A Proximal Point Algorithm for Minimum Divergence Estimators with Application to Mixture Models

Diaa Al Mohamad, Michel Broniatowski

entropy

MDPI

Article

A Proximal Point Algorithm for Minimum Divergence Estimators with Application to Mixture Models †

Diaa Al Mohamad * and Michel Broniatowski

Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie CURIE, 4 place Jussieu, 75005 Paris, France; michel.broniatowski@upmc.fr

* Correspondence: diaa.almohamad@gmail.com; Tel.: +33-7-62-59-17-73

† This paper is an extended version of our paper published in the 2nd Conference on Geometric Science of Information, Palaiseau, France, 28–30 October 2015.

Academic Editors: Frédéric Barbaresco and Frank Nielsen
Received: 11 June 2016; Accepted: 21 July 2016; Published: 27 July 2016

Abstract: Estimators derived from a divergence criterion such as $\varphi-$divergences are generally more robust than the maximum likelihood ones. We are interested in particular in the so-called minimum dual φ–divergence estimator (MDφDE), an estimator built using a dual representation of φ–divergences. We present in this paper an iterative proximal point algorithm that permits the calculation of such an estimator. The algorithm contains by construction the well-known Expectation Maximization (EM) algorithm. Our work is based on the paper of Tseng on the likelihood function. We provide some convergence properties by adapting the ideas of Tseng. We improve Tseng's results by relaxing the identifiability condition on the proximal term, a condition which is not verified for most mixture models and is hard to be verified for "non mixture" ones. Convergence of the EM algorithm in a two-component Gaussian mixture is discussed in the spirit of our approach. Several experimental results on mixture models are provided to confirm the validity of the approach.

Keywords: φ–divergences; robust estimation; EM algorithm; proximal-point algorithms; mixture models

1. Introduction

The Expectation Maximization (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 [1]). Another way to look at the EM algorithm is as a proximal point problem (see [2,3]). 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. 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 [4]). Chrétien and Hero [5] prove superlinear convergence of a proximal point algorithm derived from 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 maximum likelihood estimator (MLE) is the least robust estimator among the class of divergence-type estimators that we present below, we generalize the EM algorithm (and the version of Tseng [2]) by replacing the

log-likelihood function by an estimator of a $\varphi-$divergence between the *true distribution* of the data and the model. A φ–divergence in the sense of Csiszár [6] is defined in the same way as [7] by:

$$D_\varphi(Q,P) = \int \varphi\left(\frac{dQ}{dP}(y)\right) dP(y),$$

where φ is a nonnegative strictly convex function. Examples of such divergences are: the Kullback–Leibler (KL) divergence , the modified KL divergence, the Hellinger distanceamong others. All these well-known divergences belong to the class of Cressie-Read functions [8] defined by

$$\varphi_\gamma(x) = \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma-1)} \text{ for } \gamma \in \mathbb{R}\setminus\{0,1\}. \tag{1}$$

for $\gamma = \frac{1}{2}, 0, 1$ respectively. For $\gamma \in \{0,1\}$, the limit is calculated, and we denote $\varphi_0(x) = -\log x + x - 1$ for the case of the modified KL and $\varphi_1(x) = x\log x - x + 1$ for the KL.

Since the $\varphi-$divergence calculus uses the unknown true distribution, we need to estimate it. We consider the dual estimator of the divergence introduced independently by [9,10]. The use of this estimator is motivated by many reasons. Its minimum coincides with the MLE for $\varphi(t) = -\log(t) + t - 1$. In addition, it has the same form for discrete and continuous models, and does not consider any partitioning or smoothing.

Let $(P_\phi)_{\phi\in\Phi}$ be a parametric model with $\Phi \subset \mathbb{R}^d$, and denote ϕ^T as the *true* set of parameters. Let dy be the Lebesgue measure defined on $\mathbb{R}$. Suppose that $\forall\phi \in \Phi$, the probability measure P_ϕ is absolutely continuous with respect to dy and denote p_ϕ the corresponding probability density. The dual estimator of the $\varphi-$divergence given an $n-$sample $y_1, \cdots, y_n$ is given by:

$$\hat{D}_\varphi(p_\phi, p_{\phi_T}) = \sup_{\alpha\in\Phi} \int \varphi'\left(\frac{p_\phi}{p_\alpha}\right)(x)p_\phi(x)dx - \frac{1}{n}\sum_{i=1}^n \varphi^\#\left(\frac{p_\phi}{p_\alpha}\right)(y_i), \tag{2}$$

with $\varphi^\#(t) = t\varphi'(t) - \varphi(t)$. Al Mohamad [11] argues that this formula works well under the model; however, when we are not, this quantity largely underestimates the divergence between the true distribution and the model, and proposes the following modification:

$$\tilde{D}_\varphi(p_\phi, p_{\phi_T}) = \int \varphi'\left(\frac{p_\phi}{K_{n,w}}\right)(x)p_\phi(x)dx - \frac{1}{n}\sum_{i=1}^n \varphi^\#\left(\frac{p_\phi}{K_{n,w}}\right)(y_i), \tag{3}$$

where $K_{n,w}$ is the Rosenblatt–Parzen kernel estimate with window parameter w. Whether it is $\hat{D}_\varphi$, or $\tilde{D}_\varphi$, the minimum dual $\varphi-$divergence estimator (MDφDE) is defined as the argument of the infimum of the dual approximation:

$$\hat{\phi}_n = \arg\inf_{\phi\in\Phi} \hat{D}_\varphi(p_\phi, p_{\phi_T}), \tag{4}$$

$$\tilde{\phi}_n = \arg\inf_{\phi\in\Phi} \tilde{D}_\varphi(p_\phi, p_{\phi_T}). \tag{5}$$

Asymptotic properties and consistency of these two estimators can be found in [7,11]. Robustness properties were also studied using the influence function approach in [11,12]. The kernel-based MDφDE (5) seems to be a *better* estimator than the classical MDφDE (4) in the sense that the former is robust whereas the later is generally not. Under the model, the estimator given by (4) is, however, more efficient, especially when the true density of the data is unbounded. More investigation is needed in the context of unbounded densities, since we may use asymmetric kernels in order to improve the efficiency of the kernel-based MDφDE, see [11] for more details.

In this paper, we propose calculation of the MDφDE using an iterative procedure based on the work of Tseng [2] on the log-likelihood function. This procedure has the form of a proximal point

algorithm, and extends the EM algorithm. Our convergence proof demands some regularity (continuity and differentiability) of the estimated divergence with respect to the parameter vector ϕ) which is not simply checked using (2). Recent results in the book of Rockafellar and Wets [13] provide sufficient conditions to prove continuity and differentiability of supremal functions of the form of (2) with respect to ϕ. Differentiability with respect to ϕ still remains a very hard task; therefore, our results cover cases when the objective function is not differentiable.

The paper is organized as follows: in Section 2, we present the general context. We also present the derivation of our algorithm from the EM algorithm and passing by Tseng's generalization. In Section 3, we present some convergence properties. We discuss in Section 4 a variant of the algorithm with a theoretical global infimum, and an example of the two-Gaussian mixture model and a convergence proof of the EM algorithm in the spirit of our approach. Finally, Section 5 contains simulations confirming our claim about the efficiency and the robustness of our approach in comparison with the MLE. The algorithm is also applied to the so-called minimum density power divergence (MDPD) introduced by [14].

2. A Description of the Algorithm

2.1. 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_is are the unobserved data (labels) and the y_is are the observations. The vector of parameters ϕ is unknown and needs to be estimated. The observed data y_i are supposed to be real numbers, and the labels x_i belong to a space $\mathcal{X}$ not necessarily finite unless mentioned otherwise. The marginal density of the observed data is given by $p_\phi(y) = \int f(x, y|\phi)dx$, where dx is a measure defined on the label space (for example, the counting measure if we work with mixture models).

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. The derivatives of a real valued function ψ defined on $\mathbb{R}$ are denoted ψ', ψ'', etc. We denote ∇f the gradient of a real function f defined on $\mathbb{R}^d$. For a generic function of two (vectorial) arguments $D(\phi|\theta)$, then $\nabla_1 D(\phi|\theta)$ denotes the gradient with respect to the first (vectorial) variable. Finally, for any set A, we use $int(A)$ to denote the interior of A.

2.2. EM Algorithm and Tseng's Generalization

The EM algorithm estimates the unknown parameter vector by (see [15]):

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

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 written as:

$$\begin{aligned} \phi^{k+1} &= \arg\max_{\Phi} \sum_{i=1}^{n} \mathbb{E}\left[\log(f(X_i, Y_i|\phi)) \,\middle|\, Y_i = y_i, \phi^k\right] \\ &= \arg\max_{\Phi} \sum_{i=1}^{n} \int_{\mathcal{X}} \log(f(x, y_i|\phi)) h_i(x|\phi^k)dx, \end{aligned} \tag{6}$$

where $h_i(x|\phi^k) = \frac{f(x,y_i|\phi^k)}{p_{\phi^k}(y_i)}$ is the conditional density of the labels (at step k) provided y_i which we suppose to be positive $dx-$almost everywhere. 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,

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

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:

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

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 distance-like function defined on the conditional densities of the labels. Tseng [2] generalizes this iteration by allowing any nonnegative convex function ψ to replace the $t \mapsto -\log(t)$ function. Tseng's recurrence is defined by:

$$\phi^{k+1} = \arg\sup_{\phi} J(\phi) - D_\psi(\phi, \phi^k), \tag{7}$$

where J is the log-likelihood function and D_ψ is given by:

$$D_\psi(\phi, \phi^k) = \sum_{i=1}^{n} \int_{\mathcal{X}} \psi\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k)dx, \tag{8}$$

for any real nonnegative convex function ψ such that $\psi(1) = \psi'(1) = 0$. $D_\psi(\phi_1, \phi_2)$ is nonnegative, and $D_\psi(\phi_1, \phi_2) = 0$ if and only if $\forall i, h_i(x|\phi_1) = h_i(x|\phi_2)\ dx$ almost everywhere.

2.3. Generalization of Tseng's Algorithm

We use the relationship between maximizing the log-likelihood and minimizing the Kullback–Liebler divergence to generalize the previous algorithm. We, therefore, replace the log-likelihood function by an estimate of a $\varphi-$divergence D_φ between the true distribution and the model. We use the dual estimators of the divergence presented earlier in the introduction (2) or (3), which we denote in the same manner $\hat{D}_\varphi$, unless mentioned otherwise. Our new algorithm is defined by:

$$\phi^{k+1} = \arg\inf_{\phi} \hat{D}_\varphi(p_\phi, p_{\phi_T}) + \frac{1}{n} D_\psi(\phi, \phi^k), \tag{9}$$

where $D_\psi(\phi, \phi^k)$ is defined by (8). When $\varphi(t) = -\log(t) + t - 1$, it is easy to see that we get recurrence (7). Indeed, for the case of (2) we have:

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

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 easily get (7). The same applies for the case of (3). For notational simplicity, from now on, we redefine D_ψ with a normalization by n, i.e.,

$$D_\psi(\phi, \phi^k) = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{X}} \psi\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k)dx. \tag{10}$$

Hence, our set of algorithms is redefined by:

$$\phi^{k+1} = \arg\inf_{\phi} \hat{D}_\varphi(p_\phi, p_{\phi_T}) + D_\psi(\phi, \phi^k). \tag{11}$$

We will see later that this iteration forces the 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 (11) being a way to calculate both the MDφDE (4) and the kernel-based MDφDE (5).

3. Some Convergence Properties of ϕ^k

We show here how, according to some possible situations, one may prove convergence of the algorithm defined by (11). Let ϕ^0 be a given initialization, and define

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

which we suppose to be a subset of $int(\Phi)$. The idea of defining this set in this context is inherited from the paper Wu [16], 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 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.*

Remark 1. *In the whole paper, the subgradient is defined for any function not necessarily convex (see Definition 8.3) in [13] for more details.*

We will be using the following assumptions:

- 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. Function $\phi \mapsto \nabla\hat{D}_\varphi(p_\phi|p_{\phi_T})$ 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 that we suppose that $h_i(x|\phi) > 0, dx - a.e.$ We relax the convexity assumption of function ψ. We only suppose that ψ is nonnegative and $\psi(t) = 0$ iff $t = 1$. In addition, $\psi'(t) = 0$ if $t = 1$.

Continuity and differentiability assumptions of function $\phi \mapsto \hat{D}_\varphi(p_\phi|p_{\phi_T})$ for the case of (3) can be easily checked using Lebesgue theorems. The continuity assumption for the case of (2) can be checked using Theorem 1.17 or Corollary 10.14 in [13]. Differentiability can also be checked using Corollary 10.14 or Theorem 10.31 in the same book. In what concerns D_ψ, continuity and differentiability can be obtained merely by fulfilling Lebesgue theorems conditions. When working with mixture models, we only need the continuity and differentiability of ψ and functions h_i. The later is easily deduced from regularity assumptions on the model. For assumption A2, there is no universal method, see Section 4.2 for an Example. Assumption A3 can be checked using Lemma 2 in [2].

We start the convergence properties by proving that the objective function $\hat{D}_\varphi(p_\phi|p_{\phi_T})$ decreases alongside the the sequence $(\phi^k)_k$, and give a possible set of conditions for the existence of the sequence $(\phi^k)_k$.

Proposition 1. *(a) Assume that the sequence $(\phi^k)_k$ is well defined in Φ, then $\hat{D}_\varphi(p_{\phi^{k+1}}|p_{\phi_T}) \leq \hat{D}_\varphi(p_{\phi^k}|p_{\phi_T})$, and (b) $\forall k, \phi^k \in \Phi^0$. (c) Assume A0 and A2 are verified, then the sequence $(\phi^k)_k$ is defined and bounded. Moreover, the sequence $(\hat{D}_\varphi(p_{\phi^k}|p_{\phi^T}))_k$ converges.*

Proof. We prove (a). 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 side and that $D_\psi(\phi^{k+1}, \phi^k) \geq 0$ for the left-hand side of the previous inequality. Hence, $\hat{D}_\varphi(p_{\phi^{k+1}}, p_{\phi_T}) \leq \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T})$.

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 directly by definition of Φ^0.

We prove (c) by induction on k. For $k = 0$, clearly ϕ^0 is well defined since we choose it. The choice of the initial point ϕ^0 of the sequence may influence the convergence of the sequence. See the Example of the Gaussian mixture in Section 4.2. Suppose, for some $k \geq 0$, that ϕ^k exists. We prove that the infimum is attained in Φ^0. Let $\phi \in \Phi$ be any vector 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.

Convergence of the sequence $(\hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}))_k$ 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. □

Compactness in part (c) can be replaced by inf-compactness of function $\phi \mapsto \hat{D}_\varphi(p_\phi|p_{\phi_T})$ and continuity of D_ψ with respect to its first argument. The convergence of the sequence $(\hat{D}_\varphi(\phi^k|\phi_T))_k$ is an interesting property, since, in general, there is no theoretical guarantee, or it is difficult to prove that the whole sequence $(\phi^k)_k$ converges. It may also continue to fluctuate around a minimum. The decrease of the error criterion $\hat{D}_\varphi(\phi^k|\phi_T)$ between two iterations helps us decide when to stop the iterative procedure.

Proposition 2. *Suppose A1 verified, Φ^0 is closed and $\{\phi^{k+1} - \phi^k\} \to 0$.*

(a) If AC is verified, then any limit point of $(\phi^k)_k$ is a stationary point of $\phi \mapsto \hat{D}_\varphi(p_\phi|p_{\phi^T})$;

(b) If AC is dropped, then any limit point of $(\phi^k)_k$ is a "generalized" stationary point of $\phi \mapsto \hat{D}_\varphi(p_\phi|p_{\phi^T})$, i.e., zero belongs to the subgradient of $\phi \mapsto \hat{D}_\varphi(p_\phi|p_{\phi^T})$ calculated at the limit point.

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 1b).

Let us now show that the subsequence (ϕ^{n_k+1}) also converges to ϕ^∞. We simply have:

$$\|\phi^{n_k+1} - \phi^\infty\| \quad \leq \quad \|\phi^{n_k} - \phi^\infty\| + \|\phi^{n_k+1} - \phi^{n_k}\|.$$

Since $\phi^{k+1} - \phi^k \to 0$ and $\phi^{n_k} \to \phi^\infty$, we conclude that $\phi^{n_k+1} \to \phi^\infty$.

By definition of ϕ^{n_k+1}, it verifies the infimum in recurrence (11), 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 $\psi'(1) = 0$) for any $\phi \in \Phi$

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

which is equal to zero since $\psi'(1) = 0$. This implies that $\nabla \hat{D}_\varphi(p_{\phi^\infty}, p_{\phi_T}) = 0$.

We prove (b). We use again the definition of the arginf. As the optimized function is not necessarily differentiable at the points of the sequence $(\phi^k)_k$, 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 [13] Chapter 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:

$$\limsup_{\phi^{n_k+1} \to \phi^\infty} \partial \hat{D}_\varphi(p_{\phi^{n_k+1}}, p_{\phi_T}) \subset \partial \hat{D}_\varphi(p_{\phi^\infty}, p_{\phi_T}). \tag{12}$$

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 the (outer) limsup (see [13] Chapter 4, Definition 1 or Chapter 5B):

$$\limsup_{\phi \to \phi^\infty} \partial \hat{D}_\varphi(p_\phi, p_{\phi_T}) = \left\{ u | \exists \phi^k \to \phi^\infty, \exists u^k \to u \text{ with } u^k \in \partial \hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}) \right\}.$$

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 (12), we get our result:

$$0 \in \partial \hat{D}_\varphi(p_{\phi^\infty}, p_{\phi_T}).$$

This ends the proof. □

The assumption $\{\phi^{k+1} - \phi^k\} \to 0$ used in Proposition 2 is not easy to be checked unless one has a close formula of ϕ^k. The following proposition gives a method to prove such assumption. This method seems simpler, but it is not verified in many mixture models (see Section 4.2 for a counter Example).

Proposition 3. *Assume that A1, A2 and A3 are verified, then* $\{\phi^{k+1} - \phi^k\} \to 0$. *Thus, by Proposition 2 (according to whether AC is verified or not), any limit point of the sequence* ϕ^k *is a (generalized) stationary point of* $\hat{D}_\varphi(.|\phi_T)$.

Proof. By contradiction, let us 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 1c. 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 1, we can deduce the following inequality:

$$\hat{D}(p_{\phi^{k+1}}, p_{\phi_T}) + D_\psi(\phi^{k+1}, \phi^k) \leq \hat{D}(p_{\phi^k}, p_{\phi_T}), \tag{13}$$

which is also verified for 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 nonnegative, 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. □

Corollary 1. *Under assumptions of Proposition 3, the set of accumulation points of* $(\phi^k)_k$ *is a connected compact set. Moreover, if* $\phi \mapsto \hat{D}(p_\phi, p_{\phi_T})$ *is strictly convex in the neighborhood of a limit point of the sequence* $(\phi^k)_k$, *then the whole sequence* $(\phi^k)_k$ *converges to a local minimum of* $\hat{D}(p_\phi, p_{\phi_T})$.

Proof. Since the sequence $(\phi)_k$ is bounded and verifies $\phi^{k+1} - \phi^k \to 0$, then Theorem 28.1 in [17] 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. The remaining of the proof is a direct result of Theorem 3.3.1 from [18]. The strict concavity of the objective function around an accumulation point is replaced here by the strict convexity of the estimated divergence. □

Proposition 3 and Corollary 1 describe what we may hope to get of the sequence ϕ^k. Convergence of the whole sequence is bound by a local convexity assumption in the neighborhood of a limit point. Although simple, this assumption remains difficult to be checked since we do not know where might be the limit points. In addition, assumption A3 is very restrictive, and is not verified in mixture models.

Propositions 2 and 3 were developed for the likelihood function in the paper of Tseng [2]. Similar results for a general class of functions replacing $\hat{D}_\varphi$ and D_ψ which may not be differentiable (but still continuous) are presented in [3]. In these results, assumption A3 is essential. Although in [18] this problem is avoided, their approach demands that the log-likelihood has $-\infty$ limit as $\|\phi\| \to \infty$. This is simply not verified for mixture models. We present a similar method to the one in [18] based on the idea of Tseng [2] of using the set Φ^0 which is valid for mixtures. We lose, however, the guarantee of consecutive decrease of the sequence $(\phi^k)_k$.

Proposition 4. *Assume A1, AC and A2 verified. Any limit point of the sequence* $(\phi^k)_k$ *is a stationary point of* $\phi \to \hat{D}(p_\phi, p_{\phi_T})$. *If AC is dropped, then 0 belongs to the subgradient of* $\phi \mapsto \hat{D}(p_\phi, p_{\phi_T})$ *calculated at the limit point.*

Proof. If $(\phi^k)_k$ converges to, say, ϕ^∞, then 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 us take the subsequence $(\phi^{N_0(k)-1})_k$. This subsequence does not necessarily converge; it is still 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 (13):

$$\hat{D}(p_{\phi^{k+1}}, p_{\phi_T}) + D_\psi(\phi^{k+1}, \phi^k) \leq \hat{D}(p_{\phi^k}, p_{\phi_T}).$$

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 (10), we get that if the sum is zero, then each term is also zero since all terms are nonnegative. 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 nonnegative functions, so they vanish almost everywhere 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 (which can be ensured by a suitable choice of the initial point ϕ^0), 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:

$$\forall i \in \{1, \cdots, n\}, \quad h_i(x|\tilde{\phi}) = h_i(x|\bar{\phi}) \quad dx - a.e. \tag{14}$$

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:

$$\nabla \hat{D}(p_{\tilde{\phi}}, p_{\phi_T}) + \nabla D_\psi(\tilde{\phi}, \bar{\phi}) = 0. \tag{15}$$

Let us 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 (14), and the fact that $\psi'(1) = 0$, to deduce that:

$$\nabla D_\psi(\tilde{\phi}, \bar{\phi}) = 0.$$

This entails using (15) 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 2b. are employed. The optimality condition in (11) 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:

$$\limsup_{\phi^{k+1} \to \phi^\infty} \partial \hat{D}_\varphi(p_{\phi^{k+1}}, p_{\phi_T}) \subset \partial \hat{D}_\varphi(p_{\tilde{\phi}}, p_{\phi_T}). \tag{16}$$

By definition of the 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 the convergence of $(\hat{D}_\varphi(p_{\phi^k}, p_{\phi_T}))_k$, inequality (13) and the 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 (16), we get our result:

$$0 \in \partial \hat{D}_\varphi(p_{\tilde{\phi}}, p_{\phi_T}),$$

which ends the proof. □

The proof of the previous proposition is very similar to the proof of Proposition 2. The key idea is to use the sequence of conditional densities $h_i(x|\phi^k)$ instead of the sequence ϕ^k. According to the application, one may be interested only in Proposition 1 or in Propositions 2–4. If one is interested in the parameters, Propositions 2 to 4 should be used, since we need a stable limit of $(\phi^k)_k$. If we are only interested in minimizing an error criterion $\hat{D}_\varphi(p_\phi, p_{\phi^T})$ between the estimated distribution and the true one, Proposition 1 should be sufficient.

4. Case Studies

4.1. An Algorithm With Theoretically Global Infimum Attainment

We present a variant of algorithm (11) which ensures *theoretically* the convergence to a global infimum of the objective function $\hat{D}_\varphi(p_\phi, p_{\phi^T})$ as soon as there exists a convergent subsequence of $(\phi^k)_k$. The idea is the same as Theorem 3.2.4 in [18]. 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 1b). Suppose also that the sequence $(\beta_k)_k$ converges to 0 as k goes to infinity.

Now assumptions of Theorem 3.2.4. from [18] are verified. Thus, using the same lines from the proof of this theorem (inverting all inequalities since we are minimizing instead of maximizing), we may prove that ϕ^∞ is a global infimum of the estimated divergence, that is

$$\hat{D}_\varphi(p_{\phi^\infty}, p_{\phi^T}) \leq \hat{D}_\varphi(p_\phi, p_{\phi^T}), \qquad \forall \phi \in \Phi.$$

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 that can perfectly solve non convex optimization problems (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 $\hat{D}_\varphi(p_\phi, p_{\phi^T})$). Although in our approach, we use a similar assumption to prove the consecutive decreasing of $\hat{D}_\varphi(p_\phi, p_{\phi^T})$, we can replace the infimum calculus in (11) 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 minima 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.

4.2. The Two-Component Gaussian Mixture

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

$$p_{\lambda,\mu}(x) = \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_1)^2} + \frac{1-\lambda}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_2)^2}. \tag{17}$$

Here, $\Phi = [\eta, 1-\eta] \times \mathbb{R}^2$. The regularization term D_ψ is defined by (8) where:

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

Functions h_i are clearly of class $\mathcal{C}^1(\text{int}(\Phi))$, and so does D_ψ. We prove that Φ^0 is closed and bounded, which is sufficient to conclude its compactness, since the space $[\eta, 1-\eta] \times \mathbb{R}^2$ provided with the euclidean distance is complete.

If we are using the dual estimator of the $\varphi-$divergence given by (2), then assumption A0 can be verified using the maximum theorem of Berge [19]. 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 order to prove the continuity of the estimated divergence, we need 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, using Theorem 10.31 from [13], $\phi \mapsto \hat{D}_\varphi(p_\phi, p_{\phi^T})$ 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 4 and 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 (2) 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 (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). There is one condition which relates the window of the kernel, say w, with the value of γ. Indeed, using Formula (3), we can write

$$\hat{D}_\varphi(p_\phi, p_{\phi^T}) = \frac{1}{\gamma - 1} \int \frac{p_\phi^\gamma}{K_{n,w}^{\gamma-1}}(y)dy - \frac{1}{\gamma n} \sum_{i=1}^{n} \frac{p_\phi^\gamma}{K_{n,w}^{\gamma}}(y_i) - \frac{1}{\gamma(\gamma-1)}.$$

In order to study the continuity and the differentiability of the estimated divergence with respect to ϕ, it suffices to study the integral term. We have

$$\frac{p_\phi^\gamma}{K_{n,w}^{\gamma-1}}(y) = \frac{\left(\frac{\lambda}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(y-\mu_1)^2\right] + \frac{1-\lambda}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(y-\mu_2)^2\right]\right)^\gamma}{\left(\frac{1}{nw} \sum_{i=1}^{n} \exp\left[-\frac{(y-y_i)^2}{2w^2}\right]\right)^{\gamma-1}}.$$

The dominating term at infinity in the nominator is $\exp(-\gamma y^2/2)$, whereas it is $\exp(-(\gamma-1)y^2/(2w^2))$ in the denominator. It suffices now in order that the integrand to be bounded by an integrable function independently of $\phi = (\lambda, \mu)$ that we have $-\gamma + (\gamma-1)/w^2 < 0$. That is $-\gamma w^2 + \gamma - 1 < 0$, which is equivalent to $\gamma(w^2 - 1) < -1$. This argument also holds if we differentiate the integrand with respect to λ or either of the means μ_1 or μ_2. For $\gamma = 2$ (the Pearson's χ^2), we need $w^2 > 1/2$. For $\gamma = 1/2$ (the Hellinger), there is no condition on w.

Closedness of Φ^0 is proved similarly to the previous case. Boundedness, however, must be treated differently since Φ is not necessarily compact and is supposed to be $\Phi = [\eta, 1-\eta] \times \mathbb{R}^2$. 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 calculation of the limits when either (or both) of the means tends to infinity. If both the means go 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:

$$\hat{D}_\varphi(p_{\phi^0}, p_{\phi^T}) < \min\left(\frac{1}{\gamma(\gamma-1)}, \inf_{\lambda,\mu} \hat{D}_\varphi(p_{(\lambda,\infty,\mu)}, p_{\phi^T})\right) \text{ if } \gamma \in (0,\infty) \setminus \{1\}, \tag{18}$$

$$\hat{D}_\varphi(p_{\phi^0}, p_{\phi^T}) < \inf_{\lambda,\mu} \hat{D}_\varphi(p_{(\lambda,\infty,\mu)}, p_{\phi^T}) \qquad \text{if } \gamma < 0, \tag{19}$$

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 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. In addition, the definition of Φ^0 permits to conclude that if ϕ^0 is chosen according to conditions (18) and (19), 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 the 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 4 and 1, under conditions (18) and (19) the sequence* $(\hat{D}_\varphi(p_{\phi^k}, p_{\phi^T}))_k$ *defined through Formula (3) converges and there exists a subsequence* $(\phi^{N(k)})$ *that 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:

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

where $J_{\mathcal{N}}$ is the log-likelihood function of the Gaussian mixture model. The log-likelihood function $J_{\mathcal{N}}$ is clearly of class $\mathcal{C}^1(\text{int}(\Phi))$. We prove that Φ^0 is closed and bounded which is sufficient to conclude its compactness, since the space $[\eta, 1-\eta] \times \mathbb{R}^2$ provided with the euclidean distance is complete.

Closedness. The set Φ^0 is the inverse image by a continuous function (the log-likelihood) of a closed set. Therefore it is closed in $[\eta, 1-\eta] \times \mathbb{R}^2$.

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 $J_{\mathcal{N}}(\phi^l) \to -\infty$. Any finite initialization ϕ^0 will imply that $J_{\mathcal{N}}(\phi^0) > -\infty$ so that $\forall \phi \in \Phi^0, J_{\mathcal{N}}(\phi) \geq J_{\mathcal{N}}(\phi^0) > -\infty$. 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 (or that μ_2^l is bounded; in such case we extract a convergent subsequence) 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 $J_{\mathcal{N}}(\phi^0) > J_{\mathcal{N}}\left(0, \infty, \frac{1}{n}\sum_{i=1}^{n} y_i\right)$, 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 $J_{\mathcal{N}}(\phi^0) > J_{\mathcal{N}}\left(1, \frac{1}{n}\sum_{i=1}^{n} y_i, \infty\right)$. In conclusion, with a choice of ϕ^0 such that:

$$J_{\mathcal{N}}(\phi^0) > \max\left[J_{\mathcal{N}}\left(0, \infty, \frac{1}{n}\sum_{i=1}^{n} y_i\right),\ J_{\mathcal{N}}\left(1, \frac{1}{n}\sum_{i=1}^{n} y_i, \infty\right)\right], \tag{20}$$

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 calculating the corresponding log-likelihood value. If $J_{\mathcal{N}}(\phi^0)$ does not verify the previous condition, we draw again another random vector until satisfaction.

Conclusion 3. *Using Propositions 4 and 1, under condition (20) the sequence $(J_{\mathcal{N}}(\phi^k))_k$ 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 [2], 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 us suppose, absurdly, 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 nonnegative 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 have 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:

$$\left\{ \begin{array}{rcl} \mu_1 - \mu_2 &=& \mu_1' - \mu_2' \\ \log\left(\frac{1-\lambda}{\lambda}\right) + \frac{1}{2}\mu_1^2 - \frac{1}{2}\mu_2^2 &=& \log\left(\frac{1-\lambda'}{\lambda'}\right) + \frac{1}{2}\mu_1'^2 - \frac{1}{2}\mu_2'^2. \end{array} \right. \tag{21}$$

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. *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.*

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

5. Simulation Study

We summarize the results of 100 experiments on 100 samples by giving the average of the estimates and the error committed, and the corresponding standard deviation. The criterion error is the total variation distance (TVD), which is calculated using the $L1$ distance. Indeed, the Scheffé Lemma (see [20] (Page 129)) states that:

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

The TVD gives a measure of the maximum error we may commit when we use the estimated model in lieu of the true distribution. We consider the Hellinger divergence for estimators based on $\varphi-$divergences, which corresponds to $\varphi(t) = \frac{1}{2}(\sqrt{t} - 1)^2$. Our preference of the Hellinger divergence is that we hope to obtain robust estimators without loss of efficiency (see [21]). D_ψ is calculated with $\psi(t) = \frac{1}{2}(\sqrt{t} - 1)^2$. The kernel-based MD$\varphi$DE is calculated using the Gaussian kernel, and the window

is calculated using Silverman's rule. We included in the comparison the minimum density power divergence (MDPD) of [14]. The estimator is defined by:

$$\begin{aligned}\hat{\phi}_n &= \arg\inf_{\phi\in\Phi} \int p_\phi^{1+a}(z)dz - \frac{a+1}{a}\frac{1}{n}\sum_i^n p_\phi^a(y_i) \\ &= \arg\inf_{\phi\in\Phi} \mathbb{E}_{P_\phi}\left[p_\phi^a\right] - \frac{a+1}{a}\mathbb{E}_{P_n}\left[p_\phi^a\right], \end{aligned} \tag{22}$$

where $a \in (0,1]$. This is a Bregman divergence and is known to have good efficiency and robustness for a good choice of the tradeoff parameter. According to the simulation results in [11], the value of $a = 0.5$ seems to give a good tradeoff between robustness against outliers and a good performance under the model. Notice that the MDPD coincides with MLE when a tends to zero. Thus, our methodology presented here in this article, is applicable on this estimator and the proximal point algorithm can be used to calculate the MDPD. The proximal term will be kept the same, i.e., $\psi(t) = \frac{1}{2}(\sqrt{t}-1)^2$.

Remark 3 (Note on the robustness of the used estimators)**.** *In Section 3, we have proved under mild conditions that the proximal point algorithm (11) ensures the decrease of the estimated divergence. This means that when we use the dual Formulas (2) and (3), then the proximal point algorithm (11) returns at convergence the estimators defined by (4) and (5), respectively. Similarly, if we use the density power divergence of Basu et al. [14], then the proximal-point algorithm returns at convergence the MDPD defined by (22). The robustness properties of the dual estimators (4) and (5) are studied in [12] and [11] respectively using the influence function (IF) approach. On the other hand, the robustness properties of the MDPD are studied using the IF approach in [14]. The MDφDE (4) has generally an unbounded IF (see [12] Section 3.1), whereas the kernel-based MDφDE's IF may be bounded for example in a Gaussian model and for any φ−divergence with $\varphi = \varphi_\gamma$ with $\gamma \in (0,1)$, see [11] Example 2. On the other hand, the MDPD has generally a bounded IF if the tradeoff parameter a is positive, and, in particular, in the Gaussian model. The MDPD becomes more robust as the tradeoff parameter* a *increases (see Section 3.3 in [14]). Therefore, we should expect that the proximal point algorithm produces robust estimators in the case of the kernel-based MDφDE and the MDPD, and thus obtain better results than the MLE calculated using the EM algorithm.*

Simulations from two mixture models are given below—a Gaussian mixture and a Weibull mixture. The MLE for both mixtures was calculated using the EM algorithm.

Optimizations were carried out using the Nelder–Mead algorithm [22] under the statistical tool R [23]. Numerical integrations in the Gaussian mixture were calculated using the `distrExIntegrate` function of package `distrEx`. It is a slight modification of the standard function `integrate`. It performs a Gauss–Legendre quadrature when function `integrate` returns an error. In the Weibull mixture, we used the `integral` function from package `pracma`. 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. Although function `integral` is slow, it performs better than other functions even if the integrand has a relatively bad behavior.

5.1. The Two-Component Gaussian Mixture Revisited

We consider the Gaussian mixture (17) presented earlier with true parameters $\lambda = 0.35$, $\mu_1 = -2, \mu_2 = 1.5$ and known variances equal to 1. Contamination was done by adding in the original sample to the five lowest values random observations from the uniform distribution $\mathcal{U}[-5,-2]$. We also added to the five largest values random observations from the uniform distribution $\mathcal{U}[2,5]$. Results are summarized in Table 1. The EM algorithm was initialized according to condition (20). This condition gave good results when we are under the model, whereas it did not always result in good estimates (the proportion converged towards 0 or 1) when outliers were added, and thus the EM algorithm was reinitialized manually.

Table 1. The mean and the standard deviation of the estimates and the errors committed in a 100 run experiment of a two-component Gaussian mixture. The true set of parameters is $\lambda = 0.35$, $\mu_1 = -2$, $\mu_2 = 1.5$.

Estimation Method	λ	sd (λ)	μ_1	sd (μ_1)	μ_2	sd (μ_2)	TVD	sd (TVD)
Without Outliers								
Classical MDφDE	0.349	0.049	–1.989	0.207	1.511	0.151	0.061	0.029
New MDφDE–Silverman	0.349	0.049	–1.987	0.208	1.520	0.155	0.062	0.029
MDPD $a = 0.5$	0.360	0.053	–1.997	0.226	1.489	0.135	0.065	0.025
EM (MLE)	0.360	0.054	–1.989	0.204	1.493	0.136	0.064	0.025
With 10% Outliers								
Classical MDφDE	0.357	0.022	–2.629	0.094	1.734	0.111	0.146	0.034
New MDφDE–Silverman	0.352	0.057	–1.756	0.224	1.358	0.132	0.087	0.033
MDPD $a = 0.5$	0.364	0.056	–1.819	0.218	1.404	0.132	0.078	0.030
EM (MLE)	0.342	0.064	–2.617	0.288	1.713	0.172	0.150	0.034

Figure 1 shows the values of the estimated divergence for both Formulas (2) and (3) on a logarithmic scale at each iteration of the algorithm.

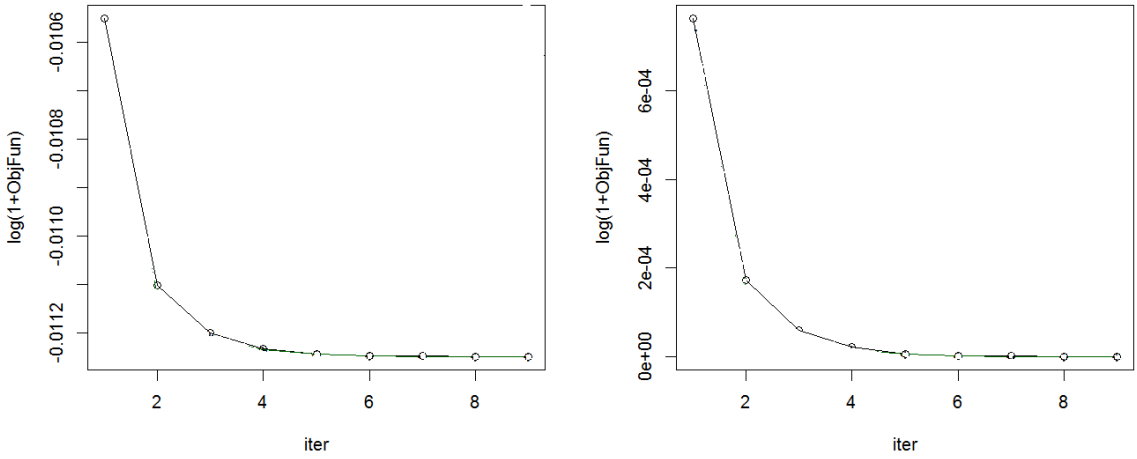

Figure 1. Decrease of the (estimated) Hellinger divergence between the true density and the estimated model at each iteration in the Gaussian mixture. The figure to the left is the curve of the values of the kernel-based dual Formula (3). The figure to the right is the curve of values of the classical dual Formula (2). Values are taken at a logarithmic scale $\log(1 + x)$.

Concerning our simulation results, the total variation of all four estimation methods is very close when we are under the model. When we added outliers, the classical MDφDE was as sensitive as the maximum likelihood estimator. The error was doubled. Both the kernel-based MDφDE and the MDPD are clearly robust since the total variation of these estimators under contamination has slightly increased.

5.2. The Two-Component Weibull Mixture Model

We consider a two-component Weibull mixture with unknown shapes $\nu_1 = 1.2, \nu_2 = 2$ and a proportion $\lambda = 0.35$. The scales are known an equal to $\sigma_1 = 0.5, \sigma_2 = 2$. The desity function is given by:

$$p_\phi(x) = 2\lambda\alpha_1(2x)^{\alpha_1 - 1}e^{-(2x)^{\alpha_1}} + (1-\lambda)\frac{\alpha_2}{2}\left(\frac{x}{2}\right)^{\alpha_2 - 1}e^{-\left(\frac{x}{2}\right)^{\alpha_2}}. \tag{23}$$

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 summarized in Table 2. Notice that it would have been better to use asymmetric kernels in order to build the kernel-based MDφDE since their use in the context of positive-supported distributions is advised in order to reduce the bias at zero, see [11] for a detailed comparison with symmetric kernels. This is not, however, the goal of this paper. In addition, the use of symmetric kernels in this mixture model gave satisfactory results.

Simulations results in Table 2 confirm once more the validity of our proximal point algorithm and the clear robustness of both the kernel-based MDφDE and the MDPD.

Table 2. The mean and the standard deviation of the estimates and the errors committed in a 100-run experiment of a two-component Weibull mixture. The true set of parameter is $\lambda = 0.35, \nu_1 = 1.2, \nu_2 = 2$.

Estimation Method	λ	sd (λ)	μ_1	sd (μ_1)	μ_2	sd (μ_2)	TVD	sd (TVD)
Without Outliers								
Classical MDφDE	0.356	0.066	1.245	0.228	2.055	0.237	0.052	0.025
New MDφDE–Silverman	0.387	0.067	1.229	0.241	2.145	0.289	0.058	0.029
MDPD $a = 0.5$	0.354	0.068	1.238	0.230	2.071	0.345	0.056	0.029
EM (MLE)	0.355	0.066	1.245	0.228	2.054	0.237	0.052	0.025
With 10% Outliers								
Classical MDφDE	0.250	0.085	1.089	0.300	1.470	0.335	0.092	0.037
New MDφDE–Silverman	0.349	0.076	1.122	0.252	1.824	0.324	0.067	0.034
MDPD $a = 0.5$	0.322	0.077	1.158	0.236	1.858	0.344	0.060	0.029
EM (MLE)	0.259	0.095	0.941	0.368	1.565	0.325	0.095	0.035

6. Conclusions

We introduced in this paper a proximal-point algorithm that permits calculation of divergence-based estimators. We studied the theoretical convergence of the algorithm and verified it in a two-component Gaussian mixture. We performed several simulations which confirmed that the algorithm works and is a way to calculate divergence-based estimators. We also applied our proximal algorithm on a Bregman divergence estimator (the MDPD), and the algorithm succeeded to produce the MDPD. Further investigations about the role of the proximal term and a comparison with direct optimization methods in order to show the practical use of the algorithm may be considered in a future work.

Acknowledgments: The authors are grateful to Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie, for financial support.

Author Contributions: Michel Broniatowski proposed use of a proximal-point algorithm in order to calculate the MDφDE. Michel Broniatowski proposed building a work based on the paper of [2]. Diaa Al Mohamad proposed the generalization in Section 2.3 and provided all of the convergence results in Section 3. Diaa Al Mohamad also conceived the simulations. Finally, Michel Broniatowski contributed to improving the text written by Diaa Al Mohamad. Both authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McLachlan, G.J.; Krishnan, T. *The EM Algorithm and Extensions*; Wiley: Hoboken, NJ, USA, 2007.
2. Tseng, P. An Analysis of the EM Algorithm and Entropy-Like Proximal Point Methods. *Math. Oper. Res.* **2004**, *29*, 27–44.
3. Chrétien, S.; Hero, A.O. Generalized Proximal Point Algorithms and Bundle Implementations. Available online: http://www.eecs.umich.edu/techreports/systems/cspl/cspl-316.pdf (acceesed on 25 July 2016).
4. Goldstein, A.; Russak, I. How good are the proximal point algorithms? *Numer. Funct. Anal. Optim.* **1987**, *9*, 709–724.

5. Chrétien, S.; Hero, A.O. Acceleration of the EM algorithm via proximal point iterations. In Proceedings of the IEEE International Symposium on Information Theory, Cambridge, MA, USA, 16–21 August 1998.
6. Csiszár, I. Eine informationstheoretische Ungleichung und ihre anwendung auf den Beweis der ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hung. Acad. Sci.* **1963**, *8*, 95–108. (In German)
7. Broniatowski, M.; Keziou, A. Parametric estimation and tests through divergences and the duality technique. *J. Multivar. Anal.* **2009**, *100*, 16–36.
8. Cressie, N.; Read, T.R.C. Multinomial goodness-of-fit tests. *J. R. Stat. Soc. Ser. B* **1984**, *46*, 440–464.
9. Broniatowski, M.; Keziou, A. Minimization of divergences on sets of signed measures. *Stud. Sci. Math. Hung.* **2006**, *43*, 403–442.
10. Liese, F.; Vajda, I. On Divergences and Informations in Statistics and Information Theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412.
11. Al Mohamad, D. Towards a better understanding of the dual representation of phi divergences. **2016**, arXiv:1506.02166.
12. Toma, A.; Broniatowski, M. Dual divergence estimators and tests: Robustness results. *J. Multivar. Anal.* **2011**, *102*, 20–36.
13. Rockafellar, R.T.; Wets, R.J.B. *Variational Analysis*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 1998.
14. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and Efficient Estimation by Minimising a Density Power Divergence. *Biometrika* **1998**, *85*, 549–559.
15. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **1977**, *39*, 1–38.
16. Wu, C.F.J. On the Convergence Properties of the EM Algorithm. *Ann. Stat.* **1983**, *11*, 95–103.
17. Ostrowski, A. *Solution of Equations and Systems of Equations*; Academic Press: Cambridge, MA, USA, 1966.
18. Chrétien, S.; Hero, A.O. On EM algorithms and their proximal generalizations. *ESAIM Probabil. Stat.* **2008**, *12*, 308–326.
19. Berge, C. *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces, and Convexity*; Dover Publications: Mineola, NY, USA, 1963.
20. Meister, A. *Deconvolution Problems in Nonparametric Statistics*; Springer: Berlin/Heidelberg, Germany, 2009.
21. Jiménz, R.; Shao, Y. On robustness and efficiency of minimum divergence estimators. *Test* **2001**, *10*, 241–248.
22. Nelder, J.A.; Mead, R. A Simplex Method for Function Minimization. *Comput. J.* **1965**, *7*, 308–313.
23. The R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.