chosen fixed value of the kernel gave always a better result. This gives rise to the question about the best window choice with respect to the estimation procedure instead of the density estimator.

- The use of asymmetric kernels is very useful and must be considered in estimation procedures which uses kernels as soon as we are working with distributions defined on a subset of $\mathbb{R}$. The choice of the kernel was not of a great importance, but the choice of the window was essential. Indeed, existing methods for the choice of the window for asymmetric kernels do not give satisfactory results and a fixed choice of the window gave clear good results almost all the time.

- We presented in the second chapter a proximal-point algorithm for the calculus of divergence-based estimators. We studied the convergence properties of this algorithm and relaxed the identifiability assumption over the proximal term.

- Our simulations show that the proximal algorithm give the same results as a direct optimization algorithm. The question is: Can the proximal algorithm give *clear* better results than direct optimization methods? We could not explore this question in the present work. We could consider a well-known model where direct optimization algorithms fail and converge to "bad" local optima and test whether our proximal algorithm succeed to give a better result.

- The role of the proximal term could also be studied. Indeed, we have noticed that the use of a proximal term of the form $\|\phi - \phi^k\|$ is not suitable for our simulations. The use of a Hellinger-type proximal term gives better results.

- We presented in the third chapter a new structure for semiparametric two-component mixture models where a component is defined through linear constraints such as moments constraints. The new structure permits the addition of a relatively general prior information about the unknown component. The new structure puts the model in between a (restrictive) fully parametric model and a complex semiparametric setup. The new structure permits to estimate the parameters of the parametric component keeping the unknown component in a neighborhood of some family of distributions. The resulting estimator is proved under standard conditions to be consistent and asymptotically normal.

- The estimation procedure presented in Chap. 3 has a linear complexity when we use the χ^2 divergence and when the constraints are polynomials in the distribution function (moments constraints). Besides, no numerical integration or smoothing are needed which permits to calculate the estimates instantly which is a clear advantage over existing methods. The later requires from several hours to several days in order to estimate the parameters of only one sample when the sample size becomes high enough (of order 10^5). Besides, it permitted in several simulated examples the identification of a parametric component even when the proportion of it is very low (of order 0.01).

- It is necessary and intriguing that we apply our new model on real data and see if we can get satisfactory results. Moreover, we should test the performance of our method on data where the true unknown component P_0^* does not verify the constraints.

- In chapter 4, we presented another structure for semiparametric two-component mixture models when one component is defined by L-moments constraints. The resulting estimator was proved to be consistent and asymptotically normal under standard conditions. In comparison to the structure introduced in Chap. 3 using moments constraints, the use of L-moments constraints shows a clear improvement of performance. In several simulations, we were able to obtain better results with L-moments constraints than with moments constraints and with a smaller number of observations.

- In the literature on L-moments, there exist some propositions and attempts to define multivariate L-moments. Our approach in Chap. 4 only treat univariate L-moments. This may be explorer in a future work.

- An important and difficult question common for both chapters 3 and 4 is: can we use less number of constraints than the number of parameters ? and even if we have an infinite number of solutions, can we ensure that these solutions are in a small neighborhood of the true set of parameters ?

Bibliography

S. M. Ali and S. D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28 (1):131–142, 1966.

Andrew R. Barron and Chyong-Hwa Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3):1347–1369, 09 1991.

Ayanendranath Basu and Bruce G. Lindsay. Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46(4):683–705, 1994.

Ayanendranath Basu and Sahadeb Sarkar. The trade-off between robustness and efficiency and the effect of model smoothing in minimum disparity inference. *Journal of Statistical Computation and Simulation*, 50:173–185, 09 1994.

Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 09 1998.

Rudolf Beran. Minimum hellinger distance estimates for parametric models. *Ann. Statist.*, 5(3):445–463, 05 1977.

C. Berge. *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces, and Convexity*. Dover books on mathematics. Dover Publications, 1963.

L. Bordes and P. Vandekerkhove. Semiparametric two-component mixture model with a known component: An asymptotically normal estimator. *Mathematical Methods of Statistics*, 19(1):22–41, 2010. ISSN 1066-5307.

Laurent Bordes, Céline Delmas, and Pierre Vandekerkhove. Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics*, 33(4):733–752, 2006.

Laurent Bordes, Didier Chauveau, and Pierre Vandekerkhove. A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis*, 51 (11):5429 – 5443, 2007. Advances in Mixture Models.

Taoufik Bouezmarni and Jeroen V.K. Rombouts. Nonparametric density estimation for multivariate bounded data. *Journal of Statistical Planning and Inference*, 140(1):139 – 152, 2010.

Taoufik Bouezmarni and Olivier Scaillet. Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data. *Econometric Theory*, 21(2):pp. 390–412, 2005.

Michel Broniatowski. Minimum divergence estimators, maximum likelihood and exponential families. *Statistics and Probability letters*, 93:27–33, 2014.

Michel Broniatowski and Alexis Decurninge. *Estimation For Models defined by conditions on their L-moments conditions*. IEEE, 2016. To appear.

Michel Broniatowski and Amor Keziou. Minimization of divergences on sets of signed measures. *Studia Sci. Math. Hungar.*, 43(4):403–442, 2006.

Michel Broniatowski and Amor Keziou. Parametric estimation and tests through divergences and the duality technique. *J. Multivariate Anal.*, 100(1):16–36, 2009a.

Michel Broniatowski and Amor Keziou. Parametric estimation and tests through divergences and the duality technique. *J. Multivariate Anal.*, 100(1):16–36, 2009b.

Michel Broniatowski and Amor Keziou. Divergences and duality for estimation and test under moment condition models. *Journal of Statistical Planning and Inference*, 142(9): 2554 – 2573, 2012.

Michel Broniatowski and Igor Vajda. Several applications of divergence criteria in continuous families. *Kybernetika*, 48(4):600–636, 2012.

Qian Chen and Richard H. Gerlach. The two-sided weibull distribution and forecasting financial tail risk. *International Journal of Forecasting*, 29(4):527 – 540, 2013.

Song Xi Chen. Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, 31(2):131 – 145, 1999.

Mohamed Cherfi. Dual φ-divergences estimation in normal models. *ArXiv e-prints*, August 2011. URL http://arxiv.org/abs/1108.2999v1.

S. Chretien and A.O. Hero. Acceleration of the em algorithm via proximal point iterations. In *Information Theory, 1998. Proceedings. 1998 IEEE International Symposium on*, pages 444–, 1998.

Stéphane Chrétien and Alfred O. Hero. Generalized proximal point algorithms and bundle implementations. Technical report, Department of Electrical Engineering and Computer Science, The University of Michigan, 1998.

Stéphane Chrétien and Alfred O. Hero. On em algorithms and their proximal generalizations. *ESAIM: Probability and Statistics*, 12:308–326, 1 2008.

Noel Cressie and Timothy R. C. Read. Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B*, 46(3):440–464, 1984. ISSN 0035-9246.

Mihai Cristea. A Note on Global Implicit Function Theorem. *Journal of Inequalities in Pure and Applied Mathematics*, 8, 2007.

I. Csiszár. Eine informationstheoretische Ungleichung und ihre anwendung auf den Beweis der ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of Hungarian Academy of Sciences*, 8:95–108, 1963.

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

Alexis Decurninge. *Univariate and multivariate quantiles, probabilistic and statistical approaches; radar applications*. Theses, Université Pierre et Marie Curie, January 2015. URL https://hal.inria.fr/tel-01129961.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, series B*, 39(1): 1–38, 1977.

David L. Donoho and Richard C. Liu. The "automatic" robustness of minimum distance functionals. *Ann. Statist.*, 16(2):552–586, 06 1988. doi: 10.1214/aos/1176350820.

W. Feller. *An introduction to probability theory and its applications*. Number vol. 2 in Wiley mathematical statistics series. Wiley, 1971.

Iva Frýdlovà, Igor Vajda, and Vàclav Kus. Modified power divergence estimators in normal models - simulation and comparative study. *Kybernetika*, 48(4):795–808, 2012.

Benedikt Funke and Rafael Kawka. Nonparametric density estimation for multivariate bounded data using two non-negative multiplicative bias correction methods. *Computational Statistics & Data Analysis*, 92:148 – 162, 2015.

Abhik Ghosh, Ian R. Harris, Avijit Maji, Ayanendranath Basu, and Leandro Pardo. A generalized divergence for statistical inference. Technical report, Byesian and Interdisciplinary Research Unit Indian Statistical Institute, 2013.

A.A. Goldstein and I.B. Russak. How good are the proximal point algorithms? *Numerical Functional Analysis and Optimization*, 9(7-8):709–724, 1987.

Mickael Guedj, Stephane Robin, Alain Celisse, and Gregory Nuel. Kerfdr: a semiparametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics*, 10(1):1–12, 2009.

J. R. M. Hosking. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(1):105–124, 1990.

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

Raùl Jiménez and Yongzhao Shao. On robustness and efficiency of minimum divergence estimators. *Test*, 10(2):241–248, 2001.

Raùl Jiménz and Yongzhao Shao. On robustness and efficiency of minimum divergence estimators. *Test*, 10(2):241–248, 2001.

R.J. Karunamuni and T. Alberts. On boundary correction in kernel density estimation. *Statistical Methodology*, 2(3):191 – 212, 2005.

R.J. Karunamuni and J. Wu. Minimum hellinger distance estimation in a nonparametric mixture model. *Journal of Statistical Planning and Inference*, 139(3):1118 – 1133, 2009.

Amor Keziou. *Utilisation des Divergences entre Mesures en Statistique Inférentielle*. Theses, Université Pierre et Marie Curie - Paris VI, November 2003. URL https://tel.archives-ouvertes.fr/tel-00004069. Patrice Bertail (rapporteur), Denis Bosq (pésident), Michel Delecroix, Dominique Picard, Ya'acov Ritov (rapporteur), Christian P. Robert, Jean-Michel Zakoian.

Arun Kumar Kuchibhotla and Ayanendranath Basu. A general set up for minimum disparity estimation. *Statistics and Probability Letters*, 96:68 – 74, 2015.

Kenneth Lange. *Optimization*. Springer Texts in Statistics. Springer-Verlag New York, 2 edition, 2013.

F. Lavancier and P. Rochet. A general procedure to combine estimators. *Computational Statistics & Data Analysis*, 94:175 – 192, 2016.

Libengué Dobélé-Kpoka Libengue Dobele-kpoka, Francial Giscard Baudin. *Non parametric method of mixed associated kernels and applications*. Theses, Université de Franche-Comté, June 2013. URL https://tel.archives-ouvertes.fr/tel-01124288.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Friedrich Liese and Igor Vajda. *Convex statistical distances*, volume 95 of *Teubner-Texte zur Mathematik [Teubner Texts in Mathematics]*. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987. ISBN 3-322-00428-7. With German, French and Russian summaries.

Bruce G. Lindsay. Efficiency versus robustness: The case for minimum hellinger distance and related methods. *Ann. Statist.*, 22(2):1081–1114, 06 1994.

Tzon-Tzer Lu and Sheng-Hua Shiou. Inverses of 2 $\times$ 2 block matrices. *Computers & Mathematics with Applications*, 43(1-2):119 – 129, 2002. ISSN 0898-1221. doi: http://dx.doi.org/10.1016/S0898-1221(01)00278-4.

Jun Ma, Sigurbjorg Gudlaugsdottir, and Graham Wood. Generalized em estimation for semi-parametric mixture distributions with discretized non-parametric component. *Statistics and Computing*, 21(4):601–612, 2011. ISSN 0960-3174.

Rostyslav Maiboroda and Olena Sugakova. Nonparametric density estimation for symmetric distributions by contaminated data. *Metrika*, 75(1):109–126, 2012.

B. Martinet. Brève communication. régularisation d'inéquations variationnelles par approximations successives. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 4(R3):154–158, 1970.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 2007.

Geoffrey McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, Inc., 2005.

Alexander Meister. *Deconvolution Problems in Nonparametric Statistics*. Lecture Notes in Statistics. Springer, 2009.

Robert Mnatsakanov and Khachatur Sarkisian. Varying kernel density estimation on $\mathbb{R}_+$. *Statistics and Probability Letters*, 82(7):1337 – 1345, 2012.

J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

Whitney K. Newey and Richard J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004. ISSN 1468-0262. doi: 10.1111/j.1468-0262.2004.00482.x. URL http://dx.doi.org/10.1111/j.1468-0262.2004.00482.x.

A.M. Ostrowski. *Solution of equations and systems of equations*. Pure and applied mathematics. Academic Press, 1966.

Art Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1): 90–120, 1990.

Leandro Pardo. *Statistical inference based on divergence measures*, volume 185 of *Statistics: Textbooks and Monographs*. Chapman & Hall/CRC, Boca Raton, FL, 2006. ISBN 978-1-58488-600-6; 1-58488-600-5.

Chanseok Park and Ayanendranath Basu. Minimum disparity estimation : Asymptotic normality and breakdown point results. *Bulletin of informatics and cybernetics*, 36: 19–33, 2004.

Rohit Kumar Patra and Bodhisattva Sen. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):869–893, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

Stéphane Robin, Avner Bar-Hen, Jean-Jacques Daudin, and Laurent Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics and Data Analysis*, 51(12):5483 – 5493, 2007. ISSN 0167-9473.

R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen. Springer, 3 edition, 1998.

Eugene F. Schuster. Estimation of a probability density function and its derivatives. *Ann. Math. Statist.*, 40(4):1187–1195, 08 1969.

Bernard W. Silverman. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.*, 6(1):177–184, 01 1978.

Douglas G. Simpson. Minimum Hellinger Distance Estimation for the Analysis of Count Data. *Journal of the American Statistical Association*, 82(399), 1987.

Seongjoo Song, Dan L. Nicolae, and Jongwoo Song. Estimating the mixing proportion in a semiparametric mixture model. *Computational Statistics and Data Analysis*, 54(10): 2276 – 2283, 2010. ISSN 0167-9473.

Stephen M. Stigler. Linear functions of order statistics with smooth weight functions. *Ann. Statist.*, 2(4):676–693, 07 1974.

Aldo Tagliani. Recovering a probability density function from its mellin transform. *Applied Mathematics and Computation*, 118(2-3):151 – 159, 2001. ISSN 0096-3003.

Roy N. Tamura and Dennis D. Boos. Minimum hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81 (393):223–229, 1986.

Qingguo Tang and Rohana J. Karunamuni. Minimum distance estimation in a finite mixture regression model. *Journal of Multivariate Analysis*, 120:185 – 204, 2013.

D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

Aida Toma. Robustness of dual divergence estimators for models satisfying linear constraints. *Comptes Rendus Mathematique*, 351(7-8):311 – 316, 2013.

Aida Toma and Michel Broniatowski. Dual divergence estimators and tests: Robustness results. *J. Multivariate Analysis*, 102(1):20–36, 2011.

Aida Toma and Samuela Leoni-Aubin. Robust tests based on dual divergence estimators and saddlepoint approximations. *Journal of Multivariate Analysis*, 101(5):1143 – 1155, 2010.

Aida Toma and Samuela Leoni-Aubin. Optimal robust m-estimators using Rényi pseudodistances. *Journal of Multivariate Analysis*, 115(C):359–373, 2013.

Paul Tseng. An Analysis of the EM Algorithm and Entropy-Like Proximal Point Methods. *Math. Oper. Res.*, 29(1):27–44, 2004.

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, 1998.

W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. Statistics and Computing. Springer New York, 2013.

Dominik Wied and Rafael Weibßach. Consistency of the kernel density estimator: a survey. *Statistical Papers*, 53(1):1–21, 2012. ISSN 0932-5026.

C. F. Jeff Wu. On the convergence properties of the em algorithm. *Ann. Statist.*, 11(1): 95–103, 03 1983.

Sijia Xiang, Weixin Yao, and Jingjing Wu. Minimum profile hellinger distance estimation for a semiparametric mixture model. *Canadian Journal of Statistics*, 42(2):246–267, 2014.

A. L. Yuille and Anand Rangarajan. The Concave-Convex Procedure (CCCP). *Neural Computation*, 15(4):915–936, 2003.

Adriano Z. Zambom and Ronaldo Dias. A Review of Kernel Density Estimation with Applications to Econometrics. *International Econometric Review (IER)*, 5(1):20–42, April 2013.