A Connection Between Half-Quadratic Criteria and EM Algorithms

Frédéric Champagnat and Jérôme Idier

***Abstract*—Iteratively Reweighted Least Squares (IRLS) and Residual Steepest descent (RSD) algorithms of robust statistics arise as special cases of half-quadratic schemes [1]. Here, we adopt a statistical framework and we show that both algorithms are instances of the EM algorithm. The augmented dataset respectively involves a *scale* and a *location* mixture of Gaussians. The sufficient conditions for the construction cover a broad class of already known robust statistics.**

***Index Terms*—EM algorithm, half-quadratic criteria, iteratively reweighted least squares (IRLS), residual steepest descent (RSD), scale mixtures.**

I. Introduction

OUR contribution deals with the minimization of criteria of the form

$$J(\boldsymbol{x}) = J_0(\boldsymbol{x}) + \sum_{m=1}^{M} \phi(u_m) \tag{1}$$

where $u_m = \boldsymbol{v}_m^{\mathrm{t}}\boldsymbol{x} - w_m, \boldsymbol{x} = \mathbf{col}\{x_n\} \in \Gamma \subset \mathbb{R}^N$ and J_0 is a quadratic function. Signal or image restoration applications is our main concern: for instance, define

$$J_0(\boldsymbol{x}) = \frac{1}{2}(z - \mathbf{H}\boldsymbol{x})^{\mathrm{t}}\mathbf{R}(z - \mathbf{H}\boldsymbol{x}) \tag{2}$$

in which case J is a penalized mean square criterion involving a linear observation equation $\boldsymbol{z} = \mathbf{H}\boldsymbol{x} + \text{noise}$, and the sum involving ϕ plays the role of a regularization function. The former setting is generic in edge preserving image restoration [1]–[3]. Another application is robust linear regression (cf. [1], Subsection II-C), in which case $J_0 = 0, \boldsymbol{w} = \mathbf{col}\{w_m\}$ is the observed data and $\mathbf{V} = [\boldsymbol{v}_1 \mid \cdots \mid \boldsymbol{v}_M]$ is the *design matrix*.

II. Half-Quadratic Criteria

Geman *et al.* [2], [3] have shown the existence of *augmented* criteria $K(\boldsymbol{x}, \boldsymbol{b})$ attached to $J(\boldsymbol{x})$, in the following sense:

$$\inf_{\boldsymbol{b}\in B} K(\cdot, \boldsymbol{b}) = J \tag{3}$$

for some set B. Identity (3) implies that J and K share the same *infimum*: $\inf_{\boldsymbol{b},\boldsymbol{x}} K(\boldsymbol{x}, \boldsymbol{b}) = \inf_{\boldsymbol{x}} J(\boldsymbol{x})$. As far as J and K are continuous functions, we have also

$$\hat{\boldsymbol{x}} = \arg\min\nolimits_{\boldsymbol{x}\in\Gamma} J \iff \exists \hat{\boldsymbol{b}}; \\ (\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}) = \arg\min\nolimits_{(\boldsymbol{x},\boldsymbol{b})\in\Gamma\times\bar{B}} K. \tag{4}$$

Thus, computation of $\hat{\boldsymbol{x}}$ may be adressed either through minimization of J, or minimization of K. [2] and [3] consider different augmented criteria

$$K_{\mathrm{GR}}(\boldsymbol{x}, \boldsymbol{b}) = J_0(\boldsymbol{x}) + \sum_m \left(\frac{1}{2} b_m u_m^2 + \psi(b_m)\right) \tag{5}$$

$$K_{\mathrm{GY}}^{\alpha}(\boldsymbol{x}, \boldsymbol{b}) = J_0(\boldsymbol{x}) + \frac{1}{\alpha}\sum_m \left(\frac{1}{2}(u_m - b_m)^2 + \zeta_\alpha(b_m)\right) \tag{6}$$

where the implicit range of summation on m is $1, \ldots, M$. In both cases, (3) holds (for $B_{\mathrm{GR}} = \mathbb{R}_+^M$ and $B_{\mathrm{GY}} = \mathbb{R}^M$, respectively), as a consequence of convex duality properties described in the following propositions [1].

Proposition 1 (GR): Let us consider an even function ϕ, continuous at zero and C^1 on $\mathbb{R} \setminus \{0\}$ such that

$$\phi(\sqrt{\cdot}) \text{ is concave on } \mathbb{R}_+ \tag{7}$$

and let $\psi(b) = \sup_{x\in\mathbb{R}}(\phi(x) - bx^2/2)$. Then

$$\inf_{b\in\mathbb{R}_+}(bx^2/2 + \psi(b)) = \phi(x) \\ \arg\min\nolimits_{b\in\mathbb{R}_+}(bx^2/2 + \psi(b)) = \phi'(x)/x \tag{8}$$

where the value of $\phi'(x)/x$ at $x = 0$ is implicitly obtained by continuity.

Proposition 2 (GY): Let $\alpha > 0$ be such that $(\,\cdot\,)^2/2 - \alpha\phi$ is a convex function, so that $\zeta_\alpha(b) = \sup_{x\in\mathbb{R}}(\alpha\phi(x) - (x-b)^2/2)$ is well defined. Then

$$\inf_{b\in\mathbb{R}}((x-b)^2/2 + \zeta_\alpha(b))/\alpha = \phi(x), \\ \arg\min\nolimits_{b\in\mathbb{R}}((x-b)^2/2 + \zeta_\alpha(b)) = x - \alpha\phi'(x) \tag{9}$$

where the latter identity holds if ϕ is differentiable.

The augmented criteria K_{GR} and K_{GY} are said to be *half-quadratic* (HQ), as quadratic functions of $\boldsymbol{x}$ when auxiliary variables $\boldsymbol{b} = \mathbf{col}\{b_m\}$ are fixed.

In the unconstrained case ($\Gamma = \mathbb{R}^N$), block coordinate descent on K_{GR} or K_{GY}^{α} corresponds to the following iterated scheme.

Manuscript received November 6, 2003; revised February 5, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Petr Tichavsky.

F. Champagnat is with Office National d'Etudes et de Recherches Aérospatiales (ONERA)/DTIM, 92322 Chatillon Cedex, France (e-mail: fchamp@onera.fr).

J. Idier is with Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN), École Centrale de Nantes, 44321 Nantes Cedex 3, France (e-mail: Jerome.Idier@irccyn.ec-nantes.fr).

Digital Object Identifier 10.1109/LSP.2004.833511

- In a first step, K is minimized as a quadratic function of $\boldsymbol{x}$, while $\boldsymbol{b}$ is held constant. The resulting updating equation reads

$$\hat{\boldsymbol{x}}_{\mathrm{GR}} = (\mathbf{H}^{\mathrm{t}}\mathbf{R}\mathbf{H} + \mathbf{V}\mathbf{B}\mathbf{V}^{\mathrm{t}})^{-1}(\mathbf{H}^{\mathrm{t}}\mathbf{R}\boldsymbol{z} + \mathbf{V}\mathbf{B}\boldsymbol{w}) \tag{10}$$
$$\hat{\boldsymbol{x}}^{\alpha}_{\mathrm{GY}} = (\alpha\mathbf{H}^{\mathrm{t}}\mathbf{R}\mathbf{H} + \mathbf{V}\mathbf{V}^{\mathrm{t}})^{-1}(\alpha\mathbf{H}^{\mathrm{t}}\mathbf{R}\boldsymbol{z} + \mathbf{V}(\boldsymbol{w} + \boldsymbol{b})) \tag{11}$$

where $\mathbf{B} = \mathrm{Diag}\{\boldsymbol{b}\}$.

- In a second step, K is minimized as a function of $\boldsymbol{b}$. In both cases, the minimizer has a closed form expression, which is a consequence of convex duality identities (8) and (9), respectively,

$$\hat{\boldsymbol{b}}_{\mathrm{GR}} = \mathbf{col}\{\phi'(u_m)/u_m\} \tag{12}$$
$$\hat{\boldsymbol{b}}_{\mathrm{GY}} = \mathbf{col}\{u_m - \phi'(u_m)\}. \tag{13}$$

It is noted in [1] that (10), (12) [resp. (11) and (13)] identify with an IRLS (resp. RSD) algorithm for minimization of J. In [4], it is shown that the GR-type algorithm can be analyzed as a majorization algorithm. Indeed, it is the same with GY-type algorithms [5]. Modern analyses of EM is also based on a majorization property [6]. In the sequel, we exhibit a tighter link between the half-quadratic approach and the EM approach: we prove that GR-type and GY-type algorithms are both instances of EM.

III. Bayesian Interpretation

Minimization of (1) amounts, in a Bayesian framework, to the maximization of the posterior probability density function (pdf) $p(\boldsymbol{x}\,|\,\boldsymbol{z}) \propto p(\boldsymbol{z}\,|\,\boldsymbol{x})\,p(\boldsymbol{x})$. Provided that $\mathbf{R}$ is invertible, J_0 given by (2) corresponds to a Gaussian density[1] $p(\boldsymbol{z}\,|\,\boldsymbol{x}) \sim \mathcal{N}(\mathbf{H}\boldsymbol{x}, \mathbf{R}^{-1})$, while

$$p_{\boldsymbol{X}}(\boldsymbol{x}) \propto \exp\left\{-\sum_m \phi(u_m)\right\} \tag{14}$$

typically corresponds to the pdf of a Markov random field. It is actually a proper pdf if $\mathbf{V}$ is a rank N matrix and $\int \exp -\phi(u)\mathrm{d}u < \infty$. These conditions are assumed thereafter.

IV. EM Algorithm

In this section, we show that performing MAP estimation within an appropriate EM setting is equivalent to minimizing J using a HQ approach. The standard EM procedure requires to "augment" the original dataset $\boldsymbol{z}$ with auxiliary "hidden" variables, in such a way that these variables would simplify the maximization. The next section exhibits such auxiliary variables.

A. Construction of a Complete Dataset

The construction depends solely on properties of a scalar random variable (RV) U whose pdf is $p_U \propto \exp(-\phi)$. In particular, a crucial step is to examine the possibility that U be Gaussian conditionally to some RV L, with a distribution function (DF) F_L, so that $p_U(u) = \int p(u\,|\,\ell)\,\mathrm{d}F_L(\ell)$ is a Gaussian mixture. For sake of generality, it will not be required that the probability of L have a pdf, that is why DF's will be manipulated instead.[2]

The following definitions introduce two different ways of building such mixture models. In particular, scale mixtures have been pioneered in the early 1970s [8]. More recent applications can be found in areas such as image modeling [9] and stochastic sampling [10], [11].

Definition 1: A scalar RV U is said a *scale mixture of Gaussians* (SMG) if $U = G/\sqrt{L}$ for a couple of independent RV's, (L, G) such that $G \sim \mathcal{N}(0, 1)$, and $L > 0$.

Definition 2: A scalar RV U is said a *location mixture of Gaussians* (LMG) if $U = \sigma G + L$ for some $\sigma > 0$ and a couple of independent RVs (L, G) such that $G \sim \mathcal{N}(0, 1)$.

Both definitions imply that U admits a pdf. Since the properties of U are naturally expressed in terms of p_U, it will be more convenient to rewrite (14) in terms of p_U

$$p_{\boldsymbol{X}}(\boldsymbol{x}) \propto \prod_m p_U(u_m). \tag{15}$$

Lemma 1 (SMG): Let $U = G/\sqrt{L}$ be a SMG, and let

$$\mathrm{d}F(\boldsymbol{x}, \boldsymbol{\ell}) \propto \prod_m p_G(u_m\sqrt{\ell_m})\sqrt{\ell_m}\,\mathrm{d}F_L(\ell_m)\,\mathrm{d}\boldsymbol{x}. \tag{16}$$

Then we have the following properties:

S1) $p_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\boldsymbol{\ell}\in\mathbb{R}_+^M} \mathrm{d}F(\boldsymbol{x}, \boldsymbol{\ell})$.

S2) $F(\boldsymbol{\ell}\,|\,\boldsymbol{x}) = \prod_m F_{L\,|\,U}(\ell_m\,|\,u_m)$.

S3) $p(\boldsymbol{x}\,|\,\boldsymbol{\ell}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{P}^{-1})$ with $\mathbf{P} = \mathbf{V}\mathbf{L}\mathbf{V}^{\mathrm{t}}, \boldsymbol{\mu} = \mathbf{P}^{-1}\mathbf{V}\mathbf{L}\boldsymbol{w}$, and $\mathbf{L} = \mathrm{Diag}\{\boldsymbol{\ell}\}$.

Proof: Since $U = G/\sqrt{L}$ is a SMG, we have

$$\mathrm{d}F_{U,L}(u, l) = p_G(u\sqrt{\ell})\sqrt{\ell}\,\mathrm{d}F_L(\ell)\mathrm{d}u \tag{17}$$
$$\Longrightarrow p_U(u) = \int_{\mathbb{R}_+} p_G(u\sqrt{\ell})\sqrt{\ell}\,\mathrm{d}F_L(\ell). \tag{18}$$

From (15), (16), and (18) it is clear that S1) holds. The two conditional properties follow from (16) straightforwardly. ■

Lemma 2 (LMG): Let $U = \sigma G + L$ be a LMG. Then

$$\mathrm{d}F(\boldsymbol{x}, \boldsymbol{\ell}) \propto \prod_m p_G\left(\frac{u_m - \ell_m}{\sigma}\right)\mathrm{d}F_L(\ell_m)\,\mathrm{d}\boldsymbol{x} \tag{19}$$

fulfills the following properties:

L1) $p_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\boldsymbol{\ell}\in\mathbb{R}^M} \mathrm{d}F(\boldsymbol{x}, \boldsymbol{\ell})$.

L2) $F(\boldsymbol{\ell}\,|\,\boldsymbol{x}) = \prod_m F_{L\,|\,U}(\ell_m\,|\,u_m)$.

L3) $p(\boldsymbol{x}\,|\,\boldsymbol{\ell}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{P}^{-1})$ with $\boldsymbol{\mu} = (\mathbf{V}\mathbf{V}^{\mathrm{t}})^{-1}\mathbf{V}(\boldsymbol{w} + \boldsymbol{\ell})$ and $\mathbf{P} = \sigma^{-2}\,\mathbf{V}\mathbf{V}^{\mathrm{t}}$.

Proof: Since U is a LMG, we have

$$p_U(u) = \frac{1}{\sigma}\int_{\mathbb{R}} p_G\left(\frac{u - \ell}{\sigma}\right)\mathrm{d}F_L(\ell). \tag{20}$$

From (15), (19), and (20), it is easy to deduce L1). The two conditional properties follow from (19) straightforwardly. ■

[1] Throughout the paper, $\mathcal{N}(\boldsymbol{\mu}, \mathbf{P})$ stands for a Gaussian probability with mean $\boldsymbol{\mu}$ and covariance $\mathbf{P}$.

[2] The DF F_V of a RV V is defined by $F_V(v) = \Pr(V \le v)$ [7]. The case with a density corresponds to $\mathrm{d}F_V(v) = p_V(v)\mathrm{d}v$, which will be also written $\mathrm{d}F(v) = p(v)\mathrm{d}v$ whenever unambiguous.

In (16) and (19), $F(\boldsymbol{x},\boldsymbol{\ell})$ is the DF of a "Compound Gauss-Markov random fields with noninteractive auxiliary variables" [12] when $p_{\boldsymbol{X}}$ is the pdf of a Markov random field.

Conditions on ϕ under which the SMG or the LMG constructions are possible are discussed in Subsections IV-C1 and IV-C2, respectively.

B. EM Iteration

Let us explicit an EM scheme for maximization of $p(\boldsymbol{x}\,|\,\boldsymbol{z})$ with $(\boldsymbol{z},\boldsymbol{\ell})$ as complete dataset. Let

$$Q(\boldsymbol{x},\boldsymbol{x}_0)=\int \log p(\boldsymbol{z},\boldsymbol{x}\,|\,\boldsymbol{\ell})\,\mathrm{d}F(\boldsymbol{\ell}\,|\,\boldsymbol{z},\boldsymbol{x}_0).$$

The iteration $\boldsymbol{x}^{(n)}=\arg\max_{\boldsymbol{x}} Q(\boldsymbol{x},\boldsymbol{x}^{(n-1)})$ defines a proper EM algorithm since Q fulfills the basic identity

$$\log\frac{p(\boldsymbol{x}\,|\,\boldsymbol{z})}{p(\boldsymbol{x}_0\,|\,\boldsymbol{z})}=Q(\boldsymbol{x},\boldsymbol{x}_0)-Q(\boldsymbol{x}_0,\boldsymbol{x}_0)+\mathrm{KL}(\boldsymbol{x}_0,\boldsymbol{x})$$

where $\mathrm{KL}(\boldsymbol{x}_0,\boldsymbol{x})$ is the Kullback-Leibler pseudo-distance between $F(\boldsymbol{\ell}\,|\,\boldsymbol{x}_0,z)$ and $F(\boldsymbol{\ell}\,|\,\boldsymbol{x},\boldsymbol{z})$.

Given $p(\boldsymbol{z},\boldsymbol{x}\,|\,\boldsymbol{\ell})=p(z\,|\,\boldsymbol{x})\,p(\boldsymbol{x}\,|\,\boldsymbol{\ell})$ and $\mathrm{d}F(\boldsymbol{\ell}\,|\,\boldsymbol{x}_0,z)=\mathrm{d}F(\boldsymbol{\ell}\,|\,\boldsymbol{x}_0)$, Q also reads

$$\begin{aligned}Q&=\log p(z\,|\,\boldsymbol{x})+\int \log p(\boldsymbol{x}\,|\,\boldsymbol{\ell})\mathrm{d}F(\boldsymbol{\ell}\,|\,\boldsymbol{x}_0)\\&=-J_0(\boldsymbol{x})-\frac{1}{2}\int(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{t}}\mathbf{P}(\boldsymbol{x}-\boldsymbol{\mu})\,\mathrm{d}F(\boldsymbol{\ell}\,|\,\boldsymbol{x}_0)+C\end{aligned}$$

where C is independent of $\boldsymbol{x}$. Note that the former derivation depends strongly on $\mathrm{d}F(\boldsymbol{\ell}\,|\,\boldsymbol{x}_0,z)=\mathrm{d}F(\boldsymbol{\ell}\,|\,\boldsymbol{x}_0)$ which means that conditionally on $\boldsymbol{x}$ the data $\boldsymbol{z}$ is independent of the "augmented data" $\boldsymbol{\ell}$: this is true here by construction of $\boldsymbol{\ell}$ but it is unusual in an EM context. The remaining integrand depends on the case considered. In the following, we denote $u_m^0=\boldsymbol{v}_m^{\mathrm{t}}\boldsymbol{x}_0-w_m$.

Theorem 1 (SMG): Under the hypothesis of Lemma 1, we have $Q(\boldsymbol{x},\boldsymbol{x}_0)=-K_{\mathrm{GR}}(\boldsymbol{x},\boldsymbol{\ell}^0)+C_{\mathrm{GR}}$ where $\boldsymbol{\ell}^0=\mathbf{col}\{\phi'(u_m^0)/u_m^0\}$ and C_{GR} does not depend on $\boldsymbol{x}$.

Proof: See Appendix A. ∎

Theorem 2 (LMG): Under the hypothesis of Lemma 2, we have $Q(\boldsymbol{x},\boldsymbol{x}_0)=-K_{\mathrm{GY}}^{\sigma^2}(\boldsymbol{x},\boldsymbol{\ell}^0)+C_{\mathrm{GY}}$ where $\boldsymbol{\ell}^0=\mathbf{col}\{u_m^0-\sigma^2\phi'(u_m^0)\}$ and C_{GY} does not depend on $\boldsymbol{x}$.

Proof: See Appendix B. ∎

It is clear now from Theorems 1 and 2 that EM iterations identify with block coordinate descent on K_{GR} or on K_{GY}, i.e., (10), (12) or (11), (13) in the unconstrained case.

C. Sufficient Conditions

The specification of $F(\boldsymbol{x},\boldsymbol{\ell})$ and the resulting equivalence results require that U be a SMG or a LMG, which corresponds to conditions on ϕ that we are going to examine more carefully in the present subsection.

1) Scale Mixtures of Gaussians:

Theorem 3 (Andrews and Mallows [8]): A scalar RV U with a symmetric pdf p_U is a SMG *iff*

$$(-\mathrm{d}/\mathrm{d}u)^n p_U(\sqrt{u})\geq 0,\quad \forall u>0, n\in\mathbb{N}. \tag{21}$$

Condition (21) is satisfied by many standard distributions, such as Laplace [8], Student t-distribution [8] and hyperbolic distributions [11]. In all such cases, L admits a pdf with an explicit expression.

On the other hand, the following lemma states that (21) implies (7), which is the main condition for the existence of the augmented criterion K_{GR} according to Proposition 1.

Lemma 3: Let U be a SMG, and let $\phi=-\log p_U$. Then $\phi(\sqrt{\cdot})$ is concave on $\mathbb{R}_+$.

Proof: Let $f=p_U(\sqrt{\cdot})$, by (21), f is such that $f>0, f'\leq 0$, and $f''\geq 0$ on $\mathbb{R}_+\setminus\{0\}$. Let $0<a<b$, so that $f(a)\geq f(b)$ and $f'(a)\leq f'(b)$. Thus $f'(a)/f(a)\leq f'(b)/f(b)$, i.e., $(\log f)'$ is increasing, which means that $\phi(\sqrt{\cdot})=-\log f$ is concave on $\mathbb{R}_+$. ∎

Then, for any pdf p_U that corresponds to a SMG, existence of $\mathrm{d}F(\boldsymbol{x},\boldsymbol{\ell})$ and of K_{GR} is guaranteed, and both constructions lead to the same algorithm.

2) Location Mixtures of Gaussians: In a similar way, if U is a LMG then the augmented criterion K_{GY} can be defined according to Proposition 2:

Lemma 4: If U is a LMG and $\phi=-\log p_U$, then $(\,\cdot\,)^2/2-\sigma^2\phi$ is a convex function.

Proof: p_U and ϕ have derivatives of any order according to (20), so we only need to check $(u^2/2-\sigma^2\phi(u))''=1-\sigma^2\phi''(u)\geq 0$. From $p_U^{(n)}=\sigma^{-n-1}\mathrm{E}[p_G^{(n)}((u-L)/\sigma)]$, $p_G'(u)=-u\,p_G(u)$ and $p_G''(u)=(u^2-1)\,p_G(u)$, we get

$$1-\sigma^2\phi''=1-\sigma^2\frac{p_U\,p_U''-(p_U')^2}{p_U^2}=\frac{p_0\,p_2-p_1^2}{\sigma^2 p_0^2},$$

where $p_n(u)=[(u-L)^n\,p_G((u-L/\sigma))$. By Cauchy-Schwarz inequality, $p_1^2\leq p_0\,p_2$, which allows to conclude. ∎

LMG distributions can be characterized in terms of characteristic function (CF). In the following, Φ_V denotes the CF of a RV V.

Theorem 4: U is a LMG *iff* $\Phi_U(\omega)\exp(\sigma^2\omega^2/2)$ is a positive definite function[3] for some $\sigma>0$

Proof: Assume $\Phi(\omega)=\Phi_U(\omega)\exp(\sigma^2\omega^2/2)$ is positive definite. By construction, it is continuous. Then by Bochner Theorem [7, Sec. XIX.2] there is a RV L such that $\Phi_L=\Phi$. Let $G\sim\mathcal{N}(0,1)$ be independent of L, so that $\Phi_{\sigma G+L}=\Phi_U$, thus $p_{\sigma G+L}=p_U$, therefore U is a LMG. Conversely, assume U is a LMG, then $U=\sigma G+L$, thus Φ is the CF of L. ∎

The LMG structure is actually very restrictive: for instance, Φ_U must converge to zero at least as fast as $\exp(-\sigma^2\omega^2/2)$ (since $\Phi_U(\omega)\exp(\sigma^2\omega^2/2)$ is bounded, as the CF of L). Indeed, LMG are scarcely found among standard distributions. An example of "usual" LMG is provided by finite Gaussian mixtures.

On the other hand one can define U indirectly through specification of L. For instance take L to have a Laplace distribution. Then p_U is log-concave as the convolution of log-concave distributions ([13], Lemma 2), therefore the resulting $\phi=-\log p_U$ is a new C^∞ variant of the Huber function. Such a "Laplace$\star$Gauss" probability seems easier to simulate than other Huber-like distributions, which is an interesting feature in the perspective of applications involving stochastic sampling.

[3]For a mathematical definition of definite positivity; see [7, Sec. XIX.2].

V. Discussion

D. Geman *et al.*'s pioneering contributions [2], [3] have introduced duality tools for the construction of *augmented, half-quadratic* criteria in the field of signal and image restoration: such criteria $K_{\rm GR}$ and $K_{\rm GY}$ share the same minimizer as the original nonquadratic criterion J in the sense of (4). More recently, it has been underlined that *reweighted least square* algorithms such as IRLS and RSD are block coordinate descent schemes that minimize $K_{\rm GR}$ and $K_{\rm GY}$, respectively [1].

In this paper, a new interpretation has been given to IRLS and RSD. In contrast to previous contributions, we have given a statistical meaning to both algorithms. More precisely, we have shown that they are particular cases of EM algorithms. The corresponding *complete datasets* involve two types of Gaussian mixtures, respectively *scale* and *location* mixtures of Gaussian.

It might seem trivial that the minimization of half-quadratic criteria amounts to maximize the likelihood of Gaussian mixtures. The result is actually correct, but its derivation is not as straightforward. In particular, the pdf of the underlying scale (resp., location) mixture of Gaussian is not proportional to $\exp(-K_{\rm GR})$ (resp., $\exp(-K_{\rm GY})$).

Surprisingly, it is only under restrictive conditions that IRLS and RSD have received a statistical interpretation in terms of EM. In particular, the location mixture construction seems far more restrictive than the effective range of application of RSD.

Appendix

A. Proof of Theorem 1

According to (S3), $(\boldsymbol{x}-\boldsymbol{\mu})^{\rm t}\mathbf{P}(\boldsymbol{x}-\boldsymbol{\mu}) = (\mathbf{V}^{\rm t}\boldsymbol{x}-\boldsymbol{w})^{\rm t}\mathbf{L}(\mathbf{V}^{\rm t}\boldsymbol{x}-\boldsymbol{w}) + C'$ where C' is independent of $\boldsymbol{x}$, so that $Q=-J_0-(1/2)(\mathbf{V}^{\rm t}\boldsymbol{x}-\boldsymbol{w})^{\rm t}\mathbf{L}^0(\mathbf{V}^{\rm t}\boldsymbol{x}-\boldsymbol{w})+C''$, where $\mathbf{L}^0 = \operatorname{Diag}\{\mathrm{E}[L_m \,|\, \boldsymbol{X}=\boldsymbol{x}_0]\}$. Two points still remain to be proved:

- $\mathrm{E}[L_m \,|\, \boldsymbol{X}=\boldsymbol{x}] = \mathrm{E}[L \,|\, U=u_m]$ results from (S2).
- $\mathrm{E}[L \,|\, U] = \phi'(U)/U$ stems from the following property, which appears in [14] without proof.

Proposition 3: If U is a SMG, then $E[L \,|\, U] = \phi'(U)/U$, where $\phi = -\log p_U$.

Proof: $\phi' = -p_U'/p_U$, differentiation of (18) under the integral yields

$$p_U'(u) = -u\int p_G(u\sqrt{\ell})\ell^{3/2}\mathrm{d}F_L(\ell). \tag{22}$$

Given (17) and $\mathrm{d}F_{U,L}(u,\ell) = \mathrm{d}F_{L\,|\,U}(\ell \,|\, u)p_U(u)\,\mathrm{d}u$,

$$\mathrm{E}[L \,|\, U=u] = \frac{1}{p_U(u)}\int p_G(u\sqrt{\ell})\,\ell^{3/2}\mathrm{d}F_L(\ell).$$

Therefore (22) also reads $p_U'(u) = -up_U(u)E[L \,|\, U=u]$. ∎

B. Proof of Theorem 2

According to (L3), we have $(\boldsymbol{x}-\boldsymbol{\mu})^{\rm t}\mathbf{P}(\boldsymbol{x}-\boldsymbol{\mu}) = \sigma^{-2}||\mathbf{V}^{\rm t}\boldsymbol{x}-\boldsymbol{w}-\boldsymbol{\ell}||^2$, thus

$$\begin{aligned} Q &= -J_0 - \frac{1}{2\sigma^2}\mathrm{E}[||\mathbf{V}^{\rm t}\boldsymbol{x}||^2 - 2\boldsymbol{x}^{\rm t}\mathbf{V}(\boldsymbol{L}+\boldsymbol{w}) \,|\, \boldsymbol{x}_0] + C' \\ &= -J_0 - \frac{1}{2\sigma^2}||\mathbf{V}^{\rm t}\boldsymbol{x}-\boldsymbol{w}-\boldsymbol{\ell}^0||^2 + C'' \end{aligned}$$

where $\boldsymbol{\ell}^0 = \mathrm{E}[\boldsymbol{L} \,|\, \boldsymbol{X}=\boldsymbol{x}_0]$, and C' and C'' are independent of $\boldsymbol{x}$. Two points still remain to be proved:

- $\mathrm{E}[L_m \,|\, \boldsymbol{X}=\boldsymbol{x}] = \mathrm{E}[L \,|\, U=u_m]$ results from (L2).
- $\mathrm{E}[L \,|\, U] = U-\sigma^2\phi'(U)$ stems from the following result.

Proposition 4: If U is a LMG, then $\mathrm{E}[L \,|\, U] = U-\sigma^2\phi'(U)$, where $\phi = -\log p_U$.

Proof: $\phi' = -p_U'/p_U$, differentiation of (20) under the integral yields

$$\begin{aligned} \sigma^3 p_U' &= \mathrm{E}\left[(u-L)p_G\left(\frac{u-L}{\sigma}\right)\right] \\ &= \sigma u p_U(u) - \mathrm{E}\left[Lp_G\left(\frac{u-L}{\sigma}\right)\right]. \end{aligned} \tag{23}$$

Since $\mathrm{d}F_{L\,|\,U}(\ell \,|\, u)p_U(u) = (1/\sigma)p_G((u-\ell)/\sigma)\mathrm{d}F_L(\ell)$, we have

$$\sigma p_U(u)E[L \,|\, U=u] = \mathrm{E}\left[Lp_G\left(\frac{u-L}{\sigma}\right)\right].$$

Combination with (23) completes the proof. ∎

Acknowledgment

The authors wish to thank the reviewers for their careful, diligent review and their useful comments.

References

[1] J. Idier, "Convex half-quadratic criteria and interacting auxiliary variables for image restoration," *IEEE Trans. Image Processing*, vol. 10, pp. 1001–1009, July 2001.

[2] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 367–383, Mar. 1992.

[3] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Processing*, vol. 4, pp. 932–946, July 1995.

[4] T. F. Chan and P. Mulet, "On the convergence of the lagged diffusivity fixed point method in total variation image restoration," in *SIAM J. Numer. Anal.*, vol. 36, 1999, pp. 354–367.

[5] M. Allain, "Approche pénalisée en tomographie hélicoïdale. Application à la conception d'une prothése personnalisée du genou," Co-Supervised Ph.D. Thesis, Univ. Paris-Sud, Orsay, France, Dec. 2002. École Polytech. Montréal, QC, Canada.

[6] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *J. Comput. Graph. Statist.*, vol. 9, no. 1, pp. 1–20, Mar. 2000. with discussion.

[7] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed. New York: Wiley, 1971, vol. 2.

[8] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. R. Statist. Soc. B*, vol. 36, pp. 99–102, 1974.

[9] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Processing*, vol. 12, pp. 1338–1351, Nov. 2003.

[10] J. Geweke, "Bayesian treatment of the independent student-t linear model," *J. Appl. Econom.*, vol. 8, pp. S19–S40, 1993.

[11] *Handbook of Heavy Tailed Distributions in Finance*, S. Rachev, Ed., Elsevier, New York, 2003, pp. 211–248. B. M. Bibby and M. Sørensen, Hyperbolic processes in finance.

[12] F. C. Jeng and J. W. Woods, "Simulated annealing in compound Gaussian random fields," *IEEE Trans. Inform. Theory*, vol. 36, pp. 94–107, Jan. 1990.

[13] I. A. Ibragimov, "On the composition of unimodal distributions," *Theory Probl. Appl.*, vol. 1, pp. 255–260, 1956.

[14] M. West, "Outlier models and prior distributions in Bayesian linear regression," *J. R. Statist. Soc. B*, vol. 46, no. 3, pp. 431–439, 1984.