Ann Inst Stat Math (2012) 64:791–809
DOI 10.1007/s10463-011-0333-x

Space alternating penalized Kullback proximal point algorithms for maximizing likelihood with nondifferentiable penalty

Stéphane Chrétien · Alfred Hero · Hervé Perdry

Received: 21 August 2008 / Revised: 13 January 2011 / Published online: 11 August 2011

Abstract The EM algorithm is a widely used methodology for penalized likelihood estimation. Provable monotonicity and convergence are the hallmarks of the EM algorithm and these properties are well established for smooth likelihood and smooth penalty functions. However, many relaxed versions of variable selection penalties are not smooth. In this paper, we introduce a new class of space alternating penalized Kullback proximal extensions of the EM algorithm for nonsmooth likelihood inference. We show that the cluster points of the new method are stationary points even when they lie on the boundary of the parameter set. We illustrate the new class of algorithms for the problems of model selection for finite mixtures of regression and of sparse image reconstruction.

Keywords EM algorithm · Maximum likelihood estimation · Sparsity · Model selection · Space alternating algorithm · Nonsmooth penalty

S. Chrétien (✉)
Mathematics Department, UMR CNRS 6623, University of Franche Comte,
UFR-ST, 16 route de Gray, Besançon 25030, France
e-mail: stephane.chretien@univ-fcomte.fr

A. Hero
Department of Electrical Engineering and Computer Science, The University of Michigan,
1301 Beal Avenue, Ann Arbor, MI 48109-2122, USA
e-mail: hero@eecs.umich.edu

H. Perdry
Université Paris-Sud and Inserm UMR-S 669, Hôpital Paul Brousse, Villejuif Cedex 94817, France
e-mail: herve.perdry@u-psud.fr

1 Introduction

The EM algorithm of Dempster et al. (1977) is a widely applicable methodology for computing likelihood maximizers or at least stationary points. It has been extensively studied over the years and many useful generalizations have been proposed, including, for instance, the stochastic EM algorithm of Delyon et al. (1999) and Kuhn and Lavielle (2004); the PX–EM accelerations of Liu et al. (1998); the MM generalization of Hunter and Lange (2004) and approaches using extrapolation such as proposed in Varadhan and Roland (2007).

In recent years, much attention has been given to the problem of variable selection for multiparameter estimation, for which the desired solution is sparse, i.e. many of the parameters are zero. Several approaches have been proposed for recovering sparse models. A large number of contributions are based on the use of non-differentiable penalties like the LASSO (Tibshirani 1996; Candès and Plan 2009), ISLE (Friedmand and Popescu 2003) and "hidden variable"-type approach developed by Figueiredo and Nowak (2003). Other contributions are for instance sparse Bayes learning (Tipping 2001), information theoretic based prior methods of Barron (1999), empirical Bayes (Johnstone and Silverman 2004). Among recent alternatives is the new Dantzig selector of Candès and Tao (2007). On the other hand, only a few attempts have been made to use of non-differentiable penalization for more complex models than the linear model; for some recent progress, see Koh et al. (2007) for the case of logistic regression; and Khalili and Chen (2007) for mixture models.

In the present paper, we develop new extensions of the EM algorithm that incorporate a non-differentiable penalty at each step. Following previous work of the first two authors, we use a Kullback proximal interpretation for the EM-iterations and prove stationarity of the cluster points of the methods using nonsmooth analysis tools. Our analysis covers coordinate by coordinate methods such as space alternating extensions of EM and Kullback proximal point (KPP) methods and such component-wise versions of EM-type algorithms can benefit from acceleration of convergence speed (Fessler and Hero 1994). The KPP method was applied to Gaussian mixture models in Celeux et al. (2001). The main result of this paper is that any cluster point of the space alternating KPP method satisfies a nonsmooth Karush–Kuhn–Tucker condition.

The paper is organized as follows: in Sect. 2, we review penalized KPP methods and introduce componentwise PKPP algorithms with new differentiable penalties. In Sect. 3, our main asymptotic results are presented. In Sect. 4, we present a space alternating implementation of the penalized EM algorithm for a problem of model selection in a finite mixture of linear regressions using the SCAD penalty introduced in Fan and Li (2001) and further studied in Khalili and Chen (2007).

2 The EM algorithm and its Kullback proximal generalizations

The problem of maximum likelihood (ML) estimation consists of solving the maximization

$$\theta_{\mathrm{ML}} = \mathrm{argmax}_{\theta\in\Theta}\, l_y(\theta), \tag{1}$$

where y is an observed sample of a random variable Y defined on a sample space $\mathcal{Y}$ and $l_y(\theta)$ is the log-likelihood function defined by

$$l_y(\theta) = \log g(y;\theta),$$

on the parameter space $\Theta \subset \mathbb{R}^p$, and $g(y;\theta)$ denotes the density of Y at y parametrized by the vector parameter θ.

The standard EM approach to likelihood maximization introduces a complete data vector X with density f. Consider the conditional density function $k(x|y;\bar{\theta})$ of X given y

$$k(x|y;\bar{\theta}) = \frac{f(x;\bar{\theta})}{g(y;\bar{\theta})}. \tag{2}$$

As is well known, the EM algorithm then consists of alternating between two steps. The first step, called the E(xpectation) step, consists of computing the conditional expectation of the complete log-likelihood given Y. Notice that the conditional density k is parametrized by the current iterate of the unknown parameter value, denoted here by $\bar{\theta}$ for simplicity. Moreover, the expected complete log-likelihood is a function of the variable θ. Thus, the second step, called the M(aximization) step, consists of maximizing the obtained expected complete log-likelihood with respect to the variable parameter θ. The maximizer is then accepted as the new current iterate of the EM algorithm and the two steps are repeated until convergence is achieved.

Consider now the general problem of maximizing a concave function $\Phi(\theta)$. The original proximal point algorithm introduced by Martinet (1970) is an iterative procedure which can be written

$$\theta^{k+1} = \operatorname{argmax}_{\theta \in D_\Phi} \left\{ \Phi(\theta) - \frac{\beta_k}{2} \|\theta - \theta^k\|^2 \right\}. \tag{3}$$

The influence of the quadratic penalty $\frac{1}{2}\|\theta - \theta^k\|^2$ is controlled by the sequence of positive parameters $\{\beta_k\}$. Rockafellar (1976) showed that superlinear convergence of this method occurs when the sequence $\{\beta_k\}$ converges to zero. A relationship between proximal point algorithms and EM algorithms was discovered in Chrétien and Hero (2000) (see also Chrétien and Hero 2008 for details). We review the EM analogy to KPP methods to motivate the space alternating generalization. Assume that the family of conditional densities $\{k(x|y;\theta)\}_{\theta \in \mathbb{R}^p}$ is regular in the sense of Ibragimov and Has'minskii (1981), in particular $k(x|y;\theta)\mu(x)$ and $k(x|y;\bar{\theta})\mu(x)$ are mutually absolutely continuous for any θ and $\bar{\theta}$ in $\mathbb{R}^p$. Then, the Radon–Nikodym derivative $\frac{k(x|y,\bar{\theta})}{k(x|y;\theta)}$ exists for all $\theta, \bar{\theta}$ and we can define the following Kullback Leibler divergence:

$$I_y(\theta,\bar{\theta}) = \mathsf{E}\left[\log \frac{k(x|y,\bar{\theta})}{k(x|y;\theta)} | y;\bar{\theta}\right]. \tag{4}$$

Let us define D_l as the domain of l_y, $D_{I,\theta}$ the domain of $I_y(\cdot,\theta)$ and D_I the domain of $I_y(\cdot,\cdot)$. Using the distance-like function I_y, the KPP algorithm is defined by

$$\theta^{k+1} = \text{argmax}_{\theta \in D_\Phi} \left\{ \Phi(\theta) - \beta_k I_y(\theta, \bar{\theta}) \right\}. \tag{5}$$

The following was proved in Chrétien and Hero (2000).

Proposition 1 (Chrétien and Hero 2000, Proposition 1). *In the case where* Φ *is the log-likelihood, the EM algorithm is a special instance of the Kullback-proximal algorithm with* Φ *equal to the penalized log-likelihood and* $\beta_k = 1$*, for all* $k \in \mathbb{N}$.

2.1 The space alternating penalized Kullback-proximal method

In what follows, and in anticipation of component-wise implementations of penalized KPP, we will use the notation $\Theta_r(\theta)$ for the local decomposition at θ defined by $\Theta_r(\theta) = \Theta \cap (\theta + \mathcal{S}_r)$, $r = 1, \ldots, R$ where $\mathcal{S}_1, \ldots, \mathcal{S}_R$ are subspaces of $\mathbb{R}^p$ and $\mathbb{R}^p = \oplus_{r=1}^R \mathcal{S}_r$.

Then, the space alternating penalized proximal point algorithm is defined as follows.

Definition 1 Let $\psi \colon \mathbb{R}^p \mapsto \mathcal{S}_1 \times \cdots \times \mathcal{S}_R$ be a continuously differentiable mapping and let ψ_r denote its r^{th} coordinate. Let $(\beta_k)_{k \in \mathbb{N}}$ be a sequence of positive real numbers and λ be a positive real vector in $\mathbb{R}^R$. Let p_n be a nonnegative possibly nonsmooth locally Lipschitz penalty function with bounded Clarke-subdifferential (see the Appendix for details) on compact sets. Then, the space alternating penalized Kullback proximal algorithm is defined by

$$\begin{aligned} \theta^{k+1} = & \text{argmax}_{\theta \in \Theta_{k-1(\text{mod } R)+1}(\theta^k) \cap D_l \cap D_{I,\theta^k}} \\ & \times \left\{ l_y(\theta) - \sum_{r=1}^R \lambda_r p_n(\psi_r(\theta)) - \beta_k I_y(\theta, \theta^k) \right\}, \end{aligned} \tag{6}$$

where D_l is the domain of l_y and $D_{I,\theta}$ is the domain of $I_y(\cdot, \theta)$.

The standard KPP algorithms as defined in Chrétien and Hero (2008) is obtained as special case by selecting $R = 1, \Theta_1 = \Theta, \lambda = 0$.

The mappings ψ_r will simply be the projection onto the subspace Θ_r, $r = 1, \ldots, R$ in the sequel, but the proofs below allow for more general mappings too.

2.2 Notations and assumptions

The notation $\|\cdot\|$ will be used to denote the norm on any previously defined space. The space on which the norm operates should be obvious from the context. For any bivariate function Φ, $\nabla_1 \Phi$ will denote the gradient with respect to the first variable. For the convergence analysis, we will make the following assumptions. For a locally Lipschitz function f, $\partial f(x)$ denotes the Clarke subdifferential of f at x (see Appendix). Regular locally Lipschitz functions are defined in the Appendix.

Assumption 1 (i) l_y is differentiable and $l_y(\theta) - \sum_{r=1}^R \lambda_r p_n(\psi_r(\theta))$ converges to $-\infty$ whenever $\|\theta\|$ tends to $+\infty$. The function p_n is locally Lipschitz and regular.

(ii) The domain $D_{I,\theta}$ of $I(\cdot,\theta)$ is a subset of the domain D_l of l.
(iii) $(\beta_k)_{k\in\mathbb{N}}$ is a convergent nonnegative sequence of real numbers whose limit is denoted by β^*.
(iv) The mappings ψ_r are such that

$$\psi_r(\theta + \epsilon d) = \psi_r(\theta)$$

for all θ in Θ, all $d \in \mathcal{S}_r^{\perp}$ and $\epsilon > 0$ sufficiently small so that $\theta + \epsilon d \in \Theta$, $r = 1, \ldots, R$. This condition is satisfied for linear projection operators.

We will also impose one of the two following sets of assumptions on the distance-like function I_y in (4).

Assumption 2 (i) There exists a finite-dimensional Euclidean space S, a differentiable mapping $t : D_l \mapsto S$ and a functional $\psi : D_\psi \subset S \times S \mapsto \mathbb{R}$ such that KL divergence (4) satisfies

$$I_y(\theta, \bar{\theta}) = \Psi(t(\theta), t(\bar{\theta})),$$

where D_ψ denotes the domain of Ψ.
(ii) For any $\{(t^k, t)_{k\in\mathbb{N}}\} \subset D_\Psi$ there exists $\rho_t > 0$ such that $\lim_{\|t^k - t\|\to\infty} I_y(t^k, t) \geq \rho_t$. Moreover, we assume that $\inf_{t\in M} \rho_t > 0$ for any bounded set $M \subset S$. For all (t', t) in D_Ψ, we will also require that
(iii) (Positivity) $\Psi(t', t) \geq 0$,
(iv) (Identifiability) $\Psi(t', t) = 0 \Leftrightarrow t = t'$,
(v) (Continuity) Ψ is continuous at (t', t) and for all t belonging to the projection of D_Ψ onto its second coordinate,
(vi) (Differentiability) the function $\Psi(\cdot, t)$ is differentiable at t.

In the case where the Kullback divergence I_y is not defined everywhere (for instance if its domain of definition is the positive orthant), we need stronger assumptions to prove the desired convergence properties.

Assumption 3 (i) There exists a differentiable mapping $t : D_l \mapsto \mathbb{R}^{n\times m}$ such that the Kullback distance-like function I_y is of the form

$$I_y(\theta, \bar{\theta}) = \sum_{1\leq i\leq n, 1\leq j\leq m} \alpha_{ij}(y_j) t_{ij}(\theta) \phi\left(\frac{t_{ij}(\bar{\theta})}{t_{ij}(\theta)}\right),$$

where for all i and j, t_{ij} is continuously differentiable on its domain of definition, α_{ij} is a function from $\mathcal{Y}$ to $\mathbb{R}_+$, the set of positive real numbers,
(ii) The function ϕ is a non negative differentiable convex function defined $\mathbb{R}_*^+$ and such that $\phi(\tau) = 0$ if and only if $\tau = 1$.
(iii) There exists $\rho > 0$ such that

$$\lim_{\mathbb{R}_+ \ni \tau\to\infty} \phi(\tau) \geq \rho.$$

(iv) The mapping t is injective on each Θ_r.

In the context of Assumptions 3, D_I is simply the set

$$D_I = \{\theta \in \mathbb{R}^p \mid t_{ij}(\theta) > 0 \quad \forall i \in \{1, \ldots, n\} \quad \text{and} \quad j \in \{1, \ldots, m\}\}^2.$$

Notice that if $t_{ij}(\theta) = \theta_i$ and $\alpha_{ij} = 1$ for all i and all j, the functions I_y turn out to reduce to the well known ϕ divergence defined in Csiszár (1967). Assumptions 3 are satisfied by most standard examples (for instance Gaussian mixtures and Poisson inverse problems) with the choice $\phi(\tau) = \tau \log(\tau) - 1$.

Assumptions 1(i) and (ii) on l_y are standard and are easily checked in practical examples, e.g. they are satisfied for the Poisson and additive mixture models.

Finally, we make the following general assumption.

Assumption 4 The Kullback proximal iteration (6) is well defined, i.e. there exists at least one maximizer of (6) at each iteration k.

In the EM case, i.e. $\beta = 1$, this last assumption is equivalent to the computability of M-steps. In practice, it suffices to show the inclusion $0 \in \nabla l_y(\theta) - \lambda \partial p_n(\psi(\theta)) - \beta_k \nabla I_y(\theta, \theta^k)$ for $\theta = \theta^{k+1}$ to prove that the solution is unique. Then, Assumption 1(i) is sufficient for a maximizer to exist.

These technical assumptions play an important role in the theory developed below. Assumption 1(i) on differentiability of the log-likelihood is important for establishing the Karush–Kuhn–Tucker optimality conditions for cluster points. The fact that the objective should decrease to negative infinity as the norm of the parameter goes to infinity is often satisfied, or can be easily imposed, and is used later to guarantee boundedness of the sequence of iterates. The fact that p_n is regular is standard since the usual choices are the ℓ_1-norm, the ℓ_p-quasi-norms for $0 < p < 1$, the SCAD penalty, etc .... Assumption 1(ii) is only needed to simplify the analysis since, otherwise, each iterate would lie in the intersection of D_l and D_I and this would lead to asymptotic complications; this assumption is always satisfied in the models we have encountered in practice. Assumption 1(iii) is standard. Assumption 1(iv) is satisfied when ψ_r is a projection onto $\mathcal{S}_r$ and simplifies the proofs. Assumption 2 imposes natural conditions on the "distance" I_y. Assumption 2(ii) ensures that the "distance" I_y is large between points whose Euclidean distance goes to $+\infty$, thus weakening the assumption that I_y should grow to $+\infty$ in such a case. Assumptions 3 are used to obtain the Karush–Kuhn–Tucker conditions in Theorem 2. For this theorem, we require I_y to behave like a standard Kullback–Leibler "distance" and therefore that I_y has a more constrained shape. Assumption 3(iii) is a simplification of Assumption 2(ii). Assumption 3(iv) is a natural injectivity requirement.

3 Asymptotic properties of the Kullback-proximal iterations

3.1 Basic properties of the penalized Kullback proximal algorithm

Under Assumptions 1, we state basic properties of the penalized KPP algorithm. The most basic property is the monotonicity of the penalized likelihood function and the boundedness of the penalized proximal sequence $(\theta^k)_{k \in \mathbb{N}}$. The proofs of the following

lemmas are given, for instance, in Chrétien and Hero (2000) for the unpenalized case ($\lambda = 0$) and their generalizations to the present context is straightforward.

We start with the following monotonicity result.

Lemma 1 *For any iteration $k \in \mathbb{N}$, the sequence $(\theta^k)_{k\in\mathbb{N}}$ satisfies*

$$l_y(\theta^{k+1}) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta^{k+1})) - (l_y(\theta^k) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta^k))) \geq \beta_k I_y(\theta^k, \theta^{k+1}) \geq 0. \tag{7}$$

Lemma 2 *The sequence $(\theta^k)_{k\in\mathbb{N}}$ is bounded.*

The next lemma will also be useful and its proof in the unpenalized case where $\lambda = 0$ is given in Chrétien and Hero (2008) Lemma 2.4.3. The generalization to $\lambda > 0$ is also straightforward.

Lemma 3 *Assume that in the space alternating KPP sequence $(\theta^k)_{k\in\mathbb{N}}$, there exists a subsequence $(\theta^{\sigma(k)})_{k\in\mathbb{N}}$ belonging to a compact set C included in D_l. Then,*

$$\lim_{k\to\infty} \beta_k I_y(\theta^{k+1}, \theta^k) = 0.$$

One important property, which is satisfied in practice, is that the distance between two successive iterates decreases to zero. This property is critical to the definition of a stopping rule for the algorithm. This property was established in Chrétien and Hero (2008) in the case $\lambda = 0$.

Proposition 2 (Chrétien and Hero 2008, Proposition 4.1.2) *The following statements hold.*

(i) *For any sequence $(\theta^k)_{k\in\mathbb{N}}$ in $\mathbb{R}^p_+$ and any bounded sequence $(\eta^k)_{k\in\mathbb{N}}$ in $\mathbb{R}^p_+$, if $\lim_{k\to+\infty} I_y(\eta^k, \theta^k) = 0$ then $\lim_{k\to+\infty} |t_{ij}(\eta^k) - t_{ij}(\theta^k)| = 0$ for all i, j such that $\alpha_{ij} \neq 0$.*

(ii) *If $\lim_{k\to+\infty} I_y(\eta^k, \theta^k) = 0$ and one coordinate of one of the two sequences $(\theta^k)_{k\in\mathbb{N}}$ and $(\eta^k)_{k\in\mathbb{N}}$ tends to infinity, so does the other's same coordinate.*

3.2 Properties of cluster points

The results of this subsection state that any cluster point θ^* such that (θ^*, θ^*) lies on the closure of D_I satisfies a modified Karush–Kuhn–Tucker type condition. We first establish this result in the case where Assumptions 2 hold in addition to Assumptions 1 and 2 for the Kullback distance-like function I_y.

For notational convenience, we define

$$F_\beta(\theta, \bar{\theta}) = l_y(\theta) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta)) - \beta I_y(\theta, \bar{\theta}). \tag{8}$$

Theorem 1 *Assume that Assumptions 1, 2 and 4 hold and if $R > 1$, then, for each $r = 1, \ldots, R$, t is injective on Θ_r. Assume that the limit of $(\beta_k)_{k\in\mathbb{N}}$, β^*, is positive. Let θ^* be a cluster point of the space alternating penalized Kullback-proximal sequence (6). Assume the mapping t is differentiable at θ^*. If θ^* lies in the interior of D_l, then θ^* is a stationary point of the penalized log-likelihod function $l_y(\theta)$, i.e.*

$$0 \in \nabla l_y(\theta^*) - \sum_{r=1}^{R} \lambda_r \partial p_n(\psi_r(\theta^*)).$$

Proof We consider two cases, namely the case where $R = 1$ and the case where $R > 1$.

A. If $R = 1$ the proof is analogous to the proof of Theorem 3.2.1 in Chrétien and Hero (2008). In particular, we have

$$F_{\beta^*}(\theta^*, \theta^*) \geq F_{\beta^*}(\theta, \theta^*)$$

for all θ such that $(\theta, \theta^*) \in D_I$. Since $I_y(\theta, \theta^*)$ is differentiable at θ^*, the result follows by writing the first order optimality condition at θ^* in (9).

B. Assume that $R > 1$ and let $(x^{\sigma(k)})_{k\in\mathbb{N}}$ be a subsequence of iterates of (6) converging to θ^*. Moreover let $r = 1, \ldots, R$ and $\theta \in \Theta_r \cap D_l$. For each k, let $\sigma_r(k)$ the smallest index greater than $\sigma(k)$, of the form $\sigma(k') - 1$, with $k' \in \mathbb{N}$ and $(\sigma(k') - 1) \pmod R + 1 = r$. Using the fact that t is injective on every Θ_r, $r = 1, \ldots, R$, Lemma 3 and the fact that $(\beta_k)_{k\in\mathbb{N}}$ converges to $\beta^* > 0$, we easily conclude that $(\theta^{\sigma_r(k)})_{k\in\mathbb{N}}$ and $(\theta^{\sigma_r(k)+1})_{k\in\mathbb{N}}$ also converge to θ^*.

For k sufficiently large, we may assume that the terms $(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)})$ and $(\theta, \theta^{\sigma_r(k)})$ belong to a compact neighborhood C^* of (θ^*, θ^*) included in D_I. By Definition 1 of the space alternating penalized Kullback proximal iterations,

$$F_{\beta_{\sigma_r(k)}}\left(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}\right) \geq F_{\beta_{\sigma_r(k)}}\left(\theta, \theta^{\sigma_r(k)}\right).$$

Therefore,

$$\begin{aligned} F_{\beta^*}\left(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}\right) &- \left(\beta_{\sigma_r(k)} - \beta^*\right) I_y\left(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}\right) \\ &\geq F_{\beta^*}\left(\theta, \theta^{\sigma_r(k)}\right) - \left(\beta_{\sigma_r(k)} - \beta^*\right) I_y\left(\theta, \theta^{\sigma(k)}\right). \end{aligned} \quad (9)$$

Continuity of F_β follows directly from the proof of Theorem 3.2.1 in Chrétien and Hero (2008), where in that proof $\sigma(k)$ has to be replaced by $\sigma_r(k)$. This implies that

$$F_{\beta^*}\left(\theta^*, \theta^*\right) \geq F_{\beta^*}\left(\theta, \theta^*\right) \quad (10)$$

for all $\theta \in \Theta_r$ such that $(\theta, \theta^*) \in C^* \cap D_I$. Finally, recall that no assumption was made on θ, and that C^* is a compact neighborhood of θ^*. Thus, using the Assumption 1(i), which asserts that $l_y(\theta)$ tends to $-\infty$ as $\|\theta\|$ tends to $+\infty$, we may deduce

that (10) holds for any $\theta \in \Theta_r$ such that $(\theta, \theta^*) \in D_I$ and, letting ϵ tend to zero, we see that θ^* maximizes $F_{\beta^*}(\theta, \theta^*)$ for all $\theta \in \Theta_r$ such that (θ, θ^*) belongs to D_I as claimed.

To conclude the proof of Theorem 1, take d in $\mathbb{R}^p$ and decompose d as $d = d_1 + \cdots + d_R$ with $d_r \in \mathcal{S}_r$. Then, Eq. (10) implies that the directional derivatives satisfy

$$F'_{\beta^*}\left(\theta^*, \theta^*; d_r\right) \leq 0 \tag{11}$$

for all $r = 1, \ldots, R$. Due to Assumption 1(iv), the directional derivative of $\sum_{r=1}^{R} \lambda_r p_n(\psi_r(\cdot))$ in the direction d is equal to the sum of the partial derivatives in the directions $d_1, \ldots, d_R$ and, since all other terms in the definition of F_β are differentiable, we obtain using (11), that

$$F'_{\beta^*}\left(\theta^*, \theta^*; d\right) = \sum_{r=1}^{R} F'_{\beta^*}\left(\theta^*, \theta^*; d_r\right) \leq 0.$$

Therefore, using the assumption that p_n is regular (see Assumption 1(i)) which says that $p_n^\circ = p'_n$, together with characterization (22) of the subdifferential in the Appendix and Proposition 2.1.5(a) in Clarke (1990), the desired result follows. □

Next, we consider the case where Assumptions 3 hold.

Theorem 2 *Assume that in addition to Assumptions 1 and 4, Assumptions 3 hold. Let θ^* be a cluster point of the space alternating penalized Kullback proximal sequence. Assume that all the functions t_{ij} are continuously differentiable at θ^*. Let $\mathcal{I}^*$ denote the index of the active constraints at θ^*, i.e. $\mathcal{I}^* = \{(i, j)$ s.t. $t_{ij}(\theta^*) = 0\}$. If θ^* lies in the interior of D_l, then θ^* satisfies the following property: there exists a family of subsets $\mathcal{I}_r^{**} \subset \mathcal{I}^*$ and a set of real numbers λ_{ij}^*, $(i, j) \in \mathcal{I}_r^{**}$, $r = 1, \ldots, R$ such that*

$$0 \in \nabla l_y(\theta^*) - \sum_{r=1}^{R} \lambda_r \partial p_n(\psi_r(\theta^*)) + \sum_{r=1}^{R} \sum_{(i,j)\in\mathcal{I}_r^{**}} \lambda_{ij}^* P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)), \tag{12}$$

where $P_{\mathcal{S}_r}$ is the projection onto $\mathcal{S}_r$.

Remark 1 The condition (12) resembles the traditional Karush–Kuhn–Tucker conditions of optimality but is in fact weaker since the vector

$$\sum_{r=1}^{R} \sum_{(i,j)\in\mathcal{I}_r^{**}} \lambda_{ij}^* \mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))$$

in Eq. (12) does not necessarily belong to the normal cone at θ^* to the set $\{\theta \mid t_{ij} \geq 0,\ i = 1, \ldots, n,\ j = 1, \ldots, m\}$.

Proof of Theorem 2 Let $\Phi_{ij}(\theta, \bar{\theta})$ denote the bivariate function defined by

$$\Phi_{ij}(\theta, \bar{\theta}) = \phi\left(\frac{t_{ij}(\bar{\theta})}{t_{ij}(\theta)}\right).$$

As in the proof of Theorem 1, let $(x^{\sigma(k)})_{k\in\mathbb{N}}$ be a subsequence of iterates of (6) converging to θ^*. Moreover let $r = 1, \ldots, R$ and $\theta \in \Theta_r \cap D_l$. For each k, let $\sigma_r(k)$ be the next index greater than $\sigma(k)$ such that $(\sigma_r(k) - 1) \pmod R + 1 = r$. Using the fact that t is injective on every $\Theta_r, r = 1, \ldots, R$, Lemma 3 and the fact that $(\beta_k)_{k\in\mathbb{N}}$ converges to $\beta^* > 0$, we easily conclude that $(\theta^{\sigma_r(k)})_{k\in\mathbb{N}}$ and $(\theta^{\sigma_r(k)+1})_{k\in\mathbb{N}}$ also converge to θ^*.

Due to Assumption 3(iv), the first order optimality condition at iteration $\sigma_r(k)$ can be written

$$0 = \mathrm{P}_{\mathcal{S}_r}\left(\nabla l_y\left(\theta^{\sigma(k)+1}\right)\right) - \lambda_r g_r^{\sigma_r(k)+1} + \beta_{\sigma_r(k)}\left(\sum_{ij}\alpha_{ij}(y_j)\mathrm{P}_{\mathcal{S}_r}\left(\nabla t_{ij}\left(\theta^{\sigma_r(k)+1}\right)\right)\right.$$
$$\left.\times\,\Phi_{ij}\left(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}\right) + \sum_{ij}\alpha_{ij}(y_j)t_{ij}\left(\theta^{\sigma_r(k)+1}\right)\mathrm{P}_{\mathcal{S}_r}\left(\nabla_1\Phi_{ij}\left(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}\right)\right)\right) \tag{13}$$

with $g_r^{\sigma_r(k)+1} \in \partial p_n(\psi_r(\theta^{\sigma_r(k)+1}))$.

Moreover, Claim A in the proof of Theorem 4.2.1 in Chrétien and Hero (2008), gives that for all (i, j) such that $\alpha_{ij}(y_j) \neq 0$

$$\lim_{k\to+\infty} t_{ij}\left(\theta^{\sigma_r(k)+1}\right)\nabla_1\Phi_{ij}\left(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}\right) = 0. \tag{14}$$

Let $\mathcal{I}_r^*$ be a subset of indices such that the family $\{\mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))\}_{(i,j)\in\mathcal{I}_r^*}$ is linearly independent and spans the linear space generated by the family of all projected gradients $\{\mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))\}_{i=1,\ldots,n,\, j=1,\ldots,m}$. Since this linear independence are preserved under small perturbations (continuity of the gradients), we may assume, without loss of generality, that the family

$$\left\{\mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^{\sigma_r(k)+1}))\right\}_{(i,j)\in\mathcal{I}_r^*}$$

is linearly independent for k sufficiently large. For such k, we may thus rewrite equation (13) as

$$0 = \mathrm{P}_{\mathcal{S}_r}(\nabla l_y(\theta^{\sigma_r(k)+1})) - \lambda_r g_r^{\sigma_r(k)+1} + \beta_{\sigma_r(k)} \left(\sum_{(i,j)\in\mathcal{I}_r^*} \pi_{ij}^{\sigma_r(k)+1}(y_j) \right.$$
$$\left. \times \mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^{\sigma_r(k)+1})) + \sum_{ij} \alpha_{ij}(y_j) t_{ij}(\theta^{\sigma_r(k)+1}) \mathrm{P}_{\mathcal{S}_r}(\nabla_1 \Phi(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)})) \right), \tag{15}$$

where

$$\pi_{ij}^{\sigma_r(k)+1}(y_j) = \alpha_{ij}(y_j)\Phi_{ij}\left(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}\right). \tag{16}$$

Claim *The sequence* $\{\pi_{ij}^{\sigma_r(k)+1}(y_j)\}_{k\in\mathbb{N}}$ *has a convergent subsequence for all* (i, j) *in* I_r^*.

Proof of the claim Since the sequence $(\theta^k)_{k\in\mathbb{N}}$ is bounded, ψ is continuously differentiable and the penalty p_n has bounded subdifferential on compact sets, there exists a convergent subsequence $\left(g_r^{\sigma_r(\gamma(k))+1}\right)_{k\in\mathbb{N}}$ with limit g_r^*. Now, using Eq. (14), this last equation implies that $\left\{\pi_{(i,j)\in\mathcal{I}_r^*}^{\sigma_r(\gamma(k))+1}(y_j)\right\}_{(i,j)\in\mathcal{I}_r^*}$ converges to the coordinates of a vector in the linearly independent family $\left\{\mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))\right\}_{(i,j)\in\mathcal{I}_r^*}$. This concludes the proof. □

The above claim allows us to finish the proof of Theorem 2. Since a subsequence $(\pi_{ij}^{\sigma_r(\gamma(k))+1}(y_j))_{(i,j)\in\mathcal{I}_r^*}$ is convergent, we may consider its limit $(\pi_{ij}^*)_{(i,j)\in\mathcal{I}_r^*}$. Passing to the limit, we obtain from Eq. (13) that

$$0 = \mathrm{P}_{\mathcal{S}_r}(\nabla l_y(\theta^*)) - \lambda_r g_r^* + \beta^* \left(\sum_{(i,j)\in\mathcal{I}_r^*} \pi_{ij}^* \mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)) \right). \tag{17}$$

Using the outer semi-continuity property of the subdifferential of locally Lipschitz functions (see Appendix) we thus obtain that $g_r^* \in \partial p_n(\psi_r(\theta^*))$. Now, summing over r in (17), we obtain

$$0 = \sum_{r=1}^{R} \mathrm{P}_{\mathcal{S}_r}(\nabla l_y(\theta^*)) - \sum_{r=1}^{R} \lambda_r g_r^* + \beta^* \sum_{r=1}^{R} \left(\sum_{(i,j)\in\mathcal{I}_r^*} \pi_{ij}^* \mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)) \right).$$

Moreover, since $\Phi_{ij}(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)})$ tends to zero if $(i, j) \notin \mathcal{I}^*$, i.e. if the constraint on component (i, j) is not active, Eq. (16) implies that

$$0 = \sum_{r=1}^{R} \mathrm{P}_{\mathcal{S}_r}(\nabla l_y(\theta^*)) - \sum_{r=1}^{R} \lambda_r g_r^* + \beta^* \sum_{r=1}^{R} \left(\sum_{(i,j)\in\mathcal{I}_r^{**}} \pi_{ij}^* \mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)) \right)$$

where $\mathcal{I}_r^{**}$ is the subset of active indices of $\mathcal{I}_r^*$, i.e. $\mathcal{I}_r^{**} = \mathcal{I}_r^* \cap \mathcal{I}^*$. Since $\sum_{r=1}^R \lambda_r g_r^* \in \sum_{r=1}^R \lambda_r \partial p_n(\psi_r(\theta^*))$, this implies that

$$0 \in \nabla l_y(\theta^*) - \sum_{r=1}^R \lambda_r \partial p_n(\psi_r(\theta^*)) + \beta^* \sum_{r=1}^R \sum_{(i,j)\in\mathcal{I}_r^{**}} \pi_{ij}^* \mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)), \quad (18)$$

which establishes Theorem 2 once we define $\lambda_{ij}^* = \lambda^* \pi_{ij}^*$. □

The result (18) can be refined to the classical Karush–Kuhn–Tucker type condition under additional conditions such as stated below.

Corollary 1 *If in addition to the assumptions of Theorem 2 we assume that either* $P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)) = \nabla t_{ij}(\theta^*)$ *or* $P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)) = 0$ *for all* $(i, j) \in \mathcal{I}^*$, *i.e. such that* $t_{ij}(\theta^*) = 0$, *then there exists a set of subsets* $\mathcal{I}_r^{**} \subset \mathcal{I}^*$ *and a family of real numbers* $\lambda_{ij}^*, (i, j) \in \mathcal{I}_r^{**}$, $r = 1, \ldots, R$ *such that the following Karush–Kuhn–Tucker condition for optimality holds at cluster point* θ^*:

$$0 \in \nabla l_y(\theta^*) - \sum_{r=1}^R \lambda_r \partial p_n(\psi_r(\theta^*)) + \sum_{r=1}^R \sum_{(i,j)\in\mathcal{I}_r^{**}} \lambda_{ij}^* \nabla t_{ij}(\theta^*).$$

4 Application: variable selection in finite mixtures of regression models

Variable subset selection in regression models is frequently performed using penalization of the likelihood function, e.g. using AIC, Akaike (1973) and BIC, Schwarz (1978) penalties. The main drawback of these approaches is lack of scalability due to a combinatorial explosion of the set of possible models as the number of variables increases. Newer methods use l_1-type penalties of likelihood functions, as in the LASSO, Tibshirani (1996) and the Dantzig selector of Candès and Tao (2007), to select subsets of variables without enumeration.

Computation of maximizers of the penalized likelihood function can be performed using standard algorithms for nondifferentiable optimization such as bundle methods, as introduced in Hiriart-Urruty and Lemaréchal (1993). However general purpose optimization methods might be difficult to implement in the situation where, for instance, log objective functions induce line-search problems. In certain cases, the EM algorithm, or a combination of EM type methods with general purpose optimization routines might be simpler to implement. Variable selection in finite mixture models, as described in Khalili and Chen (2007), represents such a case due to the presence of very natural hidden variables.

In the finite mixture estimation problem considered here, $y_1, \ldots, y_n$ are realizations of the response variable Y and $x_1, \ldots, x_n$ are the associated realizations of the P-dimensional vector of covariates X. We focus on the case of a mixture of linear

regression models sharing the same variance, as in the baseball data example of section 7.2 in Khalili and Chen (2007), i.e.

$$Y \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(X^t \beta_k, \sigma^2), \tag{19}$$

with $\pi_1, \ldots, \pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. The main problem discussed in Khalili and Chen (2007) is model selection for which a generalization of the smoothly clipped absolute deviation (SCAD) method of Fan and Li (2001, 2002) is proposed using an MM–EM algorithm in the spirit of Hunter and Lange (2004). No convergence property of the MM algorithm was established. The purpose of this section is to show that the space alternating KPP EM generalization is easily implemented and that stationarity of the cluster points is guaranteed by the theoretical analysis of Sect. 3.

The SCAD penalty, studied in Khalili and Chen (2007) is a modification of the Ll_1 penalty which is given by

$$p_n(\beta_1, \ldots, \beta_K) = \sum_{k=1}^{K} \pi_k \sum_{j=1}^{P} p_{\gamma_{nk}}(\beta_{k,j})$$

where p_{nk} is specified by

$$p'_{\gamma_{nk}}(\beta) = \gamma_{nk} \sqrt{n} 1_{\sqrt{n}|\beta| \leq \gamma_{nk}} + \frac{\sqrt{n}(a\gamma_{nk} - \sqrt{n}|\beta|)_+}{a-1} 1_{\sqrt{n}|\beta| > \gamma_{nk}}$$

for β in $\mathbb{R}$.

Define the missing data as the class labels $z_1, \ldots, z_n$ of the mixture component from which the observed data point y_n was drawn. The complete log-likelihood is then

$$l_c(\beta_1, \ldots, \beta_K, \sigma^2) = \sum_{i=1}^{n} \log(\pi_{z_i}) - \frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - x_i^t \beta_{z_i})^2}{2\sigma^2}.$$

Setting $\theta = (\pi_1, \ldots, \pi_K, \beta_1, \ldots, \beta_K, \sigma^2)$, the penalized Q-function is given by

$$Q(\theta, \bar{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}(\bar{\theta}) \left[\log(\pi_k) - \frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - x_i^t \beta_k)^2}{2\sigma^2} \right] - p_n(\beta_1, \ldots, \beta_K)$$

where

$$t_{ik}(\theta) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - X\beta_k)^2}{2\sigma^2}\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - X\beta_l)^2}{2\sigma^2}\right)}.$$

The computation of this Q-function accomplishes the E-step. Moreover, a penalty of the form $-\sum_{k=1}^{K}\sum_{j=1}^{P}|\max\{10^6, |\beta_{k,j}|\} - 10^6|$ can be added to the log-likelihood function in order to ensure that Assumptions 1(i) (convergence of the penalized log-likelihood to $-\infty$ for parameter values with norm growing to $+\infty$) is satisfied for the case where X is not invertible. Owing to the fact that the penalty p_n is a function of the mixture probabilities π_k, the M-step estimate of the π vector is not given by the usual formula

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n} t_{ik}(\bar{\theta}) \qquad k = 1, \ldots, K. \tag{20}$$

This, however, is the choice made in Khalili and Chen (2007) in their implementation. Moreover, optimizing jointly over the variables β_k and π_k is clearly a more complicated task than independently optimizing with respect to each variable. We implement a componentwise version of EM consisting of successively optimizing with respect to the π_k's and alternatively with respect to the vectors β_k. Optimization with respect to the π_k's can be easily performed using standard differentiable optimization routines and optimization with respect to the β_k's can be performed by a standard non-differentiable optimization routine, e.g. as provided by the function optim of Scilab using the 'nd' (standing for 'non-differentiable') option.

We now turn to the description of the Kullback proximal penalty I_y defined by (4). The conditional density function $k(y_1, \ldots, y_n, z_1, \ldots, z_n \mid y_1, \ldots, y_n; \theta)$ is

$$k(y_1, \ldots, y_n, z_1, \ldots, z_n \mid y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} t_{iz_i}(\theta).$$

and therefore, the Kullback distance-like function $I_y(\theta, \bar{\theta})$ is

$$I_y(\theta, \bar{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}(\bar{\theta}) \log\left(\frac{t_{ik}(\bar{\theta})}{t_{ik}(\theta)}\right). \tag{21}$$

We have $R = K + 1$ subsets of variables with respect to which optimization will be performed successively. All components of Assumptions 1 and 3 are trivially satisfied for this model. Validation of Assumption 3(iv) is provided by Lemma 1 of Celeux et al. (2001). On the other hand, since $t_{ik}(\theta) = 0$ implies that $\pi_k = 0$ and $\pi_k = 0$ implies

$$\frac{\partial t_{ik}}{\partial \beta_{jl}}(\theta) = 0$$

for all $j = 1, \ldots, p$ and $l = 1, \ldots, K$ and

$$\frac{\partial t_{ik}}{\partial \sigma^2}(\theta) = 0,$$

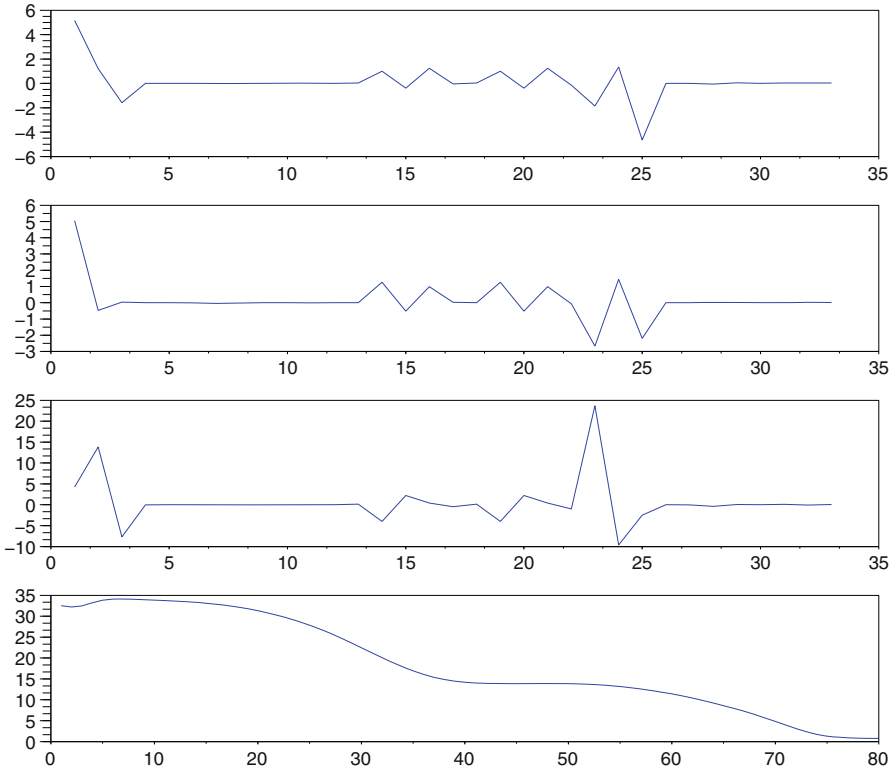

Fig. 1 Baseball data of Khalili and Chen (2007). This experiment is performed with the plain EM. The parameters are $\gamma_{nk} = 0.1$ and $a = 10$. The *first plot* is the vector β obtained for the single component model. The *second* (resp. *third*) *plot* is the vector of the optimal β_1 (resp. β_2). The *fourth plot* is the Euclidean distance to the optimal θ^* versus iteration index. The starting value of π_1 was 0.3

it follows that $P_{\mathcal{S}_r}(\nabla t_{ik}(\theta^*)) = \nabla t_{ik}(\theta^*)$ if $\mathcal{S}_r$ is the vector space generated by the probability vectors π and $P_{\mathcal{S}_r}(\nabla t_{ik}(\theta^*)) = 0$ otherwise. Therefore, Corollary 1 applies.

We illustrate this algorithm on real data (available at http://www.amstat.org/publications/jse/v6n2/datasets.watnik.html).

Khalili and Chen (2007) report that a model with only two components was selected by the BIC criterion in comparison to a three components model. Here, two alternative algorithms are compared: the approximate EM using (20) and the plain EM using the optim subroutines. The results for $\gamma_{nk} = 1$ and $a = 10$ are given in Fig. 1.

The results shown in Fig. 1 establish that the approximate EM algorithm has similar properties to the plain EM algorithm for small values of the threshold parameters γ_{nk}. Moreover, the larger the values of γ_{nk}, the closer the probability of the first component is to 1. One important fact to notice is that with the plain EM algorithm, the optimal probability vector becomes singular, in the sense that the second component has zero probability, as shown in Fig. 2. Figure 3 demonstrates that the approximate EM algorithm of Khalili and Chen (2007) does not produce optimal solutions.

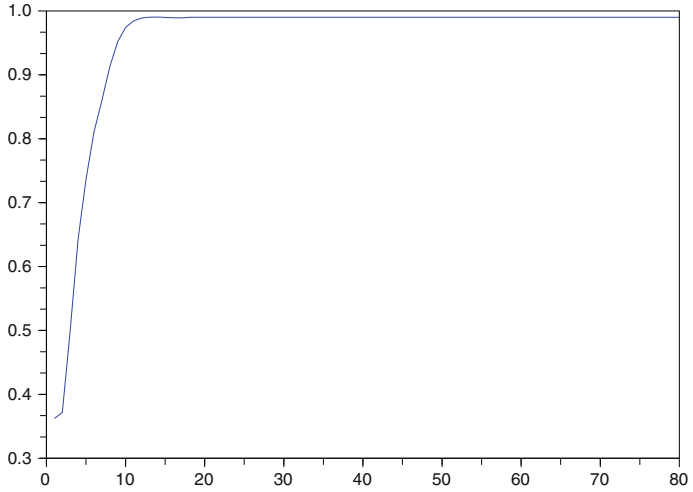

Fig. 2 This experiment is performed with the plain EM for the baseball data of Khalili and Chen (2007). The parameters are $\gamma_{nk} = 5$ and $a = 10$. The *plot* shows the probability π_1 of the first component versus iteration index. The starting value of π_1 was 0.3

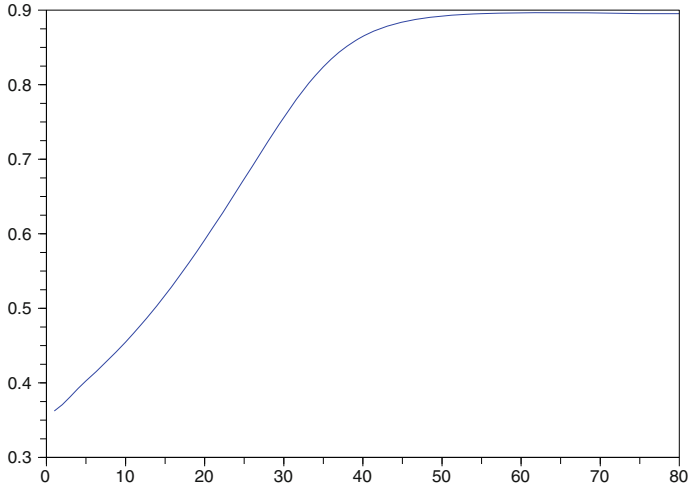

Fig. 3 Baseball data of Khalili and Chen (2007). This experiment is performed with the approximate EM. The parameters are $\gamma_{nk} = 5$ and $a = 10$. The *plot* shows the probability π_1 of the first component versus iteration index. The starting value of π_1 was 0.3

5 Conclusion

In this paper, we analyzed the expectation maximization (EM) algorithm with non-differentiable penalty. By casting the EM algorithm as a KPP iteration, we proved the stationarity of the cluster points and showed that any cluster point of the space alternating KPP method satisfies a nonsmooth Karush–Kuhn–Tucker condition.

The theory was applied to a space alternating implementation of the penalized EM algorithm for a problem of model selection in a finite mixture of linear regressions.

Acknowledgments The authors would like to thank the editor and one reviewer for their careful reading of the manuscript. Alfred Hero was partially supported by the National Science Foundation Grant Number CCF 0830490.

6 Appendix: the Clarke subdifferential of a locally Lipschitz function

Since we are dealing with non differentiable functions, the notion of generalized differentiability is required. The main references for this appendix are Clarke (1990) and Rockafellar and Wets (2004). A locally Lipschitz function $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ always has a generalized directional derivative $f^\circ(\theta,\omega)\colon \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ in the sense given by Clarke, i.e.

$$f^\circ(\theta,\omega) = \limsup_{\eta\in\mathbb{R}^p\to\theta,\, t\downarrow 0} \frac{f(\eta+t\omega)-f(\eta)}{t}.$$

A locally Lipschitz function is called *regular* if it admits a directional derivative at every point and if moreover this directional derivative coincides with Clarke's generalized directional derivative.

The Clarke subdifferential of f at θ is the convex set defined by

$$\partial f(\theta) = \{\eta \mid f^\circ(\theta,\omega) \geq \eta^t\omega,\ \ \forall\omega\}. \tag{22}$$

Proposition 3 *The function f is differentiable if and only if $\partial f(\theta)$ is a singleton.*

We now introduce another very important property of the Clarke subdifferential related to generalization of semicontinuity for set-valued maps.

Definition 2 A set-valued map Φ is said to be outer-semicontinuous if its graph

$$\text{graph}\,\Phi = \{(\theta, g) \mid g \in \Phi(\theta)\}$$

is closed, i.e. if for any sequence $(\text{graph}\Phi \ni)\ (\theta_n, g_n) \to (\theta^*, g^*)$ as $n \to +\infty$, then $(\theta^*, g^*) \in \text{graph}\Phi$.

One crucial property of the Clarke subdifferential is that it is outer-semicontinuous.

A point θ is said to be a *stationary point* of f if

$$0 \in \partial f(\theta).$$

Consider now the problem

$$\sup_{\theta\in\mathbb{R}^p} f(\theta)$$

subject to

$$g(\theta) = [g_1(\theta), \ldots, g_m(\theta)]^t \geq 0$$

where all the functions are locally Lipschitz from $\mathbb{R}^p$ to $\mathbb{R}$. Then, a necessary condition for optimality of θ is the Karush–Kuhn–Tucker condition, i.e. there exists a vector $u \in \mathbb{R}^m_+$ such that

$$0 \in \partial f(\theta) + \sum_{j=1}^{m} u_j \partial g_j(\theta).$$

Convex functions are in particular locally Lipschitz. The main references for these facts are Rockafellar (1970) and Hiriart-Urruty and Lemaréchal (1993).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory (Tsahkadsor, 1971)* (pp. 267–281). Budapest: Akadmiai Kiad.

Barron, A. R. (1999). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian statistics, 6 (Alcoceber 1998)* (Vol. 6, 2752). New York: Oxford University Press.

Candès, E., Plan, Y. (2009). Near-ideal model selection by L1 minimization. *The Annals of Statistics, 37*(5), 2145–2177.

Candès, E., Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics, 35* (6), 2313–2351.

Celeux, G., Chrétien, S., Forbes, F., Mkhadri, A. (2001). A component-wise EM algorithm for mixtures. *Journal of Computational and Graphical Statistics, 10*(4), 697–712.

Chrétien, S., Hero, A. O. (2000). Kullback proximal algorithms for maximum-likelihood estimation. Information-theoretic imaging. *IEEE Transactions on Information Theory, 46*(5), 1800–1810.

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

Clarke, F. (1990). *Optimization and nonsmooth analysis* (Vol. 5: Classics in Applied Mathematics). Philadelphia: Society for Industrial and Applied Mathematics.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica, 2*, 299–318.

Delyon, B., Lavielle, M., Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics, 27*, (1), 94–128.

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), 1–38.

Fan, J., Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*, 1348–1360.

Fan, J., Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics, 30*, 74–99.

Fessler, J. A., Hero, A. O. (1994). Space-alternating generalized expectation–maximization algorithm. *IEEE Transactions on Signal Processing, 42*(10), 2664–2677.

Figueiredo, M. A. T., Nowak, R. D. (2003). An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing, 12*(8), 906–916.

Friedmand, J., Popescu, B. E. (2003). Importance sampled learning ensembles. *Journal of Machine Learning Research* (submitted).

Hiriart-Urruty, J. B., Lemaréchal, C. (1993). *Convex Analysis and Minimization Algorithms* (Vol. 306). Grundlehren der mathematischen Wissenschaften. Berlin: Springer.

Hunter, D. R., Lange, K. (2004). A Tutorial on MM Algorithms. *The American Statistician, 58*,(1), 30–37.

Ibragimov, I.A., Hasminski, R.Z. (1981). *Statistical estimation. asymptotic theory* Translated from the Russian by Samuel Kotz. Applications of Mathematics (Vol. 16.). New York: Springer.

Johnstone, I. M., Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics, 32*(4), 1594–1649.

Khalili, A., Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association, 102*(479), 1025–1038.

Koh, K., Kim, S.-J., Boyd, S. (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research, 8*, 1519–1555.

Kuhn, E.,, Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *European Society for Applied and Industrial Mathematics Probability and Statistics, 8*, 115–131.

Liu, C., Rubin, D. B., Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika, 85*(4), 755–770.

Martinet, B. (1970). Régularisation d'inéquation variationnelles par approximations successives. *Revue Francaise d'Informatique et de Recherche Operationnelle, 3*, 154–179.

Rockafellar, R. T. (1970). *Convex analysis*. Convex analysis. Princeton Mathematical Series (No. 28) Princeton: Princeton University Press.

Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. *Society for Industrial and Applied Mathematics Journal on Control and Optimization, 14*, 877–898.

Rockafellar, R. T., Wets, R. J. B. (2004). *Variational analysis*. Variational analysis, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] (Vol. 317). Berlin: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B, 58* (1), 267–288.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research, 1*(3), 211–244.

Varadhan, R., Roland, Ch. (2007). Simple and Globally-Convergent Numerical Methods for Accelerating Any EM Algorithm. *Scandinavian Journal of Statistics, 35* (2), 335–353.