Space Alternating Penalized Kullback Proximal Point Algorithms for Maximing Likelihood with Nondifferentiable Penalty

Stéphane Chrétien [*] Alfred Hero [†] Hervé Perdry [‡]

December 30, 2008

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. The goal of this paper is to 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 on the boundary of the parameter set. Special attention has been paid to the construction of component-wise version of the method in order to ease the implementation for complicated models. Illustration for the problems of model selection for finite mixtures of regression and to sparse image reconstruction is presented.

Keywords: EM Algorithm Maximum Likelihood Estimation Sparsity Model Selection Space Alternating Algorithm Nonsmooth Penalty.

1 Introduction

The EM algorithm of Dempster Laird and Rudin (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 generalizationshave been proposed including for instance the stochastic EM algorithm of Delyon, Lavielle and Moulines (1999); Kuhn and Lavielle (2004); the PX-EM accelerations of Liu, Rubin and Wu (1998); the MM generalization of Lange and

[*] Mathematics Department, UMR CNRS 6623 and University of Franche Comte, UFR-ST, 16 route de Gray, 25030 Besançon, France. **Email**: stephane.chretien@univ-fcomte.fr

[†] department of electrical engineering and computer science, The University of Michigan, 1301 Beal Avenue, Ann Arbor,MI 48109-2122, USA. **Email**: hero@eecs.umich.edu

[‡] INSERM U535, Université Paris-Sud Pavillon Leriche Secteur Jaune - Porte 18, BP 1000, 94817 Villejuif Cedex, France.**Email**: perdry@vjf.inserm.fr

Hunter (2004) and the recent approach 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 desiered solution is spase, i.e. many of the parameters are zero. Several approaches have been proposed for recovering sparse models. The main contributions in this direction are sparse Bayes learning (Tipping ()), LASSO-like penalties penalized least squares (Tibshirani (1996)), ISLE (Friedman and Popescu (2003)), information theoretic based prior methods of Barron (1999), empirical Bayes (Johnstone and Silverman ()) and "hidden variable"-type approach developped by Figueiredo and Nowak (2003). Among recent and exciting alternatives is the new Dantzig selector of Candès and Tao (2008). Of particular interest are penalization methods by miximizing the log-likelihood function with a penalty for non-sparsity. Most approaches use non-differentiable penalization. See for example the paper of Candès and Plan (2008) for a very elegant analysis of the l_1-type penalization in the context of linear variable selection. On the other hand, only a few attempts have been made to use this type of penalization for more complex models than the linear model; for some recent progress, see Koh, Kim, and Boyd (2007) for the case of logistic regression; and Khalili and Chen (2007) for mixture models. However, the use of non-differentiable penalties can be reasonnably expected to be extended to more complex nonlinear models, the mixture model being one of the most popular instance.

The goal of the present paper is to propose new extensions of the EM algorithm that incorporate a non-differentiable penalty at each step. Following previous work of the first two authors, we develop a Kullback Proximal framework for understanding the EM-iterations and prove optimality for the cluster points of the methods using nonsmooth analysis tools. A key additional feature in our study is the consideration of Space Alternating extensions of EM and Kullback Proximal Point (KPP) methods. Such component-wise versions of EM-type algorithms can enjoy nice theoretical properties with respect to 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 our paper is that any cluster point of the Space Alternating KKP method satisfies a nonsmooth Karush-Kuhn-Tucker necessary optimality equation.

The paper is organized as follows. In section 2 we present the penalized EM-type methods that we call Penalized Kullback Proximal Point methods. In Section 3, our main asymptotic results are presented. In Section 4, two examples are presented. The first is 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). The second example is taken from Ting, Raich and Hero, (2007) in which the theoretical issue of convergence was not addressed. New asymptotic results follow from the theory developped in Section 3.

2 The EM algorithm and its Kullback proximal generalizations

The problem of maximum likelihood (ML) estimation consists of finding a solution of the form

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

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

$$l_y(\theta) = \log g(y; \theta), \tag{2}$$

defined 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 consists of introducing a complete data vector X with density f. Consider the conditional density function $k(x|y;\bar{\theta})$ of X given y

$$k(x|y;\bar{\theta}) = \frac{f(x;\bar{\theta})}{g(y;\bar{\theta})}. \tag{3}$$

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 values, 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 in a recursive manner 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

$$\theta^{k+1} = \mathrm{argmax}_{\theta \in D_\Phi} \left\{ \Phi(\theta) - \frac{\beta_k}{2} \|\theta - \theta^k\|^2 \right\}. \tag{4}$$

The influence of the quadratic penalty $\frac{1}{2}\|\theta - \theta^k\|^2$ is controled by using a sequence of positive parameters $\{\beta_k\}$. Rockafellar (1976) showed that superlinear convergence of this method is obtained when the sequence $\{\beta_k\}$ converges towards zero. A relationship between Proximal Point algorithms and the EM algorithm was discovered in Chrétien and Hero (2000) (see also Chrétien and Hero (2007) for details). We review the EM analyogy to PPK methods. Assume that the family of conditional densities $\{k(x|y;\theta)\}_{\theta \in \mathbb{R}^p}$ is regular in the sense of Ibragimov and Khasminskii (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:

$$I_y(\theta, \bar{\theta}) = \mathsf{E}\big[\log \frac{k(x|y,\bar{\theta})}{k(x|y;\theta)}|y;\bar{\theta}\,\big]. \tag{5}$$

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 Kullback Proximal Point algorithm is defined by

$$\theta^{k+1} = \text{argmax}_{\theta \in D_\Phi} \left\{ \Phi(\theta) - \beta_k I_y(\theta, \bar{\theta}) \right\}. \tag{6}$$

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

Proposition 2.1 [Chrétien and Hero (2000) Proposition 1]. *The EM algorithm is a special instance of the Kullback-proximal algorithm with $\beta_k = 1$, 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 KKP, the parameter space is decomposed into subspaces $\Theta_r = \Theta \cap \mathcal{S}_r$, $r = 1, \ldots, R$ 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, our Penalized Proximal Point Algorithm is defined as follows.

Definition 2.1 *Let ψ: $\mathbb{R}^p \mapsto \mathcal{S}_1 \times \cdots \times \mathcal{S}_R$ be a continuously differentiable mapping. 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 possibly nonsmooth 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*

$$\theta^{k+1} = \text{argmax}_{\theta \in \Theta_{k-1(\text{mod } R)+1} \cap D_l \cap D_{I,\theta^k}} l_y(\theta) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta)) - \beta_k I_y(\theta, \theta^k). \tag{7}$$

The standard Kullback-Proximal Point algorithms as defined in Chrétien and Hero (2007) is obtained as special case by selecting $R = 1$, $\Theta_1 = \Theta$, $\lambda = 0$.

In most cases, the mappings ψ_r will simply be the projection onto the subspace Θ_r, $r = 1, \ldots, R$.

2.2 Notations and assumptions

The notation $\|\cdot\|$ will be used to denote the norm on any previously defined space without more precision. The space on which it is the norm should be obvious from the context. For any bivariate function Φ, $\nabla_1 \Phi$ will denote the gradient with respect to the first variable. In the remainder of this paper we will make the following assumptions. For a locally Lipschitz function f, $\partial f(x)$ denotes the Clarke subdifferential of f at x; this notion is recalled in the Appendix.

Assumptions 1 (i) l_y is differentiable on D_l and $l_y(\theta) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta))$ converges to $-\infty$ whenever $\|\theta\|$ tends to $+\infty$.
(ii) the projection of D_I onto the first coordinate is a subset of D_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

$$\psi_r(\theta_r + \epsilon d) = \psi_r(\theta_r) \tag{8}$$

for all $d \in \mathcal{S}_r^\perp$ and $\epsilon > 0$ sufficiently small so that $\theta_r + \epsilon d \in \Theta$, $r = 1, \ldots, R$.

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

Assumptions 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 (5) 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.

Assumptions 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

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

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 for positive real numbers only 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

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

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 Csiszàr (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.

Assumptions 4 *The Kullback proximal iteration (7) is well defined, i.e. there exists at least one maximizer of (7) 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 solve the inclusion $0 \in \nabla l_y(\theta) - \lambda \partial p_n(\psi(\theta)) - \beta_k \nabla I_y(\theta, \theta^k)$ in order to prove in practice that the solution is unique. Then assumption 1(i) is sufficient to conclude that we actually have a maximizer.

3 Asymptotic properties of the Kullback-Proximal iterations

3.1 Basic properties of the penalized Kullback proximal algorithm

Under Assumptions 1, we state some basic properties of the penalized Kullback Proximal Point Algorithm. The most basic is the monotonicity of the penalized likelihood values taken by successive iterates 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 case where $\lambda = 0$ and their generalizations to the present context is straightforward.

We start with the following monotonicity result.

Lemma 3.1 *For any iteration $k \in \mathbb{N}$, the sequence $(\theta^k)_{k\in\mathbb{N}}$ satisfies*

$$l_y(\theta^{k+1}) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta^{k+1})) - (l_y(\theta^k) - \sum_{r=1}^{R} \lambda_r p_n \psi_r(\theta^k))) \geq \beta_k I_y(\theta^k, \theta^{k+1}) \geq 0. \tag{9}$$

Lemma 3.2 *The sequence $(\theta^k)_{k\in\mathbb{N}}$ is bounded.*

The next lemma will also be useful and its proof in the case where $\lambda = 0$ is given in Chrétien and Hero (2007) Lemma 2.4.3. The generalization to $\lambda > 0$ is also straightforward.

Lemma 3.3 *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 observed in pratice and onto which stopping criteria often rely is that the distance between two successive iterates decreases to zero. This fact was established in Chrétien and Hero (2007) and does not depend on the fact that $\lambda = 0$.

Proposition 3.1 [Chrétien and Hero (2007) 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 $\mathbb{R}^p_+$, the fact that $\lim_{k\to+\infty} I_y(\eta^k,\theta^k) = 0$ implies $\lim_{k\to+\infty} |t_{ij}(\eta^k) - t_{ij}(\theta^k)| = 0$ for all i,j such that $\alpha_{ij} \neq 0$.

(ii) If 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 some modified Karush-Kuhn-Tucker type conditions on the domain of the log-likelihood function. For notational convenience, we define

$$F_\beta(\theta,\bar{\theta}) = l_y(\theta) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta)) - \beta I_y(\theta,\bar{\theta}). \tag{10}$$

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.

Theorem 3.1 *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. Let θ^* be a cluster point of the Space Alternating Penalized Kullback-proximal sequence of Definition 2.1. Assume that all the functions t_{ij} are differentiable at θ^*. If θ^* lies in the interior of D_l, then θ^* is a stationary point of the 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 exactly analog to the proof of Theorem 3.2.1 in Chrétien and Hero (2007). In particular, we have

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

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 (11).

B. Assume that $R > 1$ and let $(x^{\sigma(k)})_{k\in\mathbb{N}}$ be a subsequence of iterates of (7) 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 next index greater than $\sigma(k)$ such that $(\sigma(k) - 1) \pmod{R} + 1 = r$. Using the fact that t is injective on every Θ_r, $r = 1, \ldots, R$, Lemma 3.3 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 2.1 of the Space Alternating Penalized Kullback Proximal iterations,

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

Therefore,

$$\begin{aligned} F_{\beta^*}(\theta^{\sigma(k)+1}, \theta^{\sigma(k)}) \quad & -(\beta_{\sigma_r(k)} - \beta^*) I_y(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}) \geq \\ & F_{\beta^*}(\theta, \theta^{\sigma_r(k)}) - (\beta_{\sigma_r(k)} - \beta^*) I_y(\theta, \theta^{\sigma(k)}). \end{aligned} \tag{11}$$

Continuity of F_β follows directly from the proof of Theorem 3.2.1 in Chrétien and Hero (2007) where in this proof $\sigma(k)$ has to be replaced by $\sigma_r(k)$). This implies that

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

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 any 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 (12) 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 3.1, take d in $\mathbb{R}^p$ and decompose d as $d = d_1 + \cdots + d_R$ with $d_r \in \mathcal{S}_r$. Then, equation (12) implies that the directional derivatives satisfy

$$F'_{\beta^*}(\theta^*, \theta^*; d_r) \leq 0 \tag{12}$$

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 (12), that

$$F'_{\beta^*}(\theta^*, \theta^*; d) = \sum_{r=1}^R F'_{\beta^*}(\theta^*, \theta^*; d_r) \leq 0. \tag{13}$$

Therefore, using characterization (46) in the Appendix of the subdifferential, the desired result follows. □

We now, consider the case where Assumptions 3 hold.

Theorem 3.2 *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 I^* denote the index of the active constraints at θ^*, i.e. $I^* = \{(i,j)$ s.t. $t_{i,j}(\theta^*) = 0$. If θ^* lies in the interior of D_l, then θ^* satisfies the following property: there exists a family of subsets $I_r^* \subset I^*$ and a set of real numbers λ_{ij}, $(i,j) \in \mathcal{I}_r^*$, $r = 1, \ldots, R$ such that

$$0 \in \nabla l_y(\theta^*) - \sum_{r=1}^{R} \lambda_r \partial p_n(\psi_r(\theta^*)) + \sum_{r=1}^{R} \sum_{(i,j)\in\mathcal{I}_r^{**}} \lambda_{ij}^* \mathrm{P}_{\mathcal{S}_\mathrm{r}}(\nabla t_{ij}(\theta^*)). \tag{14}$$

Remark 3.1 *The condition (14) ressembles the traditional Karush-Kuhn-Tucker conditions of optimality but are in fact weaker since the vector*

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

may not 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 3.2. Let $\Phi_{ij}(\theta, \bar{\theta})$ denote the bivariate function defined by

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

As in the proof of Theorem 3.1, let $(x^{\sigma(k)})_{k\in\mathbb{N}}$ be a subsequence of iterates of (7) 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 $(\sigma_r(k)-1)(\mathrm{mod}\, R)+1 = r$. Using the fact that t is injective on every Θ_r, $r = 1, \ldots, R$, Lemma 3.3 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

$$\begin{aligned} 0 = \quad & \mathrm{P}_{\mathcal{S}_\mathrm{r}}(\nabla l_y(\theta^{\sigma(k)+1})) - \lambda_r g_r^{\sigma_r(k)+1} + \beta_{\sigma_r(k)}\Big(\textstyle\sum_{ij} \alpha_{ij}(y_j)\mathrm{P}_{\mathcal{S}_\mathrm{r}}(\nabla t_{ij}(\theta^{\sigma_r(k)+1})) \\ & \Phi_{ij}(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}) + \textstyle\sum_{ij} \alpha_{ij}(y_j) t_{ij}(\theta^{\sigma_r(k)+1})\mathrm{P}_{\mathcal{S}_\mathrm{r}}(\nabla_1 \Phi_{ij}(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}))\Big) \end{aligned} \tag{15}$$

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 (2007), gives that for all (i,j) such that $\alpha_{ij}(y_j) \neq 0$

$$\lim_{k\to+\infty} t_{ij}(\theta^{\sigma_r(k)+1})\nabla_1 \Phi_{ij}(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}) = 0. \tag{16}$$

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

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

is linearly independent for k sufficiently large. For such k, we may thus rewrite equation (15) as

$$\begin{aligned} 0 = \quad & \mathrm{P}_{\mathcal{S}_{\mathrm{r}}}(\nabla l_y(\theta^{\sigma_r(k)+1})) - \lambda_r g_r^{\sigma_r(k)+1} + \beta_{\sigma_r(k)}\Big(\textstyle\sum_{(i,j)\in\mathcal{I}_r^*}\pi_{ij}^{\sigma_r(k)+1}(y_j) \\ & \mathrm{P}_{\mathcal{S}_{\mathrm{r}}}(\nabla t_{ij}(\theta^{\sigma_r(k)+1})) + \textstyle\sum_{ij}\alpha_{ij}(y_j)t_{ij}(\theta^{\sigma_r(k)+1})\mathrm{P}_{\mathcal{S}_{\mathrm{r}}}(\nabla_1\Phi(\theta^{\sigma_r(k)+1},\theta^{\sigma_r(k)}))\Big). \end{aligned} \tag{17}$$

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 $(g_r^{\sigma_r(\gamma(k))+1})_{k\in\mathbb{N}}$ with limit g_r^*. Now, using Equation (16), this last equation implies that the $\{\pi_{(i,j)\in\mathcal{I}_r^*}^{\sigma_r(\gamma(k))+1}(y_j)\}_{(i,j)\in\mathcal{I}_r^*}$ converges to the coordinates of a vector in the linearly independent family $\{\mathrm{P}_{\mathcal{S}_{\mathrm{r}}}(\nabla t_{ij}(\theta^*))\}_{(i,j)\in\mathcal{I}_r^*}$ which concludes the proof. □

This claim allows us to finish the proof of Theorem 3.2. 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 equation (15) that

$$0 = \mathrm{P}_{\mathcal{S}_{\mathrm{r}}}(\nabla l_y(\theta^*)) - \lambda_r g_r^* + \beta^*\Big(\sum_{(i,j)\in\mathcal{I}_r^{**}}\pi_{ij}^*\mathrm{P}_{\mathcal{S}_{\mathrm{r}}}(\nabla t_{ij}(\theta^*))\Big). \tag{18}$$

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 (18), we obtain

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

Moreover, since $\Phi_{ij}(\theta^{\sigma_r(k)+1},\theta^{\sigma_r(k)})$ tends to zeros if $(i,j)\notin I^*$, i.e. if (i,j) is not active, passing to the limit in equation equation (15) implies that

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

for I_r^{**} being the subset of active indices of I_r^*, i.e. $I_r^{**} = I_r^* \cap I^*$. 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

$$0 \in \nabla l_y(\theta^*) - \sum_{r=1}^R \lambda_r\partial p_n(\psi_r(\theta^*)) + \beta^*\sum_{r=1}^R\sum_{(i,j)\in\mathcal{I}_r^{**}}\pi_{ij}^*\mathrm{P}_{\mathcal{S}_{\mathrm{r}}}(\nabla t_{ij}(\theta^*)). \tag{19}$$

which establishes Theorem 3.2 once we take $\lambda^*_{ij} = \lambda^* \pi^*_{ij}$. □

The result (19) can be refined to the classical Karush-Kuhn-Tucker type condition under additional conditions such as stated in the next corrolary.

Corollary 3.1 *If in addition to the assumptions of Theorem 3.2 we assume that either* $\mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)) = \nabla t_{ij}(\theta^*)$ *or* $\mathrm{P}_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)) = 0$ *for all* $(i,j) \in I^*$, *i.e. such that* $t_{ij}(\theta^*) = 0$, *then there exists a set of subsets* $I^*_r \subset I^*$ *and a family of real numbers* λ_{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 Examples

In this section, we show two applications of the penalized KKP algorithm (7) to enforce sparsity in a multiple parameter estimator. We first consider the finite mixtures of linear regression models using the SCAD penalty for variable selection as studied in Khalili and Chen (2007). We then address a problem in sparse image reconstruction studied in Ting, Raich and Hero (2007).

4.1 Variable selection in finite mixtures of regression models

.

Until quite recently, variable selection in regression models was performed using penalization strategies in the maximum likelihood framework, e.g. using AIC, Akaike (1973) and BIC, Schwarz (1978) for instance. The main drawback of these approaches is the combinatorial explosion of the set of possible models in the case where the number of variables is large. Recently, new approaches have been proposed that select the subsets of variables without enumeration of all subsets of a given size. Most such methods use l_1-type penalties of likelihood functions as in the LASSO, Tibshirani (1996). The recent Dantzig selector of Candès and Tao (2007) also uses the l_1 penalty.

Computation of the maximizers of the penalized likelihood 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 functions induce line-search problems. In certain cases, the EM algorithm or its generalizations, 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), is 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.

$$Y \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(X^t \beta_k, \sigma^2), \tag{20}$$

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 is established. The purpose of this section is to show that the Space Alternating KPP EM genaralization is easily implemented and that stationarity of the cluster points is garanteed by the theoretical analysis of Section 3.

The SCAD penalty, studied in Khalili and Chen (2007) is a modification of the l_1 penalty which is given by

$$p_n(\beta_1, \ldots, \beta_K) = \sum_{k=1}^{K} \pi_k \sum_{j=1}^{P} p_{\gamma_{nk}}(\beta_{k,j}) \tag{21}$$

where p_{nk} is specified by

$$p'_{\gamma_{nk}}(\beta) = \gamma_{nk} \sqrt{n} 1_{\sqrt{n}|\beta| \leq \gamma_{nk}} + \frac{\sqrt{n}(a\gamma_{nk} - \sqrt{n}|\beta|)_+}{a-1} 1_{\sqrt{n}|\beta| > \gamma_{nk}} \tag{22}$$

for β in $\mathbb{R}$.

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

$$l_c(\beta_1, \ldots, \beta_K, \sigma^2) = \sum_{i=1}^{n} \log(\pi_{z_i}) - \frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - x_i^t \beta_{z_i})^2}{2\sigma^2}. \tag{23}$$

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

$$Q(\theta, \bar{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}(\bar{\theta}) \left[\log(\pi_k) - \frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - x_i^t \beta_k)^2}{2\sigma^2} \right] - p_n(\beta_1, \ldots, \beta_K) \tag{24}$$

where

$$t_{ik}(\theta) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_i - X\beta_k)^2}{2\sigma^2} \Big)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_i - X\beta_l)^2}{2\sigma^2} \Big)}. \tag{25}$$

The computation of this Q-function corresponds to the E-step. Due 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

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n} t_{ik}(\bar{\theta}) \quad k = 1, \ldots, K, \tag{26}$$

although this is the choice made in Khalili and Chen (2007) in their implementation. Moreover, optimizing jointly in 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 approach consisting of successively optimizing with respect to the π_k's and alternatively with respect to each vector β_k. Optimization with respect to the π_k's can be easily performed using any standard optimization routine and optimization with respect to the β_k's requires a specific algorithm for optimization of non-differentiable functions 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 (5). 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

$$I_y(\theta, \bar{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}(\bar{\theta}) \log\Big(\frac{t_{ik}(\bar{\theta})}{t_{ik}(\theta)}\Big). \tag{27}$$

We have $R = K + 1$ subsets of variables with respect to which optimization must be performed successively. All components of assumptions 1 and 3 are trivially satisfied for this model except for Assumption 3 (iv). However Assumption 3 (iv) is proved in 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

$$\frac{\partial t_{ik}}{\partial \beta_{jl}}(\theta) = 0 \tag{28}$$

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

$$\frac{\partial t_{ik}}{\partial \sigma^2}(\theta) = 0, \tag{29}$$

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 3.1 applies.

We now turn to some experiments on the real data set (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 comparision to the three components model.

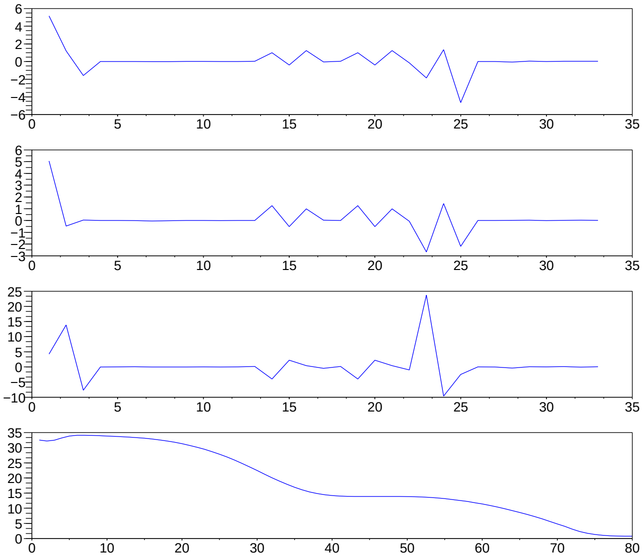

Figure 1: Baseball data of Khalili and Chen (2007). This experiment is performed with the plain EM. The parameters are $\gamma_{nk} = .1$ and $a = 10$. The first plot is the vector β obtained for the single component model. The second (resp. third) plot is the vector of the optimal β_1 (resp. β_2). The fourth plot is the euclidean distance to the optimal θ^* versus iteration index. The starting value of π_1 was .3

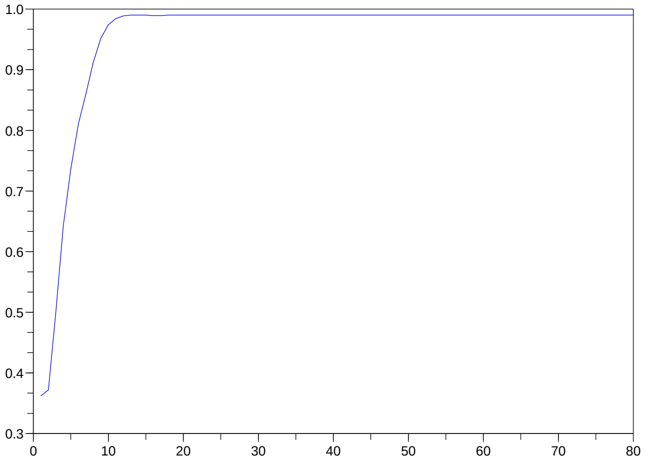

Figure 2: Baseball data of Khalili and Chen (2007). This experiment is performed with the plain EM. The parameters are $\gamma_{nk} = 5$ and $a = 10$. The plot shows the probability π_1 of the first component versus iteration index. The starting value of π_1 was .3

Here, two algorithms are compared: the approximate EM using (26) and the plain EM using the optim subroutines. The results for $\gamma_{nk} = 1$ and $a = 10$ are given in Figures 1.

The experiments shown in Figure 1 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 Figure 2 (we fixed a maximum upper bound equal to .99 in order to avoid numerical problems). Figure 3 demonstrates that this behavior is not reproduced by the approximate EM algorithm chosen by Khalili and Chen (2007).

4.2 l_2/l_1 penalized EM for sparse image reconstruction

.

In this section, we justify the convergence of the l_1-penalized EM algorithm of Ting, Raich and Hero (2007).

The image will be denoted by $\theta \in \mathbb{R}^p$ and the main problem is to reconstruct this image from a set of noisy measurements $y \in \mathbb{R}^N$, e.g. a set of noisy projections of the image. We assume that the image θ is sparse, i.e. the number of

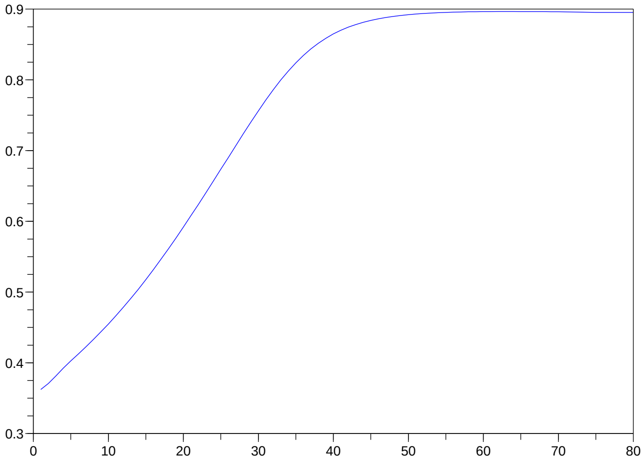

Figure 3: Baseball data of Khalili and Chen (2007). This experiment is performed with the approximate EM. The parameters are $\gamma_{nk} = 5$ and $a = 10$. The plot shows the probability π_1 of the first component versus iteration index. The starting value of π_1 was .3

nonzero pixels is small compared to the size of θ. The projection y is obtained from θ via a linear transformation with additive gaussian white noise

$$y = H\theta + w \tag{30}$$

where $w \sim \mathcal{N}(0, \sigma^2 I)$ and where $H \in \mathbb{R}^{N\times d}$.

One very successfull method for reconstruction of sparse signals is the LASSO. This method was first proposed in Alliney and Ruzinsky (1994) and then further developped in Tibshirani (1996). The LASSO estimator is given by

$$\hat{\theta}(y, \beta) = \text{argmin}_\theta \|H\theta - y\|_2^2 + \beta\|\theta\|_1 \tag{31}$$

where β is a regularization parameter that can be tuned manually. We will denote by $p(y \mid \theta)$ the density of Y whose log-likelihood is $-\frac{\|H\theta-y\|_2^2}{2\sigma^2}$.

In Ting, Raich and Hero (2007), the authors propose a more general framework allowing for more general possible penalties and in particular more tractable forms of the LAZE prior of Johnstone and Silverman (2004). The prior incorporated in the Ting Raich and Hero model is as follows. For each $i = 1, \dots, d$ define the random variables $\tilde{\theta}_i$ and I_i such that $\theta_i = \tilde{\theta}_i I_i$ with the following density

$$I_i = \begin{cases} 0 \text{ with probability } (1-w) \\ 1 \text{ with probability } w \end{cases} \tag{32}$$

$$p(\tilde{\theta}_i \mid I_i) = \begin{cases} g(\tilde{\theta}_i) \text{ if } I_i = 0 \\ \gamma(\tilde{\theta}_i; a) \text{ if } I_i = 1 \end{cases} \tag{33}$$

where $g(\cdot)$ is some p.d.f. that will be specified later on and where it is assumed that the sequence $\{(\tilde{\theta}_i, I_i)\}$ is i.i.d. The variables $\{I_i\}$ play the role of the delta function in the standard LAZE prior. The Maximum A Posteriori (MAP) reconstruction problem is then given by

$$(\hat{\tilde{\theta}}, \hat{I}, \hat{w}, \hat{a}) = \mathrm{argmax}_{\tilde{\theta}, I, w, a} \log p(\tilde{\theta}, I \mid Y, w, a). \tag{34}$$

Let $\mathcal{I}_1 = \{i \mid I_i = 1\}$ and $\mathcal{I}_0 = \{i \mid I_i = 0\}$. The MAP problem is equivalent to

$$\max -\tfrac{\|H\theta - Y\|_2^2}{2\sigma^2} \quad \begin{aligned} &+(M - \mathrm{Card}(\mathcal{I}_1)) \log(1-w) + \mathrm{Card}(\mathcal{I}_1) \log w \\ &+ \textstyle\sum_{i \in \mathcal{I}_1} \log \left(\tfrac{1}{2} a e^{-a|\tilde{\theta}_i|}\right) + \textstyle\sum_{i \in \mathcal{I}_0} \log g(\tilde{\theta}_i). \end{aligned} \tag{35}$$

Maximization is performed in a block coordinate-wise fashion, handling maximization over (w, a) and $(\tilde{\theta}, I)$ by alternating between them as described in Ting, Raich and Hero (2007) Section IV, Algorithm 1.

Two options are considered MAP1 and MAP2; we refer to Ting, Raich and Hero (2007) for more details. We only present MAP1 since our results readily apply to MAP2 once case MAP1 has been justified. Set $g(x) = \gamma(x, a)$ where $\frac{1}{2} a e^{-a|x|}$ is the Laplacian p.d.f. Maximization over (w, a) is easily obtained by

$$\hat{a} = \frac{M}{\|\hat{\tilde{\theta}}\|_1} \text{ and } \hat{w} = \frac{\mathrm{Card}(\hat{I}_1)}{M}. \tag{36}$$

The M step with respect to $(\tilde{\theta}, I)$ is obtained by applying the EM strategy. For this purpose, introduce the complete data model

$$Z = \theta + w_1 \tag{37}$$

$$Y = HZ + w_2 \tag{38}$$

where w_1 and w_2 are gaussian white noises with $w_1 \sim \mathcal{N}(0, \alpha^2 I)$ and $w = Hw_1 + w_2$ which implies that $w_2 \sim \mathcal{N}(0, \sigma^2 I - \alpha^2 HH^t)$. This representation is very interesting since it allows to decompose the problem in two tasks, the first being the one of deconvolving, the second of denoising. In this model, the hidden data is $z = \theta + \alpha w$. Assuming (w, a) fixed, the E-step of the algorithm is obtained as follows from Figueiredo and Nowak (2003), Section V.B. We can write the complete likelihood $f_{Y,Z|\theta} = f_{Y|Z,\theta} f_{Z|\theta}$. First, given Z, Y is independent of θ. Next, we have

$$\log f_{Z|\theta}(Z) = \frac{\theta^t \theta - 2\theta^t Z}{2\alpha^2} + C \tag{39}$$

where C is a constant dependent on θ. Hence, finding the function $Q(\theta, \bar{\theta})$, is equivalent to replacing Z by its conditional expectation given Y and $\bar{\theta}$. This conditional expectation is a deconvolution step and is given by

$$\mathrm{E}\Big[Z \mid Y, \bar{\theta}\Big] = \bar{\theta} + \frac{\alpha^2}{\sigma^2} H^t (Y - H\bar{\theta}). \tag{40}$$

therefore, the E-step corresponds to the computation of

$$(\text{ E-step }) \qquad Q(\theta, \bar{\theta}) = -\frac{1}{2\alpha^2} \|\theta - \mathrm{E}\Big[Z \mid Y, \bar{\theta}\Big]\|^2 - \frac{\theta^t\theta - 2\theta^t \mathrm{E}\Big[Z \mid Y, \bar{\theta}\Big]}{2\alpha^2} + C. \tag{41}$$

Now, recalling that $\theta_i = \tilde{\theta}_i I_i$, the M step is

$$(\text{ M-step }) \qquad \theta_i = \begin{cases} T_{hy}(\mathrm{E}[Z \mid Y, \bar{\theta}]; a\alpha^2 + \sqrt{2\alpha^2 \log(\frac{1-w}{w}}, a\alpha^2) & \text{if } 0 < w \le \frac{1}{2} \\ T_s(\mathrm{E}[Z \mid Y, \bar{\theta}]; a\alpha^2) & \text{if } \frac{1}{2} < w \le 1 \end{cases} \tag{42}$$

where T_{hy} denotes the hybrid soft thresholding function

$$T_{hy}(x, t_1, t_2) = (x - \mathrm{sign}(x) t_2 1_{|x| > t_1}) \tag{43}$$

and T_s denotes the usual soft thresholding function

$$T_s(x, t) = T_{hy}(x, t, t). \tag{44}$$

Finally, we obtain an EM algorithm which satisfies the assumptions of Theorem 3.1 above and the asymptotic properties asserted by Theorem 3.1 hold.

5 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: $\mathbb{R}^p \mapsto \mathbb{R}$ always has a generalized directional derivative $f^\circ(\theta, \omega)$: $\mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ in the sense given by Clarke, i.e.

$$f^\circ(\theta, \omega) = \lim\ \sup_{\eta \in \mathbb{R}^p \to \theta,\ t \downarrow 0} \frac{f(\eta + t\omega) - f(\eta)}{t}. \tag{45}$$

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

$$\partial f(\theta) = \{\eta \mid f^\circ(\theta, \omega) \ge \eta^t \omega,\ \forall \omega\}. \tag{46}$$

Proposition 5.1 *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 5.1 *A set-valued map Φ is said to be outer-semicontinuous if its graph*

$$\text{graph}\,\Phi = \{(\theta, g) \mid g \in \Phi(\theta)\} \tag{47}$$

is closed, i.e. if for any sequence $\text{graph}\Phi \supset (\theta_n, g_n) \to (\theta^*, g^*)$ *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

$$0 \in \partial f(\theta). \tag{48}$$

Consider now the problem

$$\sup_{\theta \in \mathbb{R}^p} f(\theta) \tag{49}$$

subject to

$$g(\theta) = [g_1(\theta), \ldots, g_m(\theta)]^t \geq 0 \tag{50}$$

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

$$0 \in \partial f(\theta) + \sum_{j=1}^{m} u_j \partial g_j(\theta). \tag{51}$$

Convex functions are in particular locally Lipschitz and the notions of subdifferential can the Clarke subdifferential is well defined. The main references on their particular properties are Rockafellar (1970) and Hiriart-Urruty and Lemaréchal (1993).

References

[Akaike (1973)] H. Akaike (1973). Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, eds. B.N. Petrox and F. Caski, Budapest: Akademiai Kiado, p.267.

[Alliney and Ruzinsky (1994)] S. Alliney and S. A. Ruzinsky (1994). An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian estimation. IEEE Trans. Signal Processing, vol. 42, no. 3, pp. 618–627.

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

[Berlinet and Roland (2007)] A. Berlinet, Ch. Roland (2007). Acceleration schemes with application to the EM algorithm. Comput. Statist. Data Anal., 51, 3689-3702.

[Biernacki and Chrétien (2003)] . C. Biernacki and S. Chrétien (2003). Degeneracy in the Maximum Likelihood Estimation of Univariate Gaussian Mixtures with EM. *Statistics and Probability Letters*, 61, 373-382.

[Candès and Plan (2007)] E. Candès and Y. Plan (2007). Near-ideal model selection by L1 minimization. Technical Report, California Institute of Technology.

[Candès and Tao (2007)] E. Candès and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics to be published.

[Celeux *et al.* (2001)] G. Celeux, S. Chrétien, F. Forbes and A. Mkhadri (2001). A Component-Wise EM Algorithm for Mixtures. Journal of Computational and Graphical Statistics, vol. 10, no. 4, 697-712.

[Chrétien and Hero (2000)] S. Chrétien and A. Hero (2000). Kullback proximal algorithms for maximum-likelihood estimation. Information-theoretic imaging. IEEE Trans. Inform. Theory 46, no. 5, 1800–1810.

[Chrétien and Hero (2008)] S. Chrétien and A. Hero (2008). On EM algorithms and their proximal generalizations. ESAIM P&S 12, 308–326

[Clarke (1990)] F. Clarke (1990). *Optimization and Nonsmooth Analysis*, Vol. 5, Classics in Applied Mathematics, SIAM.

[Cover and Thomas (1987)] T. Cover and J. Thomas (1987). *Elements of Information Theory*, Wiley, New York.

[Delyon, Lavielle and Moulines (1999)] B. Delyon, M. Lavielle, Marc and E. Moulines (1999). Convergence of a stochastic approximation version of the EM algorithm. Ann. Statist. 27, no. 1, 94–128.

[Dempster, Laird, and Rubin (1977)] A. P. Dempster, N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Ser. B*, vol. 39, no. 1, pp. 1–38.

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

[Fan and Li (2002)] J. Fan and R. Li (2002). Variable selection for Cox's proportional hazards model and frailty model. The annals of statistics, 30, 74–99.

[Fessler and Hero (1994)] J. A. Fessler, and A. O. Hero (1994). Space-alternating generalized expectation-maximization algorithm. IEEE Trans. Signal Processing, vol. 42, no. 10, pp. 2664–2677.

[Figueiredo and Nowak (2003)] M. A. T. Figueiredo and R. D. Nowak (2003). An EM algorithm for wavelet-based image restoration. IEEE Trans. Image Processing, vol. 12, no. 8.

[Friedmand and Popescu (2003)] J. Friedmand and B.E. Popescu (2003). Importance Sampled Learning Ensembles", Journal of Machine Learning Research.

[Green (1990)] P. J. Green (1990). On the use of the EM algorithm for penalized likelihood estimation. *J. Royal Statistical Society, Ser. B*, vol. 52, no. 2, pp. 443–452.

[Hero and Fessler (1995)] A. O. Hero and J. A. Fessler (1995). Convergence in norm for alternating expectation-maximization (EM) type algorithms. *Statistica Sinica*, vol. 5, no. 1, pp. 41–54.

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

[Hunter and Lange (2004)] D. R. Hunter and K. Lange (2004). A Tutorial on MM Algorithms. The American Statistician, Vol. 58.

[Hunter and Li (2005)] D. R. Hunter and R. Li (2005). Variable selection using MM algorithms. The Annals of Statistics, 33, 1617–1642.

[Ibragimov and Has'minskii (1981)] I. A. Ibragimov and R. Z. Has'minskii (1981). *Statistical estimation: Asymptotic theory*, Springer-Verlag, New York.

[Johnstone and Silverman (2004)] I. M. Johnstone and B. W. Silverman (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. The Annals of Statistics, vol. 32, no. 4, pp. 1594–1649.

[Khalili and Chen (2007)] A. Khalili and J. Chen (2007). Variable Selection in Finite Mixture of Regression Models. Journal of the American Statistical Association, Volume 102, Number 479, pp. 1025-1038.

[Koh, Kim and Boyd (2007)] K. Koh, S.-J. Kim, and S. Boyd (2007). An Interior-Point Method for Large-Scale l1-Regularized Logistic Regression. Journal of Machine Learning Research, 8:1519-1555.

[Kuhn and Lavielle (2004)] E. Kuhn, Estelle and M. Lavielle (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. ESAIM Probab. Stat. 8, 115–131

[Lange (1995)] K. Lange (1995). A quasi-newtonian acceleration of the EM algorithm. *Statistica Sinica*, vol. 5, no. 1, pp. 1–18.

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

[McLachlan and Peel (2000)] G.J. McLachlan and D. Peel (2000). *Finite Mixture Models*. Wiley

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

[Minty (1962)] G. J. Minty (1962). Monotone (nonlinear) operators in Hilbert space. *Duke Math. Journal*, vol. 29, pp. 341–346.

[Moreau (1965)] J. J. Moreau (1965). Proximité et dualité dans un espace Hilbertien. *Bull. Soc. Math. France*, vol. 93, pp. 273–299.

[Rockafellar (1970)] R. T. Rockafellar (1970). *Convex Analysis*, Vol. 28 of Princeton Math. Series, Princeton Univ. Press.

[Rockafellar (1976)] R. T. Rockafellar (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, vol. 14, pp. 877–898.

[Rockafellar and Wets (2004)] R. T. Rockafellar and R. J. B. Wets (2004). *Variational Analysis*. Vol. 317 Grundlehren der mathematischen Wissenschaften, Springer.

[Schwarz (1978)] G. Schwarz (1978). Estimating the dimension of a model. Annals of Statistics, vol. 6, pp. 461–464.

[Tibshirani (1996)] R. Tibshirani (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, Series B, vol. 58, no. 1, pp. 267–288.

[Ting, Raich and Hero (2007)] M. Ting, R. Raich and A. O. Hero (2007). Empirical approaches to sparse image reconstruction. Submitted to IEEE Trans. Image Processing.

[Varadhan and Roland (2007)] R. Varadhan and Ch. Roland (2007). Simple and Globally-Convergent Numerical Methods for Accelerating Any EM Algorithm. Scandinavian Journal of Statistics (to appear).

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