Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces

Axel Munk[1], Nicolai Bissantz[1,2], Lutz Dümbgen[3], and Bernd Stratmann[1]

[1] Institute for Mathematical Stochastics
Georg-August-University Göttingen, Germany
[2] Faculty of Mathematics
Ruhr-University Bochum, Germany
and
[3] Institute of Mathematical Statistics and Actuarial Science
University of Bern, Switzerland

November 13, 2006

Abstract

The computation of robust regression estimates often relies on minimization of a convex functional on a convex set. In this paper we discuss a global technique for a large class of convex functionals to compute the minimizers iteratively which is closely related to majorization-minimization algorithms. We show convergence on general convex function spaces for general coercive and convex functionals F. This includes the iteratively reweighted least squares algorithm as a special case. The algorithm is applied to an example from astrophysics which amounts to bivariate regression with unimodality constraints.

AMS 2000 subject classification: primary 62G07; secondary 62J05, 65K05, 85-08

Keywords: regression analysis, monotone regression, quantile regression, shape constraints, L^1 regression, nonparametric regression, total variation semi-norm, reweighted least squares, Fermat's problem, convex approximation, pool adjacent violators algorithm

Address for correspondence: Nicolai Bissantz, Fakultät für Mathematik, Universitätsstraße 150, Mathematik III, NA 3/70, D-44780 Bochum, Germany; email: nicolai.bissantz@rub.de

1 Introduction

The computation of robust parametric and nonparametric regression estimators often requires the minimization of (convex) functionals on a set $\mathcal{C}$ which is determined by a priori information on the model underlying the data. For example, $\mathcal{C}$ can be a linear finite-dimensional space (linear model) or the set of isotonic vectors $m = (m_1, ..., m_d) \in \mathbb{R}^d$, $m_1 \leq ... \leq m_d$, with $d \leq n$. To this end

the functional

$$F^{(\rho)}(m) = \sum_{i=1}^{n} \rho(r_i(m)) \tag{1}$$

has to be minimized over $\mathcal{C} \subset \mathbb{R}^d$. Here $r_i, i = 1, ..., n$ denote the (model-dependent) residuals of n data pairs (X_i, Y_i), $i = 1, \ldots, n$ and ρ a given function (Huber 1981). Taking $\rho(z) = z^2/2$ gives the ordinary least squares problem, and

$$\rho(z) = 2|z| \cdot \begin{cases} p & z \geq 0 \\ 1-p & z < 0 \end{cases} \tag{2}$$

with $0 < p < 1$ yields quantile regression (Koenker & Bassett 1978, Portnoy 1997). Other functions are logistic $\rho(z) = \gamma^z \log(\cosh(z/\gamma))$ (Coleman et al. 1980) or Huber's (1964) loss function

$$\rho(z) = \begin{cases} z^2/2 & |z| \leq \gamma \\ \gamma|z| - \gamma^2/2 & |z| > \gamma \end{cases}$$

for some $\gamma > 0$. An important extension of (1) are functionals

$$F(m) = F^{(\rho)}(m) + \lambda P(m)\ , \quad \lambda \geq 0, \tag{3}$$

where $P(m)$ denotes a penalizing term such as, for instance, the discrete total variation semi-norm of $m \in \mathbb{R}^d$,

$$P(m) = \sum_{j=1}^{d-1} |m_j - m_{j+1}|\ ; \tag{4}$$

see Künsch (1994), Koenker, Ng & Portnoy (1994) or Mammen & van de Geer (1997). In this paper a generalization of the iteratively reweighted least squares (IRLS) algorithm - therefore named GIRLS - is considered for minimization of a functional F as in (3) over any convex subset $\mathcal{C}$ of $\mathbb{R}^d$. This allows us to extend the IRLS algorithm for example to situations where $\mathcal{C}$ is defined as the space of monotone (or k-modal) vectors or to the problem of nonparametric regression estimates with total variation semi-norm penalization of its discrete derivative.

The general idea of the IRLS algorithm (and variants of it) is to approximate the functional F in a first step by smooth functionals F_δ such that $F_\delta \to F$ pointwise as $\delta \searrow 0$. The collection $(F_\delta)_{\delta>0}$ will be called a regularization of F (cf. Def. 1). In a second step, for each given base point $f \in \mathcal{C}$ the functional F_δ will be approximated by $G_\delta(f, .)$ (cf. Def. 2). Here $G_\delta : \mathcal{C} \times \mathbb{R}^d \to \mathbb{R}$ is a functional which is chosen such that a quick and numerically stable iterative minimization can be performed. The resulting minimizer will serve as an approximation for the minimizer m_δ^* of F_δ and hence for a minimizer m^* of F. In particular, if it is possible to choose G_δ as a polynomial of degree two, the well known iteratively reweighted least squares algorithm may result (Lejeune & Sarda 1988, McCullagh & Nelder 1989, Dodge & Jurečková 2000).

The IRLS and related algorithms are based on the idea of majorizing functionals by a sequence of quadratic approximations and subsequent minimization. These have been treated extensively in the literature, e.g. Kuhn (1972), Katz (1973), Wolke & Schwetlick (1988), O'Leary (1990), de Leeuw & Michailidis (2000), Hunter & Lange (2000), Lange, Hunter & Yang (2000), Vardi & Zhang (2000, 2001), and the references therein. However, in most cases convergence is only shown for $\mathcal{C} = \mathbb{R}^d$. This simplifies proofs notably, since the minimizers can be represented as zeros of the derivatives of the functional. For arbitrary convex $\mathcal{C}$, however, the minimizers are no longer represented solely by such equality constraints, instead inequalities occur. Notable exceptions for general convex $\mathcal{C}$ are Eckhardt (1980), where however, the convergence results are restricted to a special class of functionals, requiring e.g. $F(m) = O(\|m\|)$, Voß & Eckhardt (1980), who show convergence on convex polyhedral sets under certain regularity conditions on the functional. Our findings generalize these results to the case of $\mathcal{C}$ being an arbitrary convex closed set as well as to more general functionals and functionals which are only required to be coercive and convex. This appears to be close to the weakest possible set of assumptions required for a general proof of convergence. Our proof adopts various arguments from convex analysis.

It is interesting to note that in the numerical literature the IRLS algorithm is denoted as the Weiszfeld algorithm (Weiszfeld, 1936,1937) who suggested this algorithm to solve the Fermat-Steiner-Weber problem (Weiszfeld 1936, 1937, Kuhn 1973, Katz 1974) which is known to the statistical community as the computation of the spatial median (as mentioned in Brown 1983, Brown et al. 1997, Ducharme & Milasevic 1987).

The paper is organized as follows. First, we motivate the GIRLS algorithm for the special case of L^1-regression in Section 2. Then we introduce the GIRLS-algorithm in a general framework and prove various results about its convergence in Section 3. The algorithm is defined by

$$m_{k+1} := \operatorname*{argmin}_{m \in \mathcal{C}} G_\delta(m_k, m) \,, \qquad k \in \mathbb{N}. \tag{5}$$

Its convergence to the minimizer m_δ^* and hence to m^* as $\delta \searrow 0$ will be shown under very general assumptions. Furthermore, we give a result showing geometric, or, more precisely, at least Q-linear convergence of the sequence $(m_k)_k$ to m_δ^* under slightly different conditions (cf. Voß & Eckhardt (1980), and Böhning & Lindsay, 1988), and guidance is provided on the choice of the number of iterates in (5) and the regularization parameter δ.

In Section 4 we describe the construction of F_δ and G_δ in some specific cases explicitly. Finally, we present in Section 5 an application of the GIRLS algorithm to the estimation of the variance surface

of brightness data from astrophysical measurements. Here the data is on a two-dimensional grid and we impose a unimodal constraint in one direction and penalization of a regression function's non-smoothness. We'd like to mention that for univariate unimodal regression problems, the pool adjacent violators algorithm (PAVA) is often more efficient whenever applicable (see Robertson et al. 1988 for a comprehensive treatment). However, for regression with two- or higher dimensional predictor this is not valid anymore, whereas the GIRLS algorithm can be transferred to predictors of dimension ≥ 2. Furthermore, PAVA type algorithms are not available in general if an additional penalization term as in (3) is added.

In summary, the main advantage of the GIRLS algorithm is twofold. First, it is simple to perform and offers great flexibility for the choice of the approximating functionals G_δ. Second, it allows us to combine various restrictions and minimisation criteria (such as monotonicity constraints and roughness penalties). For such complex minimisation problems simple and quick algorithms such as PAVA or Newton type algorithms are not available in general, and more complicated and time consuming algorithms such as quadratic programming or interior point methods become necessary. Here the GIRLS-algorithm represents a feasible alternative because it typically requires in each updating step the computation of minimizers (e.g. a weighted L^2 solution), which can be obtained easily. Further, our numerical experiments have shown that a rather small number of updating steps give already satisfactory results.

2 L^1-regression with the GIRLS algorithm

As a motivating example consider the L^1 linear regression problem for observations (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n) in $\mathbb{R}^d \times \mathbb{R}$. Assuming that Y_i equals $X_i^\top m$ plus a random error, the goal is to compute

$$\hat{m} := \operatorname*{argmin}_{m \in \mathbb{R}^d} \sum_{i=1}^{n} |Y_i - X_i^\top m| = \operatorname*{argmin}_{m \in \mathbb{R}^d} F(m) \,, \tag{6}$$

an estimator of the unknown parameter vector $m \in \mathbb{R}^d$. Iteratively reweighted least squares is based on the idea that, in a first step, the L^1 norm F, being a convex functional, will be approximated (regularized) by a family of smooth convex functionals F_δ, $\delta > 0$, e.g.

$$F_\delta(m) = \sum_{i=1}^{n} h_\delta \left(Y_i - X_i^\top m \right) \,,$$

where

$$h_\delta(z) = [z^2 + \delta]^{1/2} \,. \tag{7}$$

It is supposed that minimisation of F_δ is numerically better tractable than minimisation of the original functional F in in (6). Then $\hat{m}_\delta := \operatorname{argmin}_{m\in\mathbb{R}^d} F_\delta(m)$ will be an approximation of $\hat{m}$ (cf. Theorem 1). In order to compute $\hat{m}_\delta$ the following recursion formula is iterated:

$$\hat{m}_\delta^{(k+1)} = \operatorname*{argmin}_{m\in\mathbb{R}^d} \sum_{i=1}^n \frac{(Y_i - X_i^\top m)^2}{h_\delta(Y_i - X_i^\top \hat{m}_\delta^{(k)})} \,. \tag{8}$$

Note, that in each updating step the computation of $\hat{m}_\delta^{(k+1)}$ means solving a simple diagonally reweigthed least squares minimisation problem, which can easily be done by using standard methods such as, e.g., Householder QR decomposition. As a starting value $\hat{m}_\delta^{(0)}$ any (reasonable) choice, e.g. the least squares estimator, may serve.

It is instructive to indicate a proof for this simple case. The basic idea is to approximate $h_\delta(z)$ from above for any given real number r by a quadratic function $g_\delta(r,z) = c(r) + a(r)z^2/2$ of z such that $g_\delta(r,\cdot) \geq h_\delta$ and $g_\delta(r,r) = h_\delta(r)$. This can be achieved indeed with

$$g_\delta(r,z) \;=\; h_\delta(r) + h_\delta(r)^{-1}(z^2 - r^2)/2 \,; \tag{9}$$

see also Lemma 1 in Section 4. The intrinsic reason is that h_δ is an even convex function whose second derivative h_δ'' is non-increasing on $[0,\infty)$. Thus $\hat{m}_\delta^{(k+1)}$ in (8) is the minimizer of

$$G_\delta(\hat{m}_\delta^{(k)}, m) \;:=\; \sum_{i=1}^n g_\delta\left(Y_i - X_i^\top \hat{m}_\delta^{(k)}, Y_i - X_i^\top m\right)$$

over all $m \in \mathbb{R}^d$. Note that F_δ as well as $G_\delta(\hat{m}_\delta^{(k)}, \cdot)$ are convex functions such that $F_\delta(m) \leq G_\delta(\hat{m}_\delta^{(k)}, m)$ with equality for $m = \hat{m}_\delta^{(k)}$, and their gradients satisfy $\nabla F_\delta(\hat{m}_\delta^{(k)}) = \nabla G_\delta(\hat{m}_\delta^{(k)}, \hat{m}_\delta^{(k)})$. Here and in the following the gradient of G_δ is defined with respect to the second argument. Thus

$$F_\delta(\hat{m}_\delta^{(k+1)}) \leq G_\delta(\hat{m}_\delta^{(k)}, \hat{m}_\delta^{(k+1)}) \leq G_\delta(\hat{m}_\delta^{(k)}, \hat{m}_\delta^{(k)}) = F_\delta(\hat{m}_\delta^{(k)}) \,,$$

and the last inequality is strict if, and only if, $\hat{m}_\delta^{(k)}$ differs from the solution $\hat{m}_\delta$. Consequently, $F_\delta(\hat{m}_\delta^{(k)})$ is either strictly decreasing in k, or $\hat{m}_\delta^{(k)} = \hat{m}_\delta$ for sufficiently large k. This fact was established by Lejeune & Sarda (1988) for the particular problem (6). Convergence of $\hat{m}_\delta^{(k)}$ to $\hat{m}_\delta$ as $k \to \infty$ follows from our general Theorem 2 below.

3 The GIRLS algorithm

3.1 Main theorem and convergence analysis

Now let us turn to the general setting. We always assume that our target functional $F : \mathbb{R}^d \to \mathbb{R}$ is convex and coercive, i.e. $F(x) \to \infty$ as $\|x\| \to \infty$. Moreover, let $\mathcal{C} \subset \mathbb{R}^d$ be closed and convex.

This entails that the set

$$M^* := \operatorname*{argmin}_{m \in \mathcal{C}} F(m)$$

is a nonvoid, compact and convex subset of $\mathcal{C}$. Now the first step is to approximate F by a family of strictly convex and smooth functionals F_δ, $\delta > 0$, converging pointwise to F as $\delta \searrow 0$. This is summarized in the following

Definition 1. A functional $F_\delta : \mathbb{R}^d \to \mathbb{R}$ is called *regular*, if F_δ is strictly convex, continuously differentiable and coercive. A *regular class for* F (or a *regularization of* F) consists of regular functionals F_δ, $\delta > 0$, such that F_δ converges pointwise to F as $\delta \searrow 0$.

Theorem 4 below shows that there exists always a regular class $(F_\delta)_{\delta>0}$ for F. It follows from strict convexity and coercivity of F_δ that it has a unique minimizer

$$m_\delta^* := \operatorname*{argmin}_{m \in \mathcal{C}} F_\delta(m)$$

which serves as an approximation to M^*. The next theorem provides an exact formulation of this fact.

Theorem 1. (Approximation of M^*).
Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex and coercive functional, and let $(F_\delta)_{\delta>0}$ be a regularization of F. Then, as $\delta \searrow 0$,

$$\left.\begin{matrix} F_\delta(m_\delta^*) \\ F(m_\delta^*) \end{matrix}\right\} \to \min_{x \in \mathcal{C}} F(x) \quad \textit{and} \quad d(m_\delta^*, M^*) := \inf_{y \in M^*} \|m_\delta^* - y\| \to 0 \,.$$

Before proving Theorem 1 we summarize some well known facts about convex functionals (see Rockafellar 1970) which we utilize in the subsequent proofs. A convex functional on $\mathbb{R}^d$ is automatically continuous. If a sequence of convex functionals on $\mathbb{R}^d$ converges pointwise, then the convergence is uniform on arbitrary bounded sets. Finally, if $H : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, and if $\mathcal{C} \subset \mathbb{R}^d$ is closed and convex, then $f \in \mathcal{C}$ minimizes H over $\mathcal{C}$ if, and only if,

$$\nabla H(f)^\top (m - f) \geq 0 \quad \text{for all } m \in \mathcal{C} \,. \tag{10}$$

Proof. For any set $S \subset \mathbb{R}^d$ let $\|F - F_\delta\|_S$ be the supremum norm of $F - F_\delta$ over S. For any fixed $\epsilon > 0$, the set $B_\epsilon := \{m \in \mathcal{C} : d(m, M^*) \leq \epsilon\}$ is compact. Thus $\|F - F_\delta\|_{B_\epsilon}$ tends to zero as $\delta \searrow 0$. In particular, for sufficiently small $\delta > 0$,

$$\min_{m \in \mathcal{C} \,:\, d(m, M^*) = \epsilon} F_\delta(m) > \max_{m \in M^*} F_\delta(m) \,.$$

In order to see the last inequality, note that this holds for F and use that $F_\delta \to F$ uniformly on bounded sets. But this implies that $F_\delta(m_o) > \min_{m\in B_\epsilon} F_\delta(m)$ for any $m_o \in \mathcal{C}\setminus B_\epsilon$, i.e. $m_\delta^* \in B_\epsilon$. For let m_* be the metric projection of m_o onto M^* and write $m_o = m_* + tv$ for some unit vector $v \in \mathbb{R}^d$ and a scalar $t > \epsilon$. Then it follows from convexity of F_δ that

$$F_\delta(m_o) - F_\delta(m_*) \geq (\epsilon/t)(F_\delta(m_* + \epsilon v) - F_\delta(m_*)) > 0\ .$$

Note also that in case of $m_\delta^* \in B_\epsilon$,

$$|F_\delta(m_\delta^*) - F(m_\delta^*)| \leq \|F - F_\delta\|_{B_\epsilon}\ ,$$

and

$$F(m_\delta^*) - \min_{\mathcal{C}} F \leq \max_{m\in B_\epsilon}\Big(F(m) - \min_{M^*} F\Big).$$

Finally, the r.h.s. of this inequality can be made arbitrarily small for proper choice of δ and ϵ by compactness of M^* and continuity of F. □

The second step is to determine m_δ^* via approximations $G_\delta(f,\cdot)$ of F_δ for various $f \in \mathcal{C}$ as in (5). The following definition summarizes our assumptions on G_δ.

Definition 2. Let $F_\delta : \mathbb{R}^d \to \mathbb{R}$ be a regular functional. Another functional $G_\delta : \mathcal{C}\times\mathbb{R}^d \to \mathbb{R}$ is called a *smooth approximation of F_δ from above*, if it is continuous in both arguments and satisfies the following additional properties for arbitrary $f \in \mathcal{C}$:
(i) $G_\delta(f,\cdot)$ is strictly convex and continuously differentiable,
(ii) $G_\delta(f,m) \geq F_\delta(m)$ for all $m \in \mathbb{R}^d$ with equality for $m = f$.
The functional G_δ is called a *quadratic approximation of F_δ from above* if, in addition, $G_\delta(f,\cdot)$ is always a polynomial of order two, i.e.

$$G_\delta(f,m) = F_\delta(f) + \nabla F_\delta(f)^\top(m-f) + 2^{-1}(m-f)^\top B(f)(m-f) \tag{11}$$

for some symmetric, positive definite matrix $B(f) \in \mathbb{R}^{d\times d}$.

Note that from (i) and (ii) it follows that $\nabla F_\delta(f) = \nabla G_\delta(f,f)$. The next theorem is the main result of this paper.

Theorem 2. (Convergence of the GIRLS algorithm).
Let $\mathcal{C} \subset \mathbb{R}^d$ be a closed convex set and $F_\delta : \mathbb{R}^d \to \mathbb{R}$ be a regular functional which can be smoothly approximated from above by G_δ. Then the GIRLS algorithm, defined by

$$m_\delta^{(k+1)} := \operatorname*{argmin}_{m\in\mathcal{C}} G_\delta(m_\delta^{(k)}, m) \qquad \textit{for } k = 0,1,2,\ldots \tag{12}$$

with an arbitrary starting point $m_\delta^{(0)} \in \mathcal{C}$, yields a sequence $(m_\delta^{(k)})_{k=0}^\infty$ converging to m_δ^*.

Proof. At first we prove that $F_\delta(m_\delta^{(k)})$ is decreasing in k. It follows from property (ii) in Definition 2 that the gradients $\nabla G_\delta(m_\delta^{(k)}, m)$ (with respect to the second argument) and $\nabla F_\delta(m)$ coincide for $m = m_\delta^{(k)}$. Thus it follows from (10) that $m_\delta^{(k+1)} = m_\delta^{(k)}$ if, and only if, $m_\delta^{(k)} = m_\delta^*$. Otherwise,

$$F_\delta(m_\delta^{(k+1)}) \leq G_\delta(m_\delta^{(k)}, m_\delta^{(k+1)}) < G_\delta(m_\delta^{(k)}, m_\delta^{(k)}) = F_\delta(m_\delta^{(k)}) \; .$$

By monotonicity of $(F_\delta(m_\delta^{(k)}))_k$, all points $m_\delta^{(k)}$ lie in the set $\left\{m \in \mathcal{C} : F_\delta(m) \leq F_\delta(m_\delta^{(0)})\right\}$, which is compact by continuity and coercivity of F_δ. Hence it is sufficient to show that any limit point m_o equals m_δ^*. Now, take an arbitrary convergent subsequence $(m_\delta^{(k_\ell)})_\ell$ with limit m_o. For any $v \in \mathcal{C}$,

$$\begin{aligned} F_\delta\left(m_\delta^{(k_\ell+1)}\right) &\leq G_\delta\left(m_\delta^{(k_\ell)}, m_\delta^{(k_\ell+1)}\right) \\ &\leq G_\delta\left(m_\delta^{(k_\ell)}, v\right) \\ &\to G_\delta(m_o, v) \quad \text{as } \ell \to \infty \; , \end{aligned}$$

by continuity of G_δ. But

$$\lim_{\ell\to\infty} F_\delta\left(m_\delta^{(k_\ell+1)}\right) \geq \lim_{\ell\to\infty} F_\delta\left(m_\delta^{(k_{\ell+1})}\right) = F_\delta(m_o) = G_\delta(m_o, m_o) \; .$$

Thus $G_\delta(m_o, m_o) \leq G_\delta(m_o, v)$ for all $v \in \mathcal{C}$, i.e. m_o is the unique minimizer of $G_\delta(m_o, \cdot)$. As argued above, this entails that $m_o = m_\delta^*$. □

The next theorem states that convex and coercive functionals F can always be regularized and approximated quadratically from above. Hence GIRLS is, in principle, always applicable.

Theorem 3. (Regularization and approximation of F).
Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex and coercive functional. Then there exists a regularization $(F_\delta)_{\delta>0}$ of F such that each F_δ admits a quadratic approximation G_δ from above.

In order to prove Theorem 3 we require the following result.

Theorem 4. *Let F be a nonnegative, coercive, convex functional on $\mathbb{R}^d$. Then there are strictly convex and infinitely often differentiable functionals $F_\delta \geq F$, $\delta > 0$, such that $F_\delta \to F$ pointwise as $\delta \searrow 0$.*

Proof. Let $K(x) := 1\{\|x\| < 1\} C \exp\left(-(1 - \|x\|^2)^{-1}\right)$, where C is chosen such that K integrates to one. This is a well-known example of an infinitely differentiable, nonnegative, even kernel function with compact support $\{x : \|x\| \leq 1\}$. For $\delta > 0$ we define $K_\delta(x) := \delta^{-1} K(\delta^{-1} x)$ and

$$F_\delta(x) := \int F(y) K_\delta(x - y)\, \mathrm{d}y = \int F(x + \delta z) K(z)\, \mathrm{d}z.$$

Elementary calculus shows that F_δ is convex and infinitely often differentiable with limit F pointwise. Moreover, since K is even,

$$F_\delta(x) = \int \frac{F(x + \delta z) + F(x - \delta z)}{2} K(z)\, \mathrm{d}z \geq \int F(x) K(z)\, \mathrm{d}z = F(x)\ ,$$

by convexity of F. Finally, if F_δ fails to be strictly convex, we may add to F_δ the strictly convex function $x \mapsto \delta \|x\|^2$. □

We mention that the construction of F_δ given here is mainly for theoretical purposes, and may in practice be difficult to evaluate numerically due to the high dimensionality of the integral.

Proof of Theorem 3. Let $(F_\delta)_{\delta > 0}$ be a regularization of F such that $D^2 F_\delta$ is positive definite everywhere; cf. Theorem 4 and its proof. It may happen that $\limsup_{\|m\| \to \infty} F_\delta(m)/\|m\|^2 = \infty$, rendering quadratic approximation of F_δ from above impossible. Thus we modify the functions F_δ as follows: Let

$$c_\delta := \max_{\|m\| \leq \delta^{-1}} \lambda_{\max}(D^2 F_\delta(m))$$

with $\lambda_{\max}(A)$ denoting the largest eigenvalue of a symmetric matrix $A \in \mathbb{R}^{d \times d}$. Starting from the representation

$$F_\delta(m) = F_\delta(0) + \nabla F_\delta(0)^\top m + \int_0^1 m^\top D^2 F_\delta(tm) m\, (1 - t)\, dt,$$

we define

$$\tilde{F}_\delta(m) := F_\delta(0) + \nabla F_\delta(0)^\top m + \int_0^1 m^\top \min(D^2 F_\delta(tm), c_\delta I) m\, (1 - t)\, dt.$$

Here $\min(A, c_\delta I) \in \mathbb{R}^{d \times d}$ is obtained from the spectral representation of A by replacing each eigenvalue $\lambda_i(A)$ with $\min(\lambda_i(A), c_\delta)$. Note that $\tilde{F}_\delta$ is twice continuously differentiable with $\tilde{F}_\delta(0) = F_\delta(0)$, $\nabla \tilde{F}_\delta(0) = \nabla F_\delta(0)$ and $D^2 \tilde{F}_\delta = \min(D^2 F_\delta, c_\delta I)$. The hessian matrix is positive definite with largest eigenvalue never exceeding c_δ. In addition, $\tilde{F}_\delta = F_\delta$ on $\{m : \|m\| \leq \delta^{-1}\}$. Thus for sufficiently small $\delta > 0$, $\tilde{F}_\delta$ is regular, and a quadratic approximation of $\tilde{F}_\delta$ from above is given by

$$G_\delta(f, m) := \tilde{F}_\delta(f) + \nabla \tilde{F}_\delta(f)^\top (m - f) + c_\delta \|m - f\|^2 / 2.$$

□

Remark 1. In Definition 1 we assume that F_δ is strictly convex. This property is only required for notational convenience, because it guarantees uniqueness of the minimizer m_δ^*. A careful inspection of the proof of Theorem 1 shows, however, that convergence continues to hold if strict convexity is replaced with convexity. Only the assertion $d(m_\delta^*, M^*) \to 0$ has to be replaced by

$$\sup_{x \in M_\delta^*} \inf_{y \in M^*} \|x - y\| \to 0,$$

where $M_\delta^* := \operatorname{argmin}_{m \in \mathcal{C}} F_\delta(m)$. An analogous modification holds for Theorem 2.
We close the section with the following result, which shows under additional regularity conditions on F_δ and $\mathcal{C}$ geometric, or, more precisely, at least Q-linear convergence of the GIRLS algorithm (cf. Böhning & Lindsay, 1988, Theorem 4.1, for a related result).

Theorem 5. (Geometric convergence of the GIRLS algorithm).
Let $F_\delta : \mathbb{R}^d \to \mathbb{R}$ be coercive and twice continuously differentiable with positive definite hessian matrix $D^2F(m_\delta^*) =: A$. Further let $G_\delta : \mathbb{R}^d \times \mathbb{R}^d$ be a quadratic approximation of F_δ from above with hessian matrix $B(m_\delta^*) =: B$ as in (11). Then the GIRLS algorithm yields a sequence $(m_\delta^{(k)})_{k=0}^\infty$ converging to $m_\delta^* = \operatorname{argmin}_{\mathcal{C}} F_\delta$ such that

$$\limsup_{k \to \infty} \frac{\|m_\delta^{(k+1)} - m_\delta^*\|_A}{\|m_\delta^{(k)} - m_\delta^*\|_A} \;\le\; 1 - \lambda_{\min}\left(B^{-1}A\right) \in [0, 1).$$

Here $\|v\|_A := (v^\top A v)^{1/2}$, and $\lambda_{\min}(B^{-1}A) \in (0, 1]$ denotes the smallest eigenvalue of $B^{-1}A$.

Proof. According to Theorem 2, $\lim_{k\to\infty} m_\delta^{(k)} = m_\delta^*$. Since $\mathcal{C} = \mathbb{R}^d$, $\nabla F_\delta(m_\delta^*) = 0$ and

$$\begin{aligned} m_\delta^{(k+1)} &= m_\delta^{(k)} - B(m_\delta^{(k)})^{-1} \nabla F_\delta(m_\delta^{(k)}) \\ &= m_\delta^{(k)} - B(m_\delta^{(k)})^{-1} \int_0^1 D^2 F_\delta\left((1-t)m_\delta^* + t m_\delta^{(k)}\right)(m_\delta^{(k)} - m_\delta^*)\,\mathrm{d}t \\ &= m_\delta^{(k)} - B^{-1}A(m_\delta^{(k)} - m_\delta^*) + o\Big(\|m_\delta^{(k)} - m_\delta^*\|\Big)\,. \end{aligned}$$

Thus

$$\frac{\|m_\delta^{(k+1)} - m_\delta^*\|_A}{\|m_\delta^{(k)} - m_\delta^*\|_A} = \frac{\|(I - B^{-1}A)(m_\delta^{(k)} - m_\delta^*)\|_A}{\|m_\delta^{(k)} - m_\delta^*\|_A} + o(1)\,,$$

and for any vector $v \in \mathbb{R}^d$,

$$\begin{aligned}
\frac{\|(I - B^{-1}A)v\|_A^2}{\|v\|_A^2} &= \frac{v^\top (I - AB^{-1})A(I - B^{-1}A)v}{v^\top Av} \\
&= \frac{w^\top A^{-1/2}(I - AB^{-1})A(I - B^{-1}A)A^{-1/2}w}{\|w\|^2} \quad (\text{with } w := A^{1/2}v) \\
&= \frac{w^\top C^2 w}{\|w\|^2} \quad (\text{with } C := I - A^{1/2}B^{-1}A^{1/2}) \\
&\leq \lambda_{\max}(C^2) \, .
\end{aligned}$$

It follows from property (ii) of G_δ in Definition 2 that $B - A$ is nonnegative definite, which implies that $\lambda_i(B^{-1}A) = \lambda_i(A^{1/2}B^{-1}A^{1/2}) \in (0, 1]$. This entails that C is nonnegative definite with $\lambda_{\max}(C^2) = \lambda_{\max}(C)^2 = (1 - \lambda_{\min}(B^{-1}A))^2$. □

3.2 Proper choice of δ and the number of iterations

In practical applications the points m_δ^* are never calculated exactly. Instead after finitely many, say $I(\delta)$, iterations of (12) the iteration is terminated and the regularization parameter δ is decreased, e.g. replaced with $\delta/2$. An obvious question is how to choose the iteration numbers $I(\delta)$. We found empirically in most cases that for a fixed parameter $\delta > 0$, the values $F(m_\delta^{(k)})$ are decreasing for $k \leq k(\delta)$ and increasing in $k \geq k(\delta)$ for some fixed $k(\delta) \in \mathbb{N}$. Hence we may take

$$I(\delta) := \min\left(\left\{k \in \mathbb{N}_0 : F(m_\delta^{(k+1)})/F(m_\delta^{(k)}) \geq 1 - \epsilon\right\} \cup \{k_{\max}\}\right) \tag{13}$$

for a small constant $\epsilon > 0$ and a large maximal number $k_{\max}$. In the examples discussed subsequently, we found that for $\epsilon = 10^{-5}$ and $k_{\max} = 100$, the number $I(\delta)$ was never larger than 30, which seems to compensate for the fact that the sequence $m_\delta^{(k)}$ converges only geometrically. This is similar to numerical findings of an implementation of an algorithm by Lejeune & Sarda (1988, Section 5) for the median and various parametric regression models.

Having determined $I(\delta)$ and $m_\delta^{(I(\delta))}$ for one particular $\delta > 0$, we define $m_{\delta/2}^{(0)} := m_\delta^{(I(\delta))}$ and repeat the same procedure with $\delta/2$ in place of δ, provided that $I(\delta) > 0$. We proceed as long as F is decreased, otherwise we terminate the algorithm.

4 Regularization and quadratic approximation for different types of regression problems

In the subsequent data examples the target functional $F(m)$ is always of type (1) or (3), i.e.

$$F(m) = \sum_{i=1}^n \rho(r_i(m)) + \lambda P(m) \tag{14}$$

with $\lambda \geq 0$, where each residual $r_i(m)$ is an affine linear functional of $m \in \mathbb{R}^d$. Here each summand of F is regularized and approximated separately. We will start with an auxiliary result justifying the quadratic approximation (9).

Lemma 1. *Let $h : \mathbb{R} \to \mathbb{R}$ be even and twice differentiable such that h'' is non-negative and non-increasing on $[0, \infty)$. For $r, z \in \mathbb{R}$ define*

$$g(r, z) \ := \ h(r) + (h'(r)/r)(z^2 - r^2)/2,$$

where $h'(0)/0 := h''(0)$. Then $g(r, z) \geq h(z)$ with equality if $z = \pm r$.

Proof. One verifies easily that $g(r, z)$ is even in both arguments with $g(r, r) = h(r)$. Thus it suffices to show that $g(r, z) \geq h(z)$ for any $r, z \geq 0$. Now,

$$\begin{aligned} g(r,z) - h(z) &= g(r,z) - h(r) - (h(z) - h(r)) \\ &= (h'(r)/r)(z^2 - r^2)/2 - h'(r)(z - r) - \int_r^z (h'(t) - h'(r))\, dt \\ &= (h'(r)/r)(z - r)^2/2 - \int_r^z (h'(t) - h'(r))\, dt \\ &= \int_r^z \left(\tilde{h}(r, 0) - \tilde{h}(r, t)\right)(t - r)\, dt \\ &= \int_{\min(r,z)}^{\max(r,z)} \left(\tilde{h}(r, 0) - \tilde{h}(r, t)\right)|t - r|\, dt, \end{aligned} \tag{15}$$

where $\tilde{h}(r, t) := (h'(t) - h'(r))/(t - r)$ for $t \neq r$, and $\tilde{h}(r, r) := h''(r)$. One can deduce easily from h'' being non-increasing on $[0, \infty)$ that $\tilde{h}(r, \cdot)$ has the same property. Thus the integrand of (15) is non-negative. □

Let us first describe how to approximate ρ itself in three special cases. After this we will discuss several penalisations P in (14). Finally we comment on isotonic regression, an example with $\mathcal{C} \neq \mathbb{R}^d$.

Quantile regression. Let $\rho(z)$ be given by (2). This may be rewritten as

$$\rho(z) = |z| + (2p - 1)z.$$

Hence we utilize the functions h_δ and g_δ from (7) and (9), which yields the regularization

$$z \mapsto h_\delta(z) + (2p - 1)z$$

and by means of Lemma 1 the quadratic approximation

$$z \mapsto g_\delta(r, z) + (2p - 1)z = c_\delta(r) + h_\delta(r)^{-1} z^2/2 + (2p - 1)z$$

of $z \mapsto \rho(z)$, where $c_\delta(r)$ is an irrelevant constant.

L^q–regression. Let $\rho(z) := |z|^q$ for some $q \in [1, \infty)$. If $1 \le q < 2$, one may generalize definitions (7) and (9) immediately as follows:

$$\begin{aligned} h_\delta(z) &:= (z^2 + \delta)^{q/2}, \\ g_\delta(r, z) &:= h_\delta(r) + q(r^2 + \delta)^{1-q}(z^2 - r^2)/2 \\ &= c_\delta(r) + q(r^2 + \delta)^{1-q} z^2/2. \end{aligned}$$

Again it follows from Lemma 1 that $g_\delta(r, z) \ge h_\delta(z)$ with equality for $z = \pm r$.

In case of $q > 2$, the second derivative of $z \mapsto |z|^q$ is increasing in $|z|$ and unbounded, hence Lemma 1 cannot be applied directly. To circumvent this problem, one could redefine

$$h_\delta(z) := \begin{cases} |z|^q & \text{if } |z| \le \delta^{-1} \\ a_\delta + b_\delta |z| + q(q-1)\delta^{2-q} z^2/2 & \text{otherwise} \end{cases}$$

with constants a_δ, b_δ such that h_δ is twice continuously differentiable, and then use the quadratic approximation

$$g_\delta(r, z) := h_\delta(r) + h'_\delta(r)(z - r) + q(q-1)\delta^{2-q}(z - r)^2/2. \tag{16}$$

Logistic regression. For data sets with a covariable X and a dichotomous response $Y \in \{0, 1\}$, maximum likelihood estimation of $M(X) := \log[P(Y = 1 \mid X)/P(Y = 0 \mid X)]$ involves "residuals" $z = (1/2 - Y)M(X)$ and

$$\rho(z) := h(z) + z \quad \text{with} \quad h(z) := \log[e^z + e^{-z}]\,.$$

Note that h satisfies the conditions of Lemma 1 with $h'(r) = \tanh(r)$ and $h''(r) = 1 - \tanh(r)^2$. Thus regularization is superfluous, while quadratic approximation is straightforward. In this case, the well known IRLS algorithm results (McCullagh & Nelder 1989).

Roughness penalties. Let us start with two particular examples for $P(m)$. For given real numbers $x_1 < x_2 < \cdots < x_d$ let M be a function on $[x_1, x_d]$ and $m := (M(x_j))_{j=1}^d$. Then let

$$\begin{aligned} \mathrm{TV}^{(0)}(m) &:= \sum_{j=1}^{d-1} |m_j - m_{j+1}|\,, \\ \mathrm{TV}^{(1)}(m) &:= \sum_{j=2}^{d-1} |\Delta_j m| \quad \text{with} \quad \Delta_j m := \frac{m_{j+1} - m_j}{x_{j+1} - x_j} - \frac{m_j - m_{j-1}}{x_j - x_{j-1}}\,. \end{aligned}$$

If M is continuous and piecewise linear with knots in $\{x_1, \ldots, x_d\}$, then $\mathrm{TV}^{(0)}(m)$ and $\mathrm{TV}^{(1)}(m)$ are the total variation of M and its first derivative, respectively. One could also think about

smoother functions M and approximate the total variation of its second or higher order derivative by suitable divided differences of m.

Generally, let $P(m)$ be a sum of several functionals of the form

$$m \mapsto |v^\top m|$$

with a given vector $v \in \mathbb{R}^d \setminus \{0\}$. For instance, $\mathrm{TV}^{(0)}(m)$ involves

$$v_i = v_i^{(j)} := \begin{cases} 1 & \text{if } i = j, \\ -1 & \text{if } i = j+1, \\ 0 & \text{else} \end{cases}$$

for $1 \le j < d$, while $\mathrm{TV}^{(1)}(m)$ involves

$$v_i = v_i^{(j)} := \begin{cases} (x_j - x_{j-1})^{-1} & \text{if } i = j-1, \\ -(x_j - x_{j-1})^{-1} - (x_{j+1} - x_j)^{-1} & \text{if } i = j, \\ (x_{j+1} - x_j)^{-1} & \text{if } i = j+1, \\ 0 & \text{else,} \end{cases}$$

for $1 < j < d$. Now an obvious strategy is to regularize $m \mapsto |v^\top m|$ by $m \mapsto h_\delta(v^\top m)$ and approximate this quadratically by

$$m \mapsto g_\delta(v^\top f, v^\top m) = c_\delta(v^\top f) + h_\delta(v^\top f)^{-1} \left(v^\top m\right)^2 /2.$$

Often it is desirable to work with quadratic approximations $G(f,\cdot)$ whose Hessian matrix $B(f)$ is diagonal. For that purpose one can modify the quadratic term $Q(m) := \left(v^\top m\right)^2$ as follows:

$$\begin{aligned} Q(m) &= \left(v^\top f\right)^2 + 2f^\top vv^\top(m-f) + \left(v^\top(m-f)\right)^2 \\ &\le \left(v^\top f\right)^2 + 2f^\top vv^\top(m-f) + \|v\|^2 \sum_{i:v_i \neq 0} (m_i - f_i)^2 \\ &= c(v,f) - 2w(v,f)^\top m + \|v\|^2 \sum_{i:v_i \neq 0} m_i^2 \end{aligned}$$

for some irrelevant constant $c(v,f)$ and $w(v,f)_i := 1_{v_i \neq 0}\|v\|^2 f_i - v^\top f v_i$.

Isotonic regression. In some applications one seeks to minimize a functional such as (14) over all vectors in $\mathcal{C}_\nearrow := \{m \in \mathbb{R}^d : m_1 \le \cdots \le m_d\}$. In the simplest case, $d = n$ and $\rho(r_i(m)) = (Y_i - m_i)^q$ for some $q \in [1,\infty]$, where $q = \infty$ corresponds to supremum norm of $Y - m$. For this special case it is well-known (see Barlow & Ubhaya 1971) that an explicit solution exists only for $q = 1, 2, \infty$. In general, via regularization and suitable quadratic approximation from above, each updating step of the GIRLS algorithm involves minimisation of

$$G_\delta(f,m) \;=\; C(f) + \sum_{i=1}^d w_i(f)(m_i - b_i(f))^2$$

over all vectors $m \in \mathcal{C}_\nearrow$ with certain weights $w_i(f) > 0$, an irrelevant constant $C(f)$ and certain numbers $b_i(f)$. This minimisation problem can be solved explicitly by means of the PAVA.

5 A numerical example: Unimodal regression on a two-dimensional grid

In this example we illustrate the flexibility of the GIRLS algorithm, where we combine an unimodality constraint together with TV penalization. In order to reconstruct the spatial luminosity distribution of the Milky Way, the surface brightness of the Milky Way at the sky was measured on an equidistant grid of angles running in horizontal direction from -89.25° to 89.25° and in vertical direction from -29.25° to 29.25° in steps of 1.5°, respectively. The data set is part of the DIRBE experiment on board the COBE satellite (cf. Spergel et al. 1996 for a comprehensive description of the data, proper calibration and dust correction). Bissantz & Munk (2001) fitted a parametric model of the spatial luminosity distribution of the Milky Way to this data and a map of the resulting squared residuals is displayed there. This suggests that the variability of the data increases in the outer regions of the Milky Way. Moreover, due to this pronounced heteroscedasticity of the data, the reliability of any conclusions about physical properties (such as scale lengths) of the Milky Way from the parameters of a (parametric) model would be improved substantially if the variance surface of the data is known and used in the fitting process.

In the following we will consider estimation of the variance surface from the squared residuals by a penalized least squares fit with an unimodality constraint. The assumption of unimodality is suggested by physical reasoning. Such an estimate of the variance surface can then be used in a further investigation of the surface brightness data.

In more detail, the data considered here, denoted as Y_{ij}, $i = 1, ..., 120$, $j = 1, ..., 40$, are the squared residuals of the the parametric fit in Bissantz & Munk (2001) to the observed surface brightness data.

First, we consider a least squares fit $m = (m_{ij})_{i,j}$ to the data $Y = (Y_{ij})_{i,j}$. Namely, we want to minimize the sum $F(m)$ of

$$\|Y - m\|^2 = \sum_{ij} (Y_{ij} - m_{ij})^2 \,,$$

and the total variation penalty $\lambda \,\mathrm{TV}(m)$, where $\lambda > 0$ and

$$\mathrm{TV}(m) := \sum_{i=1}^{119} \sum_{j=1}^{40} |m_{i+1,j} - m_{ij}| + \sum_{i=1}^{120} \sum_{j=1}^{39} |m_{i,j+1} - m_{ij}| \,.$$

For the regularization F_δ and quadratic approximation $G_\delta(f, \cdot)$, the least squares term is kept unchanged, while each summand $|m_{(a)} - m_{(b)}|$ of $\mathrm{TV}(m)$ is treated as described in section 4:

First we regularize it by $h_\delta(m_{(a)} - m_{(b)})$. Then as a first quadratic approximation we use

$$g_\delta(f_{(a)} - f_{(b)}, m_{(a)} - m_{(b)}) = C_{\delta,(a),(b)}(f) + \frac{(m_{(a)} - m_{(b)})^2}{2h_\delta(f_{(a)} - f_{(b)})}.$$

Alternatively, one might replace the enumerator $(m_{(a)} - m_{(b)})^2$ with

$$2\Big(m_{(a)} - \frac{f_{(a)} + f_{(b)}}{2}\Big)^2 + 2\Big(m_{(b)} - \frac{f_{(a)} + f_{(b)}}{2}\Big)^2,$$

which is never less than $(m_{(a)} - m_{(b)})^2$ with equality for $m = f$. The advantages of this latter quadratic approximation are computational simplicity and feasibility of isotonic least squares algorithms when incorporating additional monotonicity constraints. Moreover, we impose in addition unimodality in vertical direction with minimum mode at the 0-line. Hence, in each step for each vertical half line the isotonic regression has to be calculated by means of some standard algorithm calculating the isotonic weighted least squares fits like PAVA. This is feasible with our second quadratic approximation. Figure 1 shows the resulting surface. Note that smoothness along the vertical direction of the estimated variance surface is already guaranteed by the unimodality constraint, and does not rely on a proper selection of the smoothing parameter λ. This relaxes significantly the requirements for a data-driven estimation of the variance surface, and, subsequently, of the spatial luminosity distribution of the Milky Way.

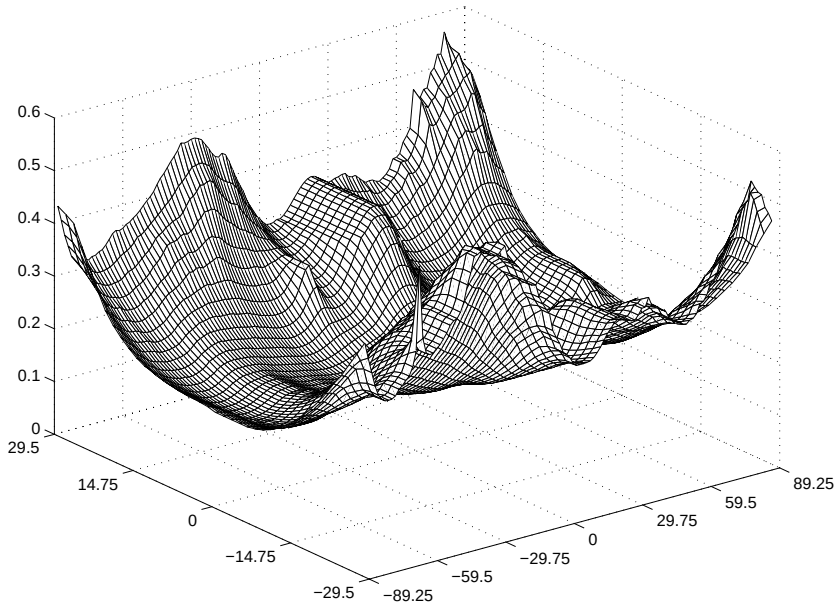

Figure 1: TV penalized fit with parameter $\lambda = 4$
under additional unimodality constraint.

ACKNOWLEDGEMENTS

The authors are thankful to both anonymous referees for their constructive criticism that helped to improve this note. Moreover, they are indepted to G. Jongbloed and Y. Vardi for helpful comments. Parts of this paper were written while L. Dümbgen was visiting Göttingen University as lecturer within the PhD Program 'Applied Statistics and Empirical Methods', and while N. Bissantz was visiting the University of Bern. Financial support of the Graduiertenkolleg 1023 "Identification in Mathematical Models" and the DAAD is gratefully acknowledged.

References

[1] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk, *Statistical Inference Under Order Restrictions. The Theory and Application of Isotonic Regression,* John Wiley & Sons, London, New York, Sydney, 1972.

[2] R. E. Barlow, and V. Ubhaya, *Isotonic approximation,* in Optimisation Methods in Statistics, J. S. Rustagi, ed., Academic Press, 1971, pp. 77-86.

[3] N. Bissantz, and A. Munk, *New statistical goodness of fit techniques in noisy inhomogeneous inverse problems. With application to the recovering of the luminosity distribution of the Milky Way,* Astron. & Astroph., 376(2001), pp. 735–744.

[4] D. Böhning, and B. Lindsay, *Monotonicity of quadratic-approximation algorithms,* Ann. Inst. Statist. Math., 40(1988), pp. 641–663.

[5] A. W. Bowman, and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations,* Oxford University Press, Oxford, UK, 1997.

[6] B. M. Brown, *Statistical uses of the spatial median,* J. Roy. Statist. Soc. Ser. B, 45(1983), pp. 25–30.

[7] B. M. Brown, P. Hall, and G. A. Young, *On the effect of inliers on the spatial median,* J. Multivariate Anal., 63(1997), pp. 88–104.

[8] D. Coleman, P. Holland, N. Kaden, V. Klema, and S. C. Peters, *A system of subroutines for iteratively reweighted least squares computations,* ACM Trans. Math. Softw., 6(1980), pp. 327–336.

[9] J. de Leeuw, and G. Michailidis, *Discussion article on the paper by Lange, Hunter & Yang (2000),* Journ. Comput. Graph. Statist., 9(2000), pp. 26–31.

[10] Y. Dodge, and J. Jurečková, *Adaptive Regression,* Springer-Verlag, New York, 2000.

[11] G. R. Ducharme, and P. Milasevic, *Spatial median and directional data,* Biometrika, 74(1987), pp. 212–215.

[12] U. Eckhardt, *Weber's problem and Weiszfeld's algorithm in general spaces,* Math. Program., 18(1980), pp. 186–196.

[13] W. Härdle, and J. S. Marron, *Bootstrap simultaneous error bars for nonparametric regression,* Ann. Statist., 19(1991), pp. 778–796.

[14] P. J. Huber, *Robust estimation of a location parameter,* Ann. Math. Stat., 35(1964), pp. 73-101.

[15] P. J. Huber, *Robust Statistics,* John Wiley & Sons Inc., New York, 1981.

[16] D. R. Hunter, and K. Lange, *Quantile Regression via an MM Algorithm,* Journ. Comput. Graph. Statist., 9(2000), pp. 60-77.

[17] I. N. Katz, *Local convergence in Fermat's problem,* Math. Program., 6(1974), pp. 89–104.

[18] K. Lange, D. R. Hunter, and I. Yang, *Optimization Transfer Using Surrogate Objective Functions,* Journ. Comput. Graph. Statist., 9(2000), pp. 1–20.

[19] R. Koenker, and G. Bassett, *Regression quantiles,* Econometrica, 46(1978), pp. 33–50.

[20] R. Koenker, P. Ng, and S. Portnoy, *Quantile smoothing splines,* Biometrika, 81(1994), pp. 673–680.

[21] H. R. Künsch, *Robust priors for smoothing and image restoration,* Ann. Inst. Statist. Math., 46(1994), pp. 1–19.

[22] H. W. Kuhn, *A note on Fermat's problem,* Math. Program., 4(1973), pp. 98–107.

[23] M. G. Lejeune, and P. Sarda, *Quantile regression: a nonparametric approach,* Comput. Statist. Data Anal., 6(1988), pp. 229–239.

[24] E. Mammen, J. S.Marron, B. A. Turlach, and M. P. Wand, *A general projection framework for constrained smoothing,* Statist. Sci., 16(2001), pp. 232–248.

[25] E. Mammen, and S. van de Geer, *Locally adaptive regression splines,* Ann. Statist., 25(1997), pp. 387–413.

[26] P. McCullagh, and J. Nelder, *Generalized Linear Models, 2nd ed.,* Chapman & Hall, London, 1989.

[27] D. P. O'Leary, *Robust regression computation using iteratively reweighted least squares,* SIAM J. Matrix Anal. Appl., 11(1990), pp. 466–480.

[28] S. Portnoy, *Local asymptotics for quantile smoothing splines,* Ann. Statist., 25(1997), pp. 414–434.

[29] D. A. Ratkowsky, *Nonlinear Regression Modelling,* Dekker, New York, 1983.

[30] T. Robertson, and P. Waltman, *On estimating monotone parameters,* Ann. Math. Statist., 39(1968), pp. 1030–1039.

[31] T. Robertson, F. T. Wright, and R. L. Dykstra, *Order Restricted Statistical inference,* John Wiley & Sons Ltd., Chichester, UK, 1988.

[32] R. T. Rockafellar, *Convex Analysis,* Princeton University Press, Princeton, N. J., 1970.

[33] D. N. Spergel, S. Malhorta, and L. Blitz, *Towards a three dimensional model of the Galaxy,* in ESO/MPA Workshop on spiral Galaxies in the Near-I, D. Minniti and H. W. Rix, eds., Springer-Verlag, 1996, pp. 128–137.

[34] Y. Vardi, and C.-H. Zhang, *The multivariate L_1-median and associated data depth,* Proc. Natl. Acad. Sci. USA, 97(2000), pp. 1423–1426.

[35] Y. Vardi, and C-H. Zhang, *A modified Weiszfeld algorithm for the Fermat-Weber location problem,* Math. Program. (Ser. A), 90(2001), pp. 559–566.

[36] H. Voß, and U. Eckhardt, *Linear Convergence of Generalized Weiszfeld's Method,* Computing, 25(1980), pp. 243-251.

[37] E. Weiszfeld, *Sur un problème de minimum dans l'espace,* Tohoku Math. J., 42(1936), pp. 274–280.

[38] E. Weiszfeld, *Sur le point pour lequel la somme des distances de n points donnés est minimum,* Tohoku Math. J., 43(1937), pp. 355–386.

[39] R. Wolke, and H. Schwetlick, *Iteratively reweighted least squares: algorithms, convergence analysis, and numerical comparisons,* SIAM J. Sci. Statist. Comput., 9(1988), pp. 907–921.