A two-step proximal-point algorithm for the calculus of divergence-based estimators in finite mixture models

Michel Bronialowski, Diaa Al Mohamad

The Canadian Journal of Statistics
Vol. 47, No. 3, 2019, Pages 392–408
La revue canadienne de statistique

A two-step proximal-point algorithm for the calculus of divergence-based estimators in finite mixture models

Diaa AL MOHAMAD[1*] and Michel BRONIATOWSKI[2]

[1]*Department of Biomedical Data Sciences, Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands*
[2]*Laboratoire de Probabilités, Statistique et Modélisation, University of Pierre and Marie Curie (Sorbonne University), 4 place Jussieu, 75252 Paris CEDEX 05, France*

Key words and phrases: EM algorithm; mixture model; proximal-point algorithm; robustness; statistical divergence.

MSC 2010: Primary 62F35; secondary 62F10, 65K10, 65C60.

Abstract: Estimators derived from the expectation-maximization (EM) algorithm are not robust since they are based on the maximization of the likelihood function. We propose an iterative proximal-point algorithm based on the EM algorithm to minimize a divergence criterion between a mixture model and the unknown distribution that generates the data. The algorithm estimates in each iteration the proportions and the parameters of the mixture components in two separate steps. Resulting estimators are generally robust against outliers and misspecification of the model. Convergence properties of our algorithm are studied. The convergence of the introduced algorithm is discussed on a two-component Weibull mixture entailing a condition on the initialization of the EM algorithm in order for the latter to converge. Simulations on Gaussian and Weibull mixture models using different statistical divergences are provided to confirm the validity of our work and the robustness of the resulting estimators against outliers in comparison to the EM algorithm. An application to a dataset of velocities of galaxies is also presented. *The Canadian Journal of Statistics* 47: 392–408; 2019 © 2019 Statistical Society of Canada

Résumé: Les estimateurs obtenus par l'algorithme EM ne sont pas robustes, car ils sont basés sur la maximisation de la vraisemblance. Les auteurs proposent un algorithme itératif de type proximal fondé sur l'algorithme EM et qui vise à minimiser une divergence statistique entre un modèle de mélange et la distribution inconnue des données. À chaque itération, l'algorithme estime en deux étapes distinctes les proportions et les paramètres décrivant les composantes du mélange. Les estimateurs obtenus sont généralement robustes contre les points aberrants et le mauvais choix du modèle. Les auteurs étudient les propriétés de convergence de leur algorithme. Ils illustrent ces propriétés sur un exemple de mélange de Weibull à deux composantes et déterminent une condition sur l'initialisation de l'algorithme EM pour ce modèle pour qu'il converge. Ils illustrent également la convergence et la robustesse de leur algorithme sur deux mélanges à deux composantes issus des lois gaussienne et de Weibull. Ils l'appliquent enfin sur un jeu de données réelles de vitesses de galaxies. *La revue canadienne de statistique* 47: 392–408; 2019 © 2019 Société statistique du Canada

Additional Supporting Information may be found in the online version of this article at the publisher's website.

*Author to whom correspondence may be addressed.
E-mail: diaa.almohamad@gmail.com

1. INTRODUCTION

The expectation-maximization (EM) algorithm (Dempster, Laird & Rubin, 1977) is a well-known method for calculating the maximum likelihood estimator (MLE) of a model where incomplete data are 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 are available; see McLachlan & Krishnan (2007). Another way to look at the EM algorithm is as a proximal-point problem; see Chrétien & Hero (1998) and Tseng (2004). Indeed, we may rewrite the conditional expectation of the complete log-likelihood as the log-likelihood function of the model (the objective) plus a proximal term. Generally, the proximal term has a regularization effect on the objective function so that the algorithm becomes more stable, could avoid some saddle points and frequently outperforms classical optimization algorithms; see Goldstein & Russak (1987) and Chrétien & Hero (2008). Chrétien & Hero (1998) prove 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 which we present below, we generalize the EM algorithm (and the version in Tseng, 2004) by replacing the log-likelihood function with an estimator of a statistical divergence between the true distribution of the data and the model. We are particularly interested in φ-divergences and the density power divergence (DPD) which is a Bregman divergence. The DPD introduced and studied by Basu et al. (1998) is defined for $a > 0$ as

$$D_a(g, f) = \int_{\mathbb{R}} \left\{ f^{1+a}(y) - \frac{a+1}{a} g(y) f^a(y) + \frac{1}{a} g^{1+a}(y) \right\} dy, \tag{1}$$

for two probability density functions f and g. Given a random sample $Y_1, \ldots, Y_n$ distributed according to some probability measure P_T with density p_T with respect to Lebesgue measure, and given a model $(p_\phi)_{\phi \in \Phi}$ with $\Phi \subset \mathbb{R}^d$, the minimum density power divergence (MDPD) estimator is defined by:

$$\hat{\phi}_n = \arg\min_{\phi \in \Phi} \left\{ \int_{\mathbb{R}} p_\phi^{1+a}(z)\, dz - \frac{a+1}{a} \frac{1}{n} \sum_i^n p_\phi^a(Y_i) \right\}. \tag{2}$$

This estimator is robust for $a > 0$, and when a goes to zero, we obtain the MLE.

A φ-divergence in the sense of Csiszár (Csiszár, 1963; Broniatowski & Keziou, 2009) is defined by:

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

where φ is a nonnegative strictly convex function with $\varphi(1) = 0$, and Q and P are two probability measures such that Q is absolutely continuous with respect to P. Examples (among others) of such divergences are: the Kullback–Leibler (KL) divergence when $\varphi_1(t) = t \log t + 1 - t$, the modified KL divergence when $\varphi_0(t) = -\log t + t - 1$, and the Hellinger distance when $\varphi_{0.5}(t) = (\sqrt{t} - 1)^2$. All these well-known divergences belong to the family of φ_γ-divergences generated by the class of Cressie–Read functions defined by:

$$\varphi_\gamma(t) = \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)}, \quad \gamma \in \mathbb{R} \setminus \{0, 1\}, \tag{3}$$

and defining φ_0 and φ_1 as the limit as γ tends to 0 and 1, respectively. We consider the dual estimator of the φ-divergence (DφDE) introduced independently by Broniatowski & Keziou (2006) and Liese & Vajda (2006). The use of this estimator is motivated by many reasons. Its minimum coincides with the MLE for $\varphi(t) = -\log t + t - 1$. Besides, it does not take into account any partitioning or smoothing and has the same form for discrete and continuous models, which is not the case for other estimators considered by Beran (1977), Park & Basu (2004) and Basu & Lindsay (1994), who use kernel density estimators. For ϕ in Φ, the DφDE is given by:

$$\hat{D}_{\varphi}(\phi) = \sup_{\alpha\in\Phi} \left\{ \int_{\mathbb{R}} \varphi' \left(\frac{p_{\phi}(x)}{p_{\alpha}(x)} \right) p_{\phi}(x)\, dx - \frac{1}{n} \sum_{i=1}^{n} \varphi^{\#} \left(\frac{p_{\phi}(Y_i)}{p_{\alpha}(Y_i)} \right) \right\}, \tag{4}$$

with $\varphi^{\#}(t) = t\varphi'(t) - \varphi(t)$. Al Mohamad (2018) argues that while this formula works well under the model, it underestimates the divergence between the true distribution and the model under misspecification of the model or contamination in the data, and proposes the following simpler estimator:

$$\tilde{D}_{\varphi}(\phi) = \int_{\mathbb{R}} \varphi' \left(\frac{p_{\phi}(x)}{K_{n,w}(x)} \right) p_{\phi}(x)\, dx - \frac{1}{n} \sum_{i=1}^{n} \varphi^{\#} \left(\frac{p_{\phi}(Y_i)}{K_{n,w}(Y_i)} \right), \tag{5}$$

where $K_{n,w}$ is a nonparametric estimator of the true distribution P_T. In this paper, $K_{n,w}$ is a kernel density estimator. The resulting new estimator is robust against outliers. It also permits getting rid of the supremal form from the dual estimator (4). The minimum dual φ-divergence estimator (MDφDE) is defined by:

$$\text{Classical MD}\varphi\text{DE} = \arg\min_{\phi\in\Phi} \hat{D}_{\varphi}(\phi),$$

$$\text{Kernel-based MD}\varphi\text{DE} = \arg\min_{\phi\in\Phi} \tilde{D}_{\varphi}(\phi).$$

Asymptotic properties and consistency of these two estimators can be found in Broniatowski & Keziou (2009), Toma & Broniatowski (2011) and Al Mohamad (2018).

We propose to calculate the two MDφDEs and the MDPD when p_{ϕ} is a mixture model using an iterative procedure based on the work of Tseng (2004) on the log-likelihood function. This procedure has the form of a proximal-point algorithm and extends the EM algorithm. A similar algorithm was introduced in Al Mohamad & Broniatowski (2015, 2016). Here, in each iteration we have two steps: a step to calculate the proportion and a step to calculate the parameters of the mixture components. The goal of this simplification is to reduce the dimension over which we optimize, since in lower dimensions optimization procedures are generally more efficient. Our convergence proof requires some regularity of the estimated divergence with respect to the parameter vector which is not easily checked using Equation (4). Results in Rockafellar & Wets (1998) provide sufficient conditions to solve this problem. Differentiability of $\hat{D}_{\varphi}(\phi)$ with respect to ϕ may remain a very hard task in many situations.

The paper is organized as follows. We explain in Section 2 the context and indicate the mathematical notation which may be nonstandard. We also present the progression and the derivation of our set of algorithms from the EM algorithm and Tseng's generalization. In Section 3, we prove some convergence properties of the sequence generated by our algorithm. We show in Section 4 a case study of a Weibull mixture including a convergence proof of the EM algorithm. Finally, Section 5 provides simulations confirming the robustness of the resulting estimator in comparison to the EM algorithm. The proofs of the main results are in the Appendix.

2. A DESCRIPTION OF THE ALGORITHM

2.1. General Context and Notation

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be n realizations drawn from the joint probability density function $f(x, y|\phi)$ parameterized by a real vector $\phi \in \Phi \subset \mathbb{R}^d$. The x_i's are the unobserved data (labels) and the y_i's are the observations. 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 by dx a measure defined on the label space $\mathcal{X}$ (it is counting measure if $\mathcal{X}$ is discrete). The marginal density of the observed data is given by $p_\phi(y) = \int_{\mathbb{R}} f(x, y|\phi)\, dx$, which is assumed to be a finite mixture model of the form

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

where $s > 1, \phi = (\lambda_1, \ldots, \lambda_{s-1}, \theta)$ and $\forall i, \lambda_i \in (0, 1)$ such that $\lambda_s = 1 - \sum_{i=1}^{s-1} \lambda_i$. For a parameterized function f with a parameter a, we write $f(x|a)$. We use the notation ϕ^k for sequences with the index k. For a set Φ, $\operatorname{Int}\Phi$ denotes its interior.

2.2. EM Algorithm for Mixture Models

Let $h_i(x|\phi^k)$ be the conditional density of the labels (at step k) given y_i:

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

Let ϕ^0 be an initial vector. Let $\psi(t) = -\log t$ (or $\psi(t) = -\log t + t - 1$). The EM algorithm estimates the unknown parameter vector by generating the sequence:

$$\begin{aligned}\phi^{k+1} &= \arg\max_{\phi} \left\{ \sum_{i=1}^{n} \log\left(p_\phi(y_i)\right) + \sum_{i=1}^{n} \int_{\mathcal{X}} \psi\left(\frac{h_i(x|\phi)}{h_i(x|\phi^k)}\right) h_i(x|\phi^k)\, dx \right\} \\ &= \arg\max_{\phi} \left\{ J(\phi) - D_\psi(\phi, \phi^k) \right\}. \end{aligned} \tag{6}$$

The formulation (6) is a proximal-point algorithm which was proposed by Tseng (2004) who studied the convergence properties for any convex nonnegative function ψ. When p_ϕ is a finite mixture model, the EM algorithm has the two-step form

$$\left(\lambda_1^{k+1}, \ldots, \lambda_{s-1}^{k+1}\right) = \arg\max_{\lambda_1 \geq 0, \ldots, \lambda_{s-1} \geq 0} \sum_{i=1}^{n} \sum_{j=1}^{s-1} \log(\lambda_j) h_i(j|\phi^k) + \sum_{i=1}^{n} \log\left(1 - \sum_{j=1}^{s-1} \lambda_j\right) h_i(s|\phi^k), \tag{7}$$

$$\left(\theta_1^{k+1}, \ldots, \theta_s^{k+1}\right) = \arg\max_{(\theta_1, \ldots \theta_s) \in \Theta} \sum_{i=1}^{n} \sum_{j=1}^{s-1} \log\left(\lambda_j p(y_i|\theta_j)\right) h_i(j|\phi^k) + \sum_{i=1}^{n} \log\left(p(y_i|\theta_j)\right) h_i(s|\phi^k). \tag{8}$$

The EM algorithm and its generalization (6) for any convex nonnegative function ψ produce estimators based on the likelihood function, which are not robust against outliers or misspecification of the model. Estimators calculated using statistical divergences such as the Hellinger or the chi-squared are known to be robust. Let $D(\lambda, \theta)$ be some statistical divergence calculated between the model and the true distribution of the data, and let $\hat{D}(\lambda, \theta)$ be its estimator. We propose to estimate the mixture model through the sequences

$$\lambda^{k+1} = \underset{\lambda\in[0,1]^s,\text{s.t.}(\lambda,\theta^k)\in\Phi}{\arg\min} \left\{\hat{D}(\lambda,\theta^k) + D_\psi((\lambda,\theta^k),(\lambda^k,\theta^k))\right\}, \tag{9}$$

$$\theta^{k+1} = \underset{\theta\in\Theta,\text{s.t.}(\lambda^{k+1},\theta)\in\Phi}{\arg\min} \left\{\hat{D}(\lambda^{k+1},\theta) + D_\psi((\lambda^{k+1},\theta),(\lambda^k,\theta^k))\right\}, \tag{10}$$

where D_ψ is as defined in (6). Examples of statistical divergences include φ-divergences (Broniatowski & Keziou, 2009), DPDs (Basu et al., 1998), S-divergences (Gosh et al., 2013) and Rényi pseudodistances (e.g., Toma & Leoni-Aubin, 2013). They all include the MLE for a suitable choice of the tuning parameter or the generating function so that the sequence (9) and (10) coincides with the sequence (7) and (8).

Our two-step algorithm in Equations (9) and (10) coincides with the one-step proximal-point algorithm introduced by Al Mohamad & Broniatowski (2015) for general models if we omit the optimization over the proportions. In other words, the one-step proximal-point algorithm is given by

$$\phi^{k+1} = \underset{\phi}{\arg\min} \left\{\hat{D}(\phi) + D_\psi(\phi,\phi^k)\right\}. \tag{11}$$

The remainder of the paper is devoted entirely to the study of the convergence of the sequences generated by the set of algorithms (9) and (10).

3. SOME CONVERGENCE PROPERTIES OF ϕ^K

We adapt the ideas given by Tseng (2004) to develop proofs of convergence for our proximal algorithm as k goes to infinity while n is held fixed. The proofs are deferred to the Appendix. Let $\phi^0 = (\lambda^0, \theta^0)$ be a given initialization for the parameters, and define the set

$$\Phi^0 = \{\phi = (\lambda,\theta) \in \Phi : \hat{D}(\phi) \le \hat{D}(\phi^0)\}.$$

We suppose that Φ^0 is a subset of $\operatorname{Int}\Phi$. The idea of defining such a set in this context is inherited from Wu (1983). We use the set of assumptions A0–A4 provided in Appendix. They are verifiable using Lebesgue theorems and the approaches provided in the Supplementary Material.

Proposition 1. *Assume that recurrences (9) and (10) are well defined in* Φ. *For both algorithms, the sequence* $(\phi^k)_k$ *satisfies the following properties:*

(a) $\hat{D}(\phi^{k+1}) \le \hat{D}(\phi^k)$.
(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* $\left\{\hat{D}(\phi^k)\right\}_k$ *converges as k goes to infinity.*

The interest of Proposition 1 is that the objective function is ensured, under mild assumptions, to decrease alongside the sequence $(\phi^k)_k$. This permits to build a stopping 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 neighbourhood of an optimum. The following result provides a first characterization of the properties of the limit of the sequence $(\phi^k)_k$ as a stationary point of the estimated divergence.

Proposition 2. *Suppose that A1 is true, and assume that* Φ^0 *is closed and* $\{\phi^{k+1} - \phi^k\} \to 0$ *as k goes to infinity. If A4 is satisfied, then the limit of every convergent subsequence is a stationary point of* $\phi \mapsto \hat{D}(\phi)$.

Proposition 3. *Assume that A1, A2 and A3 hold, then* $\{\phi^{k+1} - \phi^k\} \to 0$ *as k goes to infinity, which implies, by Proposition 2, that any limit point of the sequence* ϕ^k *is a stationary point of* $\phi \mapsto \hat{D}(\phi)$.

We can go further in exploring the properties of the sequence $(\phi^k)_k$ by imposing additional assumptions. The following corollary provides a convergence result of the whole sequence. The convergence holds also towards a local minimum as long as the estimated divergence is locally strictly convex.

Corollary 1. *Under the assumptions of Proposition 3, the set of accumulation points of* $(\phi^k)_k$ *defined by (9) and (10) is a connected compact set. Moreover, if* $\hat{D}(\phi)$ *is strictly convex in a neighbourhood of a limit point of the sequence* $(\phi^k)_k$*, then the whole sequence* $(\phi^k)_k$ *converges to a local minimum of* $\hat{D}(\phi)$ *as k goes to infinity.*

Although Proposition 3 provides a general solution to assess that $\{\phi^{k+1} - \phi^k\} \to 0$ as k goes to infinity, the identifiability assumption over the proximal term is hard to be fulfilled. It does not hold in most simple mixtures such as a two component Gaussian mixture (Tseng, 2004). This is the reason behind our next result. A similar idea is employed by Chrétien & Hero (2008). Their work however requires that the log-likelihood approaches $-\infty$ as $\|\phi\| \to \infty$, which is not satisfied by usual mixture models (e.g., the Gaussian mixture model). Our result treats the problem from another perspective using the set Φ^0.

Proposition 4. *Assume that A1 and A2 hold. For the algorithm defined by (9) and (10), if* $\|\theta^{k+1} - \theta^k\| \to 0$ *as k goes to infinity, then any convergent subsequence* $\{\lambda^{N(k)}, \theta^{N(k)}\}_k$ *converges to a stationary point of the objective function* $(\lambda, \theta) \to \hat{D}(\lambda, \theta)$ *as k goes to infinity.*

Proposition 4 requires a condition on the distance between two consecutive members of the sequence $\{\theta^k\}_k$ which is weaker than the same condition on the whole sequence $\phi^k = (\lambda^k, \theta^k)$. Still, as the regularization term D_ψ does not satisfy the identifiability condition A3, it remains an open problem for further work. It is interesting to notice that the 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 (9) and (10).

Following Chrétien & Hero (2008), we can define a proximal-point algorithm which converges to a global infimum. Let $\{\beta_k\}_k$ be a sequence of positive numbers which decreases to zero ($\beta_k = 1/k$ does the job). Define

$$\lambda^{k+1} = \underset{\lambda \in [0,1]^s,\ \text{s.t. } (\lambda, \theta^k) \in \Phi}{\arg\min} \left\{ \hat{D}(\lambda, \theta^k) + \beta_k D_\psi((\lambda, \theta^k), (\lambda^k, \theta^k)) \right\},$$

$$\theta^{k+1} = \underset{\theta \in \Theta,\ \text{s.t. } (\lambda^{k+1}, \theta) \in \Phi}{\arg\min} \left\{ \hat{D}(\lambda^{k+1}, \theta) + \beta_k D_\psi((\lambda^{k+1}, \theta), (\lambda^k, \theta^k)) \right\}.$$

The justification of such a variant falls directly from Theorem 3.2.4 of Chrétien & Hero (2008). The problem with this approach is that the infimum on each step of the algorithm needs to be calculated exactly, which does not happen in general unless the function $\phi \mapsto \hat{D}(\phi) + \beta_k D_\psi(\phi, \phi^k)$ is strictly convex.

4. CASE STUDY: A TWO-COMPONENT WEIBULL MIXTURE

Let p_ϕ be the two-component Weibull mixture

$$p_\phi(x) = 2\lambda\phi_1(2x)^{\phi_1 - 1} e^{-(2x)^{\phi_1}} + (1 - \lambda)\frac{\phi_2}{2}\left(\frac{x}{2}\right)^{\phi_2 - 1} e^{-(x/2)^{\phi_2}}, \tag{12}$$

where $\phi = (\lambda, \phi_1, \phi_2)$. We have $\Phi = [\eta, 1-\eta] \times \mathbb{R}_+^* \times \mathbb{R}_+^*$ for some $\eta > 0$ in order to avoid degeneracy. We will be interested only in power divergences defined through the Cressie–Read class of functions $\varphi = \varphi_\gamma$ given by (3). Functions h_i are given by

$$h_i(1|\phi) = \frac{2\lambda\phi_1(2x)^{\phi_1-1}e^{-(2x)^{\phi_1}}}{2\lambda\phi_1(2x)^{\phi_1-1}e^{-(2x)^{\phi_1}} + (1-\lambda)\phi_2\,(1/2)^{\phi_2-1}\,e^{-(x/2)^{\phi_2}/2}},$$

$$h_i(2|\phi) = 1 - h_i(1|\phi).$$

It is clear that functions h_i are in class $C^1(\mathrm{Int}\,\Phi)$ and so is $\phi \mapsto D_\psi(\phi, \phi')$ for any $\phi' \in \Phi$.

Use the DPD defined by Equation (1): If we use the DPD (1), the continuity and differentiability of the estimated divergence $\hat{D}_a$ (the optimized function in Equation (2)) can be treated using Lebesgue theorems. To prove that Φ^0 is compact, we prove that it is closed and bounded in the complete space $[\eta, 1-\eta] \times \mathbb{R}_+^2$. We add zero to the values of the shape parameter so that the space becomes complete. Closedness is an immediate result of the continuity of the estimated divergence since Φ^0 is the inverse image of the closed set $(-\infty, \hat{D}_a(\phi^0)]$. To ensure boundedness of Φ^0, we need to choose carefully the initial point $(\lambda^0, \phi_1^0, \phi_2^0)$ of the algorithm. Since λ is bounded by 0 and 1, we only need to verify the boundedness of the shapes. If both shapes ϕ_1 and ϕ_2 go to $\pm\infty$, then $\hat{D}_a(\phi) \to 0$. If either of the shapes goes to infinity, then the corresponding component vanishes. In order to prevent the shapes from growing to infinity, we start at a point where the estimated divergence is lower than those extremities. Then, because of the decreasing property of the algorithm and the definition of Φ^0, the algorithm never goes back to any of the unbounded situations. We thus identify a condition on the initialization of the algorithm in order to make Φ^0 bounded:

$$\hat{D}_a(p_{\lambda,\phi}) < \min\left\{0, \inf_{\phi_1>0, \lambda\in[\eta,1-\eta]} \hat{D}_a(p_{(\lambda,\phi_1,\infty)})\right\}. \tag{13}$$

Conclusion 1. *Using Proposition 1 and under condition (13), the sequence* $\{\hat{D}_a(\phi^k)\}_k$ *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* $\hat{D}_a$.

Use the dual estimator defined by Equation (4): If we use (4) to determine the estimator, then only continuity of the estimated divergence with respect to the parameters can be obtained. Write $\hat{D}_\varphi(\phi) = \sup_\alpha f(\alpha, \phi)$. We list the following 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 $\gamma \in (0, 1)$, the function $f(\alpha, \phi)$ is continuous.
3. For $\gamma < 0$, the function $f(\alpha, \phi)$ is continuous and well defined for $\phi_1 < \alpha_1\gamma^{-1}(\gamma - 1)$ 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, if Φ is compact, then using Theorem 1.17 of Rockafellar & Wets (1998), $\phi \mapsto \hat{D}_\varphi(\phi)$ is continuous. Differentiability, however, is difficult to prove and requires more investigation on the form of the estimated divergence and the model used. We conclude that if Φ is compact, then Proposition 1 can be used to deduce that the sequence $\hat{D}_\varphi(\phi^k)$ converges, but no information about the convergence towards stationary points could be obtained.

Use the kernel-based dual estimator given by Equation (5): If we use (5) to define the estimator, then the continuity of $\hat{D}_{\varphi}(\phi)$ depends on the tail of the kernel to be used and the value of γ. For example, if we use a Gaussian kernel and for $\gamma \in (0, 1)$, then the estimated divergence is $C^1(\mathrm{Int}\,\Phi)$. A similar condition to (13) can be obtained and we have the same conclusion as Conclusion 1.

Use the likelihood of the model: If we use $\varphi(t) = \phi(t) = -\log t + t - 1$, we obtain the EM algorithm. Assumptions A1 and A4 are clearly satisfied. Let $L(\phi)$ be the likelihood function, and $J(\phi) = \log L(\phi)$. The set Φ^0 is given by

$$\Phi^0 = \left\{\phi \in \Phi : J(\phi) \geq J(\phi^0)\right\}.$$

We will show that under suitable conditions, the set Φ^0 is compact. Suppose that the shape parameter can have values in $\mathbb{R}_+$. The set Φ^0 becomes the inverse image of $[L(\phi^0), \infty)$ by the likelihood function which is continuous, and thus Φ^0 is closed in the space $[\eta, 1-\eta] \times \mathbb{R}_+ \times \mathbb{R}_+$. Similarly to the previous cases, in order to prove boundedness we need to avoid the cases where either of the shape parameters tends to infinity. For example, when ϕ_1 goes to infinity,

$$J(\lambda, \infty, \phi_2) = \sum_{i=1}^{n} \log\left((1-\lambda)\frac{\phi_2}{2}\left(\frac{y_i}{2}\right)^{\phi_2 - 1} e^{-(y_i/2)^{\phi_2}}\right)$$

which is bounded almost everywhere. We then choose the initial point of the algorithm ϕ^0 in such a way that

$$J(\phi^0) > \max\left\{\sup_{\lambda,\phi_2} J(\lambda, \infty, \phi_2), \sup_{\lambda,\phi_1} J(\lambda, \phi_1, \infty)\right\},$$

and the set Φ^0 hence becomes bounded and therefore compact. The same conclusion as Conclusion 1 holds for the Weibull mixture model.

Note that the verification of assumption A3 is a hard task, because it results in a set of n nonlinear equations in y_i and cannot be treated in a similar way to the Gaussian mixture in Tseng (2004) or Al Mohamad & Broniatowski (2016).

5. EXPERIMENTAL RESULTS

We summarize the results of 100 experiments on 100 samples (with and without outliers) from two-component Gaussian and Weibull mixtures. We measure the error of replacing the true distribution of the data with the model using the total variation distance (TVD) which is calculated using the L^1 distance by the Scheffé lemma (e.g., Meister, 2009, p. 129).

$$\mathrm{TVD}(\phi) = \sup_{a<b}\left|dP_{\phi}([a,b]) - dP_T([a,b])\right| = \frac{1}{2}\int_{\mathbb{R}} |p_{\phi}(x) - p_T(x)|\,dx.$$

We also provide for the Gaussian mixture the values of the square root of the χ^2 divergence between the estimated model and the true mixture (it gives infinite values for the Weibull experiment because of the sensitivity of the χ^2 to differences in the tail of the distribution). The χ^2 criterion is defined by:

$$\chi^2(\phi) = \int_{\mathbb{R}} \frac{\left\{p_{\phi}(x) - p_T(x)\right\}^2}{p_T(x)}dx.$$

We also apply our algorithms to a dataset of the velocities of galaxies where only estimates of the parameters are provided.

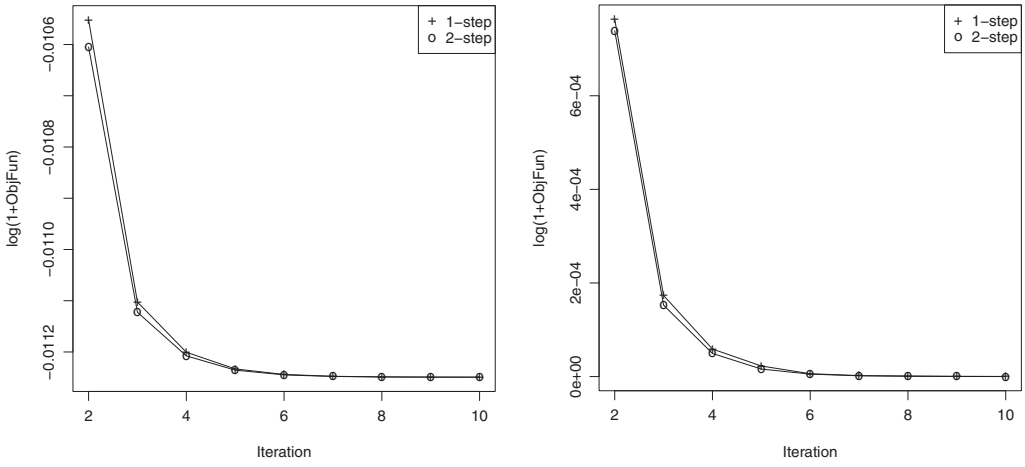

FIGURE 1: Decrease of the estimated Hellinger divergence in the Gaussian mixture. The figure to the left corresponds to estimator (5). The figure to the right corresponds to estimator (4).

We use different φ-divergences, namely the Hellinger, the Pearson and the Neyman χ^2. For the MDPD, we use $a = 0.5$, a choice which gave the best tradeoff between robustness and efficiency suggested by the simulation results of Al Mohamad (2018). For the proximal term, we use $\psi(t) = (\sqrt{t} - 1)^2$. The methods are compared with the EM algorithm. All the experiments are carried out using the statistical tool R (R Core Team, 2013).

5.1. A Two-Component Gaussian Mixture

We consider a Gaussian mixture with true parameters $\lambda = 0.35, \mu_1 = 2, \mu_2 = 1.5$ and fixed variances $\sigma_1^2 = \sigma_2^2 = 1$. Figure 1 shows the values of the estimated divergence for both formulas (4) and (5) on a logarithmic scale at each iteration of the algorithms (11), and (9) and (10) until convergence. The one-step algorithm refers to algorithm (11), whereas the two-step algorithm refers to algorithms (9) and (10). The results are presented in Table 1.

Contamination is done by adding in the original sample to the 5 lowest values random observations from the uniform distribution $\mathcal{U}[-5, -2]$. We also add to the 5 largest values random observations from the uniform distribution $\mathcal{U}[2, 5]$. Results are presented in Table 2. It is clear that both the MDPD and the kernel-based MDφDE are more robust than the EM algorithm and the classical MDφDE.

5.2. The Two-Component Weibull Mixture Model Revisited

We consider the Weibull mixture (12) with $\phi_1 = 0.5, \phi_2 = 3$ and $\lambda = 0.35$. We let $\phi = (\phi_1, \phi_2)$ denote the shape parameters of the Weibull mixture model $p_{(\lambda,\phi)}$, and $\alpha = (\alpha_1, \alpha_2)$ for $p_{(\lambda,\alpha)}$. Contamination is done by replacing 10 observations of each sample chosen randomly by 10 i.i.d. observations drawn from a Weibull distribution with shape 0.9 and scale 3. Results are presented in Table 3.

When there are no outliers, all estimation methods have the same performance. The results show a clear robustness of the MDPD and the kernel-based MDφDE with the Hellinger divergence in comparison to the other estimators. Using the Neymann χ^2 divergence, the classical MDφDE (4) shows better robustness than the kernel-based MDφDE. 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 2 in Al Mohamad (2018).

TABLE 1: Estimation error for the two-component Gaussian mixture when there are no outliers.

Divergence	Algorithm	Estimator	$\sqrt{\chi^2}$	sd$\left(\sqrt{\chi^2}\right)$	TVD	sd(TVD)
Chi-squared	Algorithm (11)	MDφDE	0.108	0.052	0.061	0.029
		Kernel MDφDE	0.118	0.052	0.066	0.027
	Algorithms (9) and (10)	MDφDE	0.108	0.052	0.061	0.029
		Kernel MDφDE	0.118	0.051	0.066	0.027
Hellinger	Algorithm (11)	MDφDE	0.108	0.052	0.050	0.025
		Kernel MDφDE	0.113	0.044	0.064	0.025
	Algorithms (9) and (10)	MDφDE	0.108	0.052	0.061	0.029
		Kernel MDφDE	0.113	0.045	0.064	0.025
DPD($a = 0.5$)	Algorithm (11)	MDPD	0.117	0.049	0.065	0.025
DPD($a = 0.5$)	Algorithms (9) and (10)	MDPD	0.117	0.047	0.065	0.025
Log-likelihood	EM	MLE	0.113	0.044	0.064	0.025

TABLE 2: Estimation error for the two-component Gaussian mixture in the presence of 10% outliers.

Divergence	Algorithm	Estimator	$\sqrt{\chi^2}$	sd$\left(\sqrt{\chi^2}\right)$	TVD	sd(TVD)
Chi-squared	Algorithm (11)	MDφDE	0.334	0.097	0.146	0.036
		Kernel MDφDE	0.149	0.059	0.084	0.033
	Algorithms (9) and (10)	MDφDE	0.333	0.097	0.149	0.033
		Kernel MDφDE	0.149	0.059	0.084	0.033
Hellinger	Algorithm (11)	MDφDE	0.321	0.096	0.146	0.034
		Kernel MDφDE	0.155	0.059	0.087	0.033
	Algorithms (9) and (10)	MDφDE	0.322	0.097	0.147	0.034
		Kernel MDφDE	0.156	0.059	0.087	0.033
DPD($a = 0.5$)	Algorithm (11)	MDPD	0.129	0.049	0.065	0.025
DPD($a = 0.5$)	Algorithms (9) and (10)	MDPD	0.138	0.053	0.078	0.030
Log-likelihood	EM	MLE	0.335	0.102	0.150	0.034

We see no significant difference between the results obtained using the one-step algorithm (11) and those obtained using the two-step algorithms (9) and (10) using the Hellinger divergence. Differences appear when we use the Neymann χ^2-divergence with the classical MDφDE. This shows again the difficulty in handling the supremal form of the dual formula (4).

5.3. The Galaxies Dataset

We study a dataset of velocities of 82 galaxies in the Corona Borealis region at which they move away from our galaxy (Figure 2). The dataset is available from the R package `MASS`. The objective of the study is to figure out if the distribution of the velocities is multimodal or if

TABLE 3: Estimation error for the two-component Weibull mixture with and without outliers.

			No outliers		10% outliers	
Divergence	Algorithm	Estimator	TVD	sd(TVD)	TVD	sd(TVD)
Neymann Chi-squared	Algorithm (11)	MDφDE	0.114	0.032	0.085	0.036
		Kernel MDφDE	0.057	0.028	0.138	0.066
	Algorithms (9) and (10)	MDφDE	0.131	0.042	0.096	0.057
		Kernel MDφDE	0.056	0.026	0.127	0.056
Hellinger	Algorithm (11)	MDφDE	0.059	0.024	0.120	0.034
		Kernel MDφDE	0.057	0.029	0.068	0.034
	Algorithms (9) and (10)	MDφDE	0.061	0.026	0.121	0.034
		Kernel MDφDE	0.057	0.029	0.068	0.034
DPD($a = 0.5$)	Algorithm (11)	MDPD	0.056	0.029	0.060	0.029
DPD($a = 0.5$)	Algorithms (9) and (10)	MDPD	0.056	0.029	0.061	0.029
Log-likelihood	EM	MLE	0.059	0.024	0.129	0.046

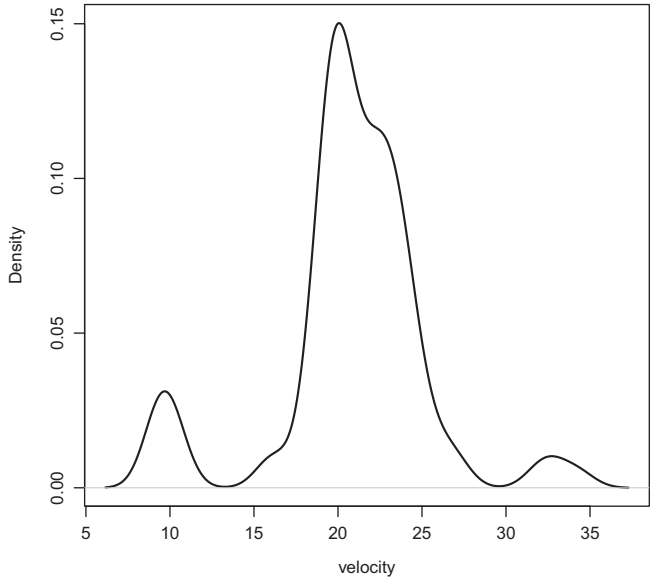

FIGURE 2: Density estimation of galaxy velocities (in 1,000 km s^{-1}) distribution.

there are superclusters in these galaxies. More details about this dataset can be found in Roeder (1990). Roeder (1990) estimates the number of clusters to be between three and seven modes and a test of unimodality is rejected at level 0.01. Using the R package `mclust`, we find that a mixture with four components best fits the data according to the BIC criterion. Therefore, we fix the number of components at four and assume that all components have the same variance to avoid degeneracy. We estimate a mixture of four Gaussian components using our algorithms. For φ-divergences, we use $\varphi(t) = \varphi_{0.5}$ which corresponds to the Hellinger divergence. For the MDPD, we use $a = 0.5$. The results are provided in Table 4.

TABLE 4: Estimation of Gaussian mixture model with four components for the Galaxy velocities.

Divergence	Algorithm	Estimator	λ_1	λ_2	λ_3	μ_1	μ_2	μ_3	μ_4	σ
Hellinger	Algorithm (11)	Kernel MDφDE	0.08	0.69	0.20	10.10	20.70	24.40	33.90	1.50
		MDφDE	0.09	0.83	0.05	10.10	21.10	26.30	33.80	1.80
	Algorithms (9) and (10)	Kernel MDφDE	0.09	0.51	0.37	9.70	20.00	23.30	33.00	1.00
		MDφDE	0.09	0.64	0.23	10.10	20.50	24.20	33.90	1.40
DPD($a = 0.5$)	Algorithm (11)	MDPD	0.11	0.40	0.43	10.20	19.40	23.70	34.30	1.50
DPD($a = 0.5$)	Algorithms (9) and (10)	MDPD	0.02	0.53	0.44	10.23	19.90	22.90	32.70	0.66
Log-likelihood	EM	MLE	0.09	0.53	0.35	9.70	20.00	23.50	33.00	1.70

The results obtained with the one-step algorithm are a little bit different from the results obtained with the two-step algorithm for the estimates of the proportions and the variance. The difference is almost negligible in the centres of the clusters. On the other hand, all the results support that there is a cluster with high proportion at velocity around 20×10^3 km s^{-1}.

6. CONCLUSIONS

We presented in this paper a two-step proximal-point algorithm whose objective was the minimization of (an estimate of) a statistical divergence for a mixture model. The EM algorithm constituted a special case. We established some convergence properties of the algorithm under mild conditions. Our simulation results showed that the proximal algorithm worked. The two-step algorithms (9) and (10) showed no difference from its one-step competitor (11) which was very encouraging, especially since the dimension of the optimization was reduced at each step in the two-step algorithm. Simulations showed again the robustness of φ-divergences and the DPD against outliers in comparison to the MLE calculated from the EM algorithm. The role of the proximal term and its influence on the convergence of the algorithm were not discussed here and might be considered in future work.

APPENDIX

The following are necessary assumptions required for the convergence results.

A0. Functions $\phi \mapsto \hat{D}(\phi)$ and $\phi \mapsto D_\psi(\phi, v)$ are lower semicontinuous on Φ for all v in Φ.

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

A2. Φ^0 is a compact subset of $\operatorname{Int} \Phi$.

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

A4. $\phi \mapsto \nabla \hat{D}(\phi)$ is defined and continuous on Φ.

Proof of Proposition 1.

(a) Recurrence (9) and the definition of the argmin give:

$$\begin{aligned} \hat{D}(\lambda^{k+1}, \theta^k) + D_\psi((\lambda^{k+1}, \theta^k), (\lambda^k, \theta^k)) &\leq \hat{D}(\lambda^k, \theta^k) + D_\psi((\lambda^k, \theta^k), (\lambda^k, \theta^k)) \\ &\leq \hat{D}(\lambda^k, \theta^k). \end{aligned} \tag{A.1}$$

The second inequality is obtained using the fact that $D_\psi(\phi, \phi) = 0$. Using recurrence (10), we get:

$$\hat{D}(\lambda^{k+1}, \theta^k) + D_\psi((\lambda^{k+1}, \theta^k), (\lambda^k, \theta^k)) \geq \hat{D}(\lambda^{k+1}, \theta^{k+1}) + D_\psi((\lambda^{k+1}, \theta^{k+1}), (\lambda^k, \theta^k))$$
$$\geq \hat{D}(\lambda^{k+1}, \theta^{k+1}). \tag{A.2}$$

The second inequality is obtained using the fact that $D(\phi, \phi') \geq 0$. The conclusion is reached by combining the two inequalities (A.1) and (A.2).

(b) Using the decreasing property previously proved in (a), we have by recurrence

$$\forall k, \hat{D}(\lambda^{k+1}, \theta^{k+1}) \leq \hat{D}(\lambda^k, \theta^k) \leq \cdots \leq \hat{D}(\lambda^0, \theta^0).$$

The result follows directly by definition of Φ^0.

(c) By induction on k. For $k = 0$, clearly $\phi^0 = (\lambda^0, \theta^0)$ is well defined (a choice we make). Suppose for some $k \geq 0$ that $\phi^k = (\lambda^k, \theta^k)$ exists. The infimum in (9) can be calculated on λ's such that $(\lambda, \theta^k) \in \Phi^0$. Indeed, suppose there exists a λ such that

$$\hat{D}(\lambda, \theta^k) + D_\psi((\lambda, \theta^k), (\lambda^k, \theta^k)) \leq \hat{D}(\lambda^k, \theta^k) + D_\psi((\lambda^k, \theta^k), (\lambda^k, \theta^k)) = \hat{D}(\lambda^k, \theta^k).$$

Then

$$\hat{D}(\lambda, \theta^k) \leq \hat{D}(\lambda, \theta^k) + D_\psi((\lambda, \theta^k), (\lambda^k, \theta^k)) \leq \hat{D}(\lambda^k, \theta^k) \leq \hat{D}(\phi^0).$$

This means that $(\lambda, \theta^k) \in \Phi^0$ and that the infimum need not be calculated for all values of $\lambda \in \Phi$, and can be restricted to values which satisfy $(\lambda, \theta^k) \in \Phi^0$. 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 clearly compact since Φ^0 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 satisfying this infimum.

The second part of the proof treats the definition of θ^{k+1} and is carried out analogously to λ^{k+1}. Convergence of the sequence $\{\hat{D}(\phi^k)\}_k$ in both algorithms results from the fact that it is nonincreasing and bounded. It is nonincreasing by virtue of (a). Boundedness results from the lower semicontinuity of $\phi \mapsto \hat{D}(\phi)$ and the compactness of the set Φ^0. ■

Proof of Proposition 2. Let $\{(\lambda^{n_k}, \theta^{n_k})\}_k$ be a convergent subsequence of $\{(\lambda^k, \theta^k)\}_k$ which converges to $(\lambda^\infty, \theta^\infty)$. First of all, $(\lambda^\infty, \theta^\infty) \in \Phi^0$, because Φ^0 is closed and the subsequence $\{(\lambda^{n_k}, \theta^{n_k})\}_k$ is a sequence of elements of Φ^0 (proved in Proposition 1(b)). Let us show that the subsequence $(\lambda^{n_k+1}, \theta^{n_k+1})$ also converges to $(\lambda^\infty, \theta^\infty)$. We simply have:

$$\|(\lambda^{n_k+1}, \theta^{n_k+1}) - (\lambda^\infty, \theta^\infty)\| \leq \|(\lambda^{n_k}, \theta^{n_k}) - (\lambda^\infty, \theta^\infty)\| + \|(\lambda^{n_k+1}, \theta^{n_k+1}) - (\lambda^{n_k}, \theta^{n_k})\|.$$

Since $(\lambda^{k+1}, \theta^{k+1}) - (\lambda^k, \theta^k) \to 0$ and $(\lambda^{n_k}, \theta^{n_k}) \to (\lambda^\infty, \theta^\infty)$, we conclude that $\phi^{n_k+1} \to \phi^\infty$. By definition of λ^{n_k+1} and θ^{n_k+1}, they achieve the infimum respectively in recurrences (9) and (10). Therefore, the gradient of the optimized function is zero for each step. In other words:

$$\nabla_\lambda \hat{D}(\lambda^{n_k+1}, \theta^{n_k}) + \nabla_\lambda D_\psi((\lambda^{n_k+1}, \theta^{n_k}), \phi^{n_k}) = 0,$$
$$\nabla_\theta \hat{D}(\lambda^{n_k+1}, \theta^{n_k+1}) + \nabla_\theta D_\psi((\lambda^{n_k+1}, \theta^{n_k+1}), \phi^{n_k}) = 0.$$

Since both $\{\phi^{n_k+1}\}$ and $\{\phi^{n_k}\}$ converge to the same limit ϕ^∞, then setting $\phi^\infty = (\lambda^\infty, \theta^\infty)$, we get that both λ^{n_k+1} and λ^{n_k} tend to λ^∞. We also have that both θ^{n_k+1} and θ^{n_k} tend to θ^∞.

The continuity of the two gradients (assumptions A1 and A4) implies that:

$$\nabla_\lambda \hat{D}(\lambda^\infty, \theta^\infty) + \nabla_\lambda D_\psi((\lambda^\infty, \theta^\infty), \phi^\infty) = 0,$$
$$\nabla_\theta \hat{D}(\lambda^\infty, \theta^\infty) + \nabla_\theta D_\psi((\lambda^\infty, \theta^\infty), \phi^\infty) = 0.$$

However, $\nabla D_\psi(\phi, \phi) = 0$, so that $\nabla \hat{D}(\phi^\infty) = 0$. ■

Proof of Proposition 3. By contradiction, let's suppose that $\phi^{k+1} - \phi^k$ does not converge to 0. We can prove using the compactness of Φ^0 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

$$\phi^{N(k)+1} \to \tilde{\phi}, \quad \phi^{N(k)} \to \bar{\phi} \text{ with } \bar{\phi} \neq \tilde{\phi}.$$

The real sequence $\hat{D}(\phi^k)_k$ converges as proved in Proposition 1(c) so that both sequences $\hat{D}(\phi^{N(k)+1})$ and $\hat{D}(\phi^{N(k)})$ converge to the same limit. In the proof of Proposition 1, we can deduce the following inequality:

$$\hat{D}(\lambda^{k+1}, \theta^{k+1}) + D_\psi((\lambda^{k+1}, \theta^{k+1}), \phi^k) \leq \hat{D}(\lambda^k, \theta^k) \tag{A.3}$$

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, $D_\psi(\tilde{\phi}, \bar{\phi}) > 0$, so that it becomes zero. Using assumption A3, $D_\psi(\tilde{\phi}, \bar{\phi}) = 0$ implies that $\tilde{\phi} = \bar{\phi}$. This contradicts the assumption that $\phi^{k+1} - \phi^k$ does not converge to 0. The second part of the proposition is a direct result of Proposition 2. ■

Proof of Corollary 1. Since the sequence $(\phi)_k$ is bounded and satisfies $\phi^{k+1} - \phi^k \to 0$, then Theorem 28.1 of 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. The remainder of the proof is a direct result of Theorem 3.3.1 of Chrétien & Hero (2008). The strict concavity of the objective function around an accumulation point is replaced here by the strict convexity of the estimated divergence. ■

Proof of Proposition 4. If $\{\phi^k\}_k$ converges to, say, ϕ^∞, then the result follows simply from Proposition 2. Suppose now that $(\phi^k)_k$ 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; 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.

Keeping previous notation, suppose that

$$\phi^{k+1} = (\lambda^{k+1}, \theta^{k+1}) \to \tilde{\phi} = (\tilde{\lambda}, \tilde{\theta})$$

and

$$\phi^k = (\lambda^k, \theta^k) \to \bar{\phi} = (\bar{\lambda}, \bar{\theta}).$$

We use again inequality (A.3)

$$\hat{D}(\lambda^{k+1}, \theta^{k+1}) + D_\psi((\lambda^{k+1}, \theta^{k+1}), (\lambda^k, \theta^k)) \leq \hat{D}(\lambda^k, \theta^k).$$

By taking the limits of the two parts of the inequality as k tends to infinity, and using the continuity of the two functions $\hat{D}$ and D_ψ, we have

$$\hat{D}(\tilde{\phi}) + D_\psi(\tilde{\phi}, \bar{\phi}) \leq \hat{D}(\bar{\phi}).$$

Recall that under assumptions A1 and A2, the sequence $\left\{\hat{D}(\phi^k)\right\}_k$ converges due to Proposition 1, so that it has the same limit for any subsequence, that is, $\hat{D}(\tilde{\phi}) = \hat{D}(\bar{\phi})$. We also use the fact that the distance-like function D_ψ is nonnegative to deduce that $D_\psi(\tilde{\phi}, \bar{\phi}) = 0$. Writing explicitly this equation gives

$$\sum_{i=1}^{n} \int_{\mathcal{X}} \psi\left(\frac{h_i(x|\tilde{\phi})}{h_i(x|\bar{\phi})}\right) h_i(x|\bar{\phi})\, dx = 0.$$

This is a sum of nonnegative terms. Thus, each term is also zero, that is

$$\forall i \in \{1, \ldots, n\}, \quad \int_{\mathcal{X}} \psi\left(\frac{h_i(x|\tilde{\phi})}{h_i(x|\bar{\phi})}\right) h_i(x|\bar{\phi})\, dx = 0.$$

The integrands are nonnegative functions, so that almost everywhere we have

$$\forall i \in \{1, \ldots, n\}, \quad \psi\left(\frac{h_i(x|\tilde{\phi})}{h_i(x|\bar{\phi})}\right) h_i(x|\bar{\phi}) = 0, \quad dx\text{-a.e.}$$

Since $h_i(x|\bar{\phi}) > 0, dx$-a.e. we have

$$\psi\left(\frac{h_i(x|\tilde{\phi})}{h_i(x|\bar{\phi})}\right) = 0, \quad dx\text{-a.e.}$$

On the other hand, ψ is chosen in a way that $\psi(z) = 0$ iff $z = 1$, therefore

$$\forall i \in \{1, \ldots, n\}, \quad h_i(x|\tilde{\phi}) = h_i(x|\bar{\phi}), \quad dx\text{-a.e.} \tag{A.4}$$

We will prove that $\frac{\partial}{\partial\theta}\hat{D}(\tilde{\lambda}, \tilde{\theta}) = 0$. Due to recurrence (10), the partial derivative at θ^{k+1} is zero, that is

$$\frac{\partial}{\partial\theta}\hat{D}(\lambda^{k+1}, \theta^{k+1}) + \frac{\partial}{\partial\theta} D_\psi((\lambda^{k+1}, \theta^{k+1}), (\lambda^k, \theta^k)) = 0.$$

By making k go to infinity and using the continuity assumption of the gradients (assumptions A1 and A4), we get

$$\frac{\partial}{\partial\theta}\hat{D}(\tilde{\lambda}, \tilde{\theta}) + \frac{\partial}{\partial\theta} D_\psi((\tilde{\lambda}, \tilde{\theta}), (\bar{\lambda}, \bar{\theta})) = 0.$$

The partial derivative with respect to θ of D_ψ is also a sum of terms of the form

$$\int_{\mathbb{R}} A(\tilde{\phi}, \bar{\phi}) \psi'\left(\frac{h_i(x|\tilde{\phi})}{h_i(x|\bar{\phi})}\right) dx.$$

Due to equalities (A.4), all terms in $\frac{\partial}{\partial\theta} D_\psi((\tilde{\lambda}, \tilde{\theta}), (\bar{\lambda}, \bar{\theta}))$ equal zero. Thus $\frac{\partial}{\partial\theta}\hat{D}(\lambda^{k+1}, \theta^{k+1}) = 0$.

We prove now that $\frac{\partial}{\partial\lambda}\hat{D}(\tilde{\lambda},\tilde{\theta})=0$. Using recurrence (9), λ^{k+1} is an optimum so that the gradient of the objective function is zero:

$$\frac{\partial}{\partial\lambda}\hat{D}(\lambda^{k+1},\theta^{k})+\frac{\partial}{\partial\lambda}D_{\psi}((\lambda^{k+1},\theta^{k}),(\lambda^{k},\theta^{k}))=0,\quad\forall k.$$

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:

$$\frac{\partial}{\partial\lambda}\hat{D}(\tilde{\lambda},\bar{\theta})+\frac{\partial}{\partial\lambda}D_{\psi}((\tilde{\lambda},\tilde{\theta}),(\bar{\lambda},\bar{\theta}))=0.$$

Since the derivative of D_{ψ} is a sum of terms which all depend on $\psi'\left(\frac{h_i(x|\tilde{\lambda},\bar{\theta})}{h_i(x|\bar{\lambda},\bar{\theta})}\right)$, and using identities (A.4), we conclude that

$$\psi'\left(\frac{h_i(x|\tilde{\lambda},\bar{\theta})}{h_i(x|\bar{\lambda},\bar{\theta})}\right)=\psi'(1)=0$$

and

$$\frac{\partial}{\partial\lambda}D_{\psi}((\tilde{\lambda},\bar{\theta}),\bar{\lambda},\bar{\theta})=0.$$

Finally, $\bar{\theta}=\tilde{\theta}$ implies that $\frac{\partial}{\partial\lambda}\hat{D}(\tilde{\lambda},\hat{\theta})=0$. ■

ACKNOWLEDGEMENTS

The authors would like to thank the editor, the associate editor and two anonymous referees for their valuable detailed comments and suggestions, which greatly improved the paper.

BIBLIOGRAPHY

Al Mohamad, D. (2018). Towards a better understanding of the dual representation of phi divergences. *Statistical Papers*, 59, 1205–1253.

Al Mohamad, D. & Broniatowski, M. (2015). Generalized EM algorithms for minimum divergence estimation. In *Geometric Science of Information: Second International Conference*, Palaiseau, France, 28–30 October 2015, Springer, pp. 417–426.

Al Mohamad, D. & Broniatowski, M. (2016). A proximal point algorithm for minimum divergence estimators with application to mixture models. *Entropy*, 18, 277.

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, 85, 549–559.

Basu, A. & Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46, 683–705.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, 5, 445–463.

Broniatowski, M. & Keziou, A. (2006). Minimization of divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43, 403–442.

Broniatowski, M. & Keziou, A. (2009). Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100, 16–36.

Chrétien, S. & Hero, A. O. (1998). Acceleration of the EM algorithm via proximal point iterations. In *Proceedings. 1998 IEEE International Symposium on Information Theory*, Cambridge, MA, 16–21 August 1998, p. 444.

Chrétien, S. & Hero, A. O. (2008). On EM algorithms and their proximal generalizations. *European Series in Applied and Industrial Mathematics (ESAIM): Probability and Statistics*, 12, 308–326.

Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 8, 95–108.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Goldstein, A. & Russak, I. (1987). How good are the proximal point algorithms? *Numerical Functional Analysis and Optimization*, 9, 709–724.

Ghosh, A., Harris, I. R., Maji, A., Basu, A., & Pardo, L. (2013). A generalized divergence for statistical inference. *Technical report, Byesian and Interdisciplinary Research Unit Indian Statistical Institute*.

Liese, F. & Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52, 4394–4412.

McLachlan, G. & Krishnan, T. (2007). *The EM Algorithm and Extensions*. Wiley, New York.

Meister, A. (2009). Deconvolution problems in nonparametric statistics. In *Lecture Notes in Statistics*: Springer, New York.

Ostrowski, A. (1966). *Solution of Equations and Systems of Equations*. Academic Press.

Park, C. & Basu, A. (2004). Minimum disparity estimation: Asymptotic normality and breakdown point results. *Bulletin of Informatics and Cybernetics*, 36, 19–33.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rockafellar, R. T. & Wets, R. J. B. (1998). *Variational Analysis*. Springer.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85, 617–624.

Toma, A. & Broniatowski, M. (2011). Dual divergence estimators and tests: Robustness results. *Journal of Multivariate Analysis*, 102, 20–36.

Toma, A. & Leoni-Aubin, S. (2013). Optimal robust M-estimators using Rényi pseudodistances. *Journal of Multivariate Analysis*, 115, 359–373.

Tseng, P. (2004). An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29, 27–44.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11, 95–103.

Received 12 January 2018
Accepted 04 February 2019