Applied Numerical Mathematics 157 (2020) 590–601

Contents lists available at ScienceDirect

Applied Numerical Mathematics

www.elsevier.com/locate/apnum

Large-scale regression with non-convex loss and penalty

Alessandro Buccini [a], Omar De la Cruz Cabrera [b], Marco Donatelli [c], Andrea Martinelli [c], Lothar Reichel [b,*]

[a] *Department of Mathematics and Computer Science, University Cagliari, 09124 Cagliari, Italy*
[b] *Department of Mathematical Sciences, Kent State University, Kent, OH 44242, USA*
[c] *Department of Science and High Technology, University of Insubria, 22100 Como, Italy*

ARTICLE INFO

Article history:
Received 20 March 2020
Received in revised form 26 May 2020
Accepted 4 July 2020
Available online 21 July 2020

Keywords:
Regression
Regularization
Robustness
Non-convex Optimization

ABSTRACT

We describe a computational method for parameter estimation in linear regression, that is capable of simultaneously producing sparse estimates and dealing with outliers and heavy-tailed error distributions. The method used is based on the image restoration method proposed in Huang et al. (2017) [13]. It can be applied to problems of arbitrary size. The choice of certain parameters is discussed. Results obtained for simulated and real data are presented.

1. Introduction

Linear regression is one of the basic tools of Statistics (see, e.g., [30]). The classical theory, based on ordinary least squares (OLS) estimates, has a long history and is well understood; however, it has important limitations. Two of those limitations are the requirement for non-collinear predictors and sensitivity to outliers.

Regularization has been used in Statistics, under different names, for a few decades. The goals have been diverse, but concentrated on three aspects: improving the stability of estimates, improving generalizability and prediction accuracy, and model selection. In linear regression, regularization has been successful in dealing with collinear predictors, producing "shrunken" and even sparse coefficient estimates (see, e.g., [29]).

Outliers can strongly affect parameter estimates in OLS; this is also true in regularized methods that use quadratic loss for the fidelity (fitness) term. In multiple regression, manual outlier identification can be particularly challenging, as several outliers can mask each other (see, e.g., [14]). This makes it imperative to use methods that can deal with outliers automatically.

In this article we propose a method for parameter estimation capable of simultaneously producing sparse estimates in the presence of collinearity and dealing with outliers and heavy-tailed error distributions. Our method is based on solving a non-convex optimization problem, and our algorithm can be applied to the solution of small and large problems.

The minimization method described has previously been applied to image reconstruction problems [13,17]. It is the purpose of the present paper to explore its application to regression problems and compare its performance under different

[*] Corresponding author.
E-mail addresses: alessandro.buccini@unica.it (A. Buccini), odelacru@kent.edu (O. De la Cruz Cabrera), marco.donatelli@uninsubria.it (M. Donatelli), andrea.martinelli@uninsubria.it (A. Martinelli), reichel@math.kent.edu (L. Reichel).

https://doi.org/10.1016/j.apnum.2020.07.006
0168-9274/© 2020 IMACS. Published by Elsevier B.V. All rights reserved.

parameter choices. We therefore present the method with notation that is commonly used in discussions on regression problems, which varies slightly from the notation in image reconstruction.

Remarks about notation. Throughout this paper, we define $\|\boldsymbol{z}\|_p = \left(\sum_{j=1}^{m} |z_j|^p\right)^{1/p}$ for $\boldsymbol{z} = [z_1, z_2, \ldots, z_m]^T \in \mathbb{R}^m$ and $p > 0$. The function $\boldsymbol{z} \mapsto \|\boldsymbol{z}\|_p$ is a norm for $p \geq 1$; we will be particularly interested in the situation when $0 < p < 1$, and will, for convenience, abuse terminology and refer to $\|\boldsymbol{z}\|_p$ as a norm also for the latter values of p, even though this is not a norm for $p < 1$. The superscript ^{T} denotes transposition.

2. Formulation of the estimation criterion

2.1. The regression problem

Consider the linear regression problem

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ with } E(\boldsymbol{\varepsilon}) = \boldsymbol{0} \text{ and } \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n, \tag{1}$$

where $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times m}$ contains the values of the m covariates on the n observations, $\boldsymbol{y} \in \mathbb{R}^n$ is the response, and I_n denotes the identity matrix of order n.

The OLS estimator, given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^m} ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2, \tag{2}$$

has several desirable properties that have made it extremely popular: it is the *best linear unbiased estimator* (BLUE) for model (1) by the Gauss-Markov Theorem (here “best” means minimum variance). Moreover, it coincides with the maximum likelihood estimator if the errors are Gaussian. In this case, the solution easily can be computed by QR factorization or singular value decomposition of the matrix X. The solution $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is unique only if $\mathrm{rank}(X) = m$, which, in particular, implies that $m \leq n$. Furthermore, the estimate becomes unstable (numerically) and highly variable (statistically) if X is badly conditioned due to approximate multiple collinearity among the covariates.

It is often of interest to find a solution to the regression problem also in situations when $m > n$, or when the covariates are highly correlated and exhibit multiple collinearity [29]. One approach is to get rid of some covariates until X becomes better conditioned, but this can involve an arduous selection problem. Regularized regression offers a way to obtain useful solutions to the regression problem and sometimes provides a method for model selection, i.e., selection of a subset of covariates to be included in the model.

Regularized regression can be understood as constraining the possible solutions $\boldsymbol{\beta}$. By Lagrangian relaxation, it results in the addition of a penalty term, multiplied by a regularization parameter. It can sometimes also be recast as a Bayesian approach, by imposing a prior distribution on $\boldsymbol{\beta}$. In general, these methods produce biased estimates, and one of the goals is to trade the increase in bias for a reduction in variance. Two important and popular methods are ridge regression [11] and Lasso [28].

Ridge regression [11] consists in adding a penalty term based on the ℓ_2-norm. Instead of (2), we consider the optimization problem

$$\hat{\boldsymbol{\beta}}_{\mathrm{ridge}(\lambda)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^m} \left\{ ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2 \right\}, \tag{3}$$

for some $\lambda > 0$. The solution can be expressed as

$$\hat{\boldsymbol{\beta}}_{\mathrm{ridge}(\lambda)} = (X^T X + \lambda I)^{-1} X^T \boldsymbol{y}.$$

Adding a positive multiple of the identity I to $X^T X$ is a straightforward way to obtain an invertible matrix even if $X^T X$ is singular. In the context of ill-posed problems, this approach is known as *Tikhonov regularization* [10]. The effect of the ridge penalty is to shrink the estimates towards zero; the shrinkage is larger when λ is larger.

Lasso [28] is a regularized regression method based on the optimization problem

$$\hat{\boldsymbol{\beta}}_{\mathrm{lasso}(\lambda)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^m} \left\{ ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \right\} \tag{4}$$

for $\lambda > 0$. The Lasso solution $\hat{\boldsymbol{\beta}}_{\mathrm{lasso}(\lambda)}$ tends to be sparse, i.e., it tends to have entries exactly equal to zero (larger values of λ lead to greater sparsity). This property has been exploited as a method for model selection. However, the minimization (4) also shrinks the non-zero components of the solution. A larger value of λ results in more shrinkage. Achieving the appropriate level of sparseness might lead to excessive shrinkage and bias [20].

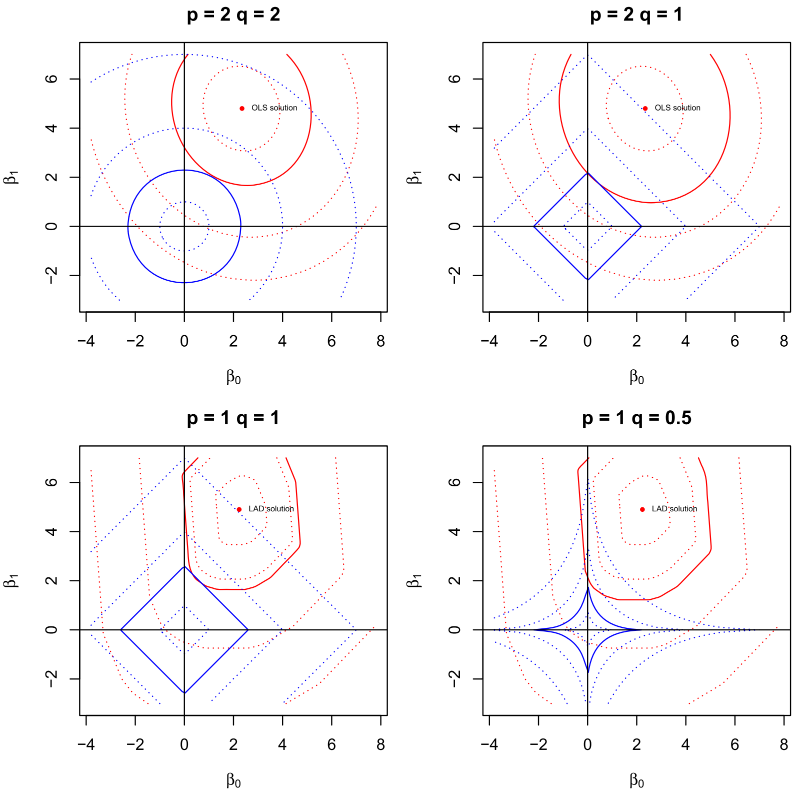

Fig. 1. Level curves of the fitness (red) and penalty (blue) terms, in a penalized regression setting. A solution to the optimization problem (5) is obtained when level curves meet and are tangent to one another (by the Lagrange multipliers method); the center of the fitness term level curves corresponds to the non-penalized solution: OLS for $p = 2$ and LAD (least absolute deviations) for $p = 1$. The top left picture corresponds to ridge regression, while the top right corresponds to Lasso. The bottom two figures use an ℓ_1-fidelity term ($p = 1$), producing non-ellipsoidal level curves; the sharper spikes in the penalty level curves obtained with $q < 1$ increase the chance of obtaining zero entries in $\hat{\boldsymbol{\beta}}$, even when the fitness level sets are non-ellipsoidal. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

2.2. ℓ_p-ℓ_q estimation

This estimation method encompasses (3) and (4) as particular cases (rescaling λ if necessary). We define the *ℓ_p-ℓ_q estimator for $\boldsymbol{\beta}$* as the solution to the optimization problem

$$\hat{\boldsymbol{\beta}}_{\mathrm{lplq}(p,q,\lambda)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^m} \left\{ \frac{1}{p} ||\boldsymbol{y} - X\boldsymbol{\beta}||_p^p + \frac{\lambda}{q} ||\boldsymbol{\beta}||_q^q \right\}, \tag{5}$$

where p, q, and λ are positive constants. The first term inside the argmin function is referred to as the *fidelity term* and the second as the *regularization term*. We are primarily interested in minimization problems (5) with $q < 1$ and $1 \le p \le 2$. Note that when $q < 1$ or $p < 1$, the minimization problem is non-convex.

2.3. Geometric intuition

Fig. 1 shows the geometric intuition behind ℓ_p-ℓ_q regularized estimation, including ridge regression and Lasso. By the Lagrange multiplier method, a solution to the optimization problem (5) is found where level curves for the fitness term and level curves for the penalty term meet in an “osculatory manner”. This includes the cases when one of the curves has a “spike” or a sharp point at the intersection. Since $p < 2$ leads to non-ellipsoidal level sets, the sharper spikes appearing on the level sets of the penalty term when $q < 1$ help increase the probability of zeros appearing in the solution vector, i.e., the solution being at the tip of a spike.

2.4. Bayesian justification

The definition of $\hat{\boldsymbol{\beta}}_{\mathrm{lplq}(p,q,\lambda)}$ in (5) can be justified with a Bayesian argument by using the generalized error distribution (GED) and maximum-a-posteriori estimation. The GED, first studied by Subbotin [1,27], is a location-scale family of distributions with density function

$$\xi_{\nu,\mu,\sigma}(x) = c_{\nu,\sigma}\exp\left(-\frac{|x-\mu|^\nu}{\nu\sigma^\nu}\right).$$

It includes the Gaussian and Laplacian (or double exponential) distributions as particular cases, for $\nu = 2$ and $\nu = 1$, respectively. When $\nu < 2$, we obtain distributions with heavier tails than the Gaussian distribution. This makes the GED useful for modeling situations in which the error term has a propensity for outliers [1].

We now set up a GED-based Bayesian version of the regression model (1). Assume that $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_m]^T$ is a random vector with prior distribution

$$\beta_j \overset{\text{i.i.d}}{\sim} f_{\boldsymbol{\beta}} = f_{q,0,\sigma_{\boldsymbol{\beta}}}, \qquad j = 1, 2, \ldots, m,$$

that $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are either fixed or independent of $\boldsymbol{\beta}$, and that the conditional density for $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^T$ is $f_{\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{x}} = \xi_{p,\boldsymbol{x}^T\boldsymbol{\beta},\sigma_{\boldsymbol{y}}}$.

After observing independent samples $y_1, y_2, \ldots, y_n$ with corresponding covariate vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, the prior density for $\boldsymbol{\beta}$ is updated to the posterior as follows:

$$\begin{aligned} f_{\boldsymbol{\beta}|\boldsymbol{y},X} &\propto \exp\left(-\frac{1}{p\sigma_{\boldsymbol{y}}^p}\sum_{i=1}^n |y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|^p\right)\exp\left(-\frac{1}{q\sigma_{\boldsymbol{\beta}}^q}\sum_{j=1}^m |\beta_j|^q\right) \\ &= \exp\left(-\frac{1}{p\sigma_{\boldsymbol{y}}^p}||\boldsymbol{y} - X\boldsymbol{\beta}||_p^p - \frac{1}{q\sigma_{\boldsymbol{\beta}}^q}||\boldsymbol{\beta}||_q^q\right). \end{aligned}$$

Then the maximum-a-posteriori estimate for $\boldsymbol{\beta}$ is obtained by minimizing

$$\frac{1}{p}||\boldsymbol{y} - X\boldsymbol{\beta}||_p^p + \frac{\sigma_{\boldsymbol{y}}^p}{\sigma_{\boldsymbol{\beta}}^q}\frac{1}{q}||\boldsymbol{\beta}||_q^q,$$

which is the optimization problem (5) with $\lambda = \sigma_{\boldsymbol{y}}^p/\sigma_{\boldsymbol{\beta}}^q$.

2.5. ℓ_p-ℓ_q regularization and thresholding

Notice that both the prior and posterior for $\boldsymbol{\beta}$ are continuous distributions in Section 2.4, and therefore assign probability zero to the event of having any component of $\boldsymbol{\beta}$ exactly equal to zero. However, the use of maximum-a-posteriori estimation results in a positive probability for such an outcome whenever $q \le 1$, since the objective function in (5) is not differentiable, as illustrated by the contour plots in Fig. 1. The following example, extremely simple as it is, clarifies this point and establishes a connection with thresholding functions.

Consider this simple situation: Let $y = \beta + \varepsilon$, with $\beta \sim \xi_{q,0,\lambda}$ and $\varepsilon \sim \xi_{p,0,1}$, independently. The maximum-a-posteriori estimate for β given y corresponds to the ℓ_p-ℓ_q estimate for a regression problem with intercept but no covariates, and only one observation: the objective function in (5) is

$$g_y(\beta) = \frac{1}{p}|y - \beta|^p + \lambda\frac{1}{q}|\beta|^q,$$

and the posterior distribution of β given y is proportional to $\exp(-g_y(\beta))$.

We now describe the behavior of $g_y(\beta)$ and its minimizer $\hat{\beta}_{\text{lplq}} = \hat{\beta}_{\text{lplq}(p,q,\lambda,y)}$, and how that behavior changes with y. Fig. 2 shows the behavior of the function g_y for different values of p and q. Without loss of generality, we may assume that $y > 0$ (the case $y < 0$ is symmetric). The function g_y has no critical points for $\beta < 0$ or $\beta > y$. If $q \le 1$, then g_y is not differentiable at $\beta = 0$, and if $p \le 1$, then it is not differentiable at $\beta = y$, while for $0 < \beta < y$, we have

$$g_y'(\beta) = -(y - \beta)^{p-1} + \lambda\beta^{q-1}.$$

If either $p, q > 1$ or $p, q < 1$, so that $p - 1$ and $q - 1$ have the same sign, then there is a unique critical point of g_y in $(0, y)$: when $p, q > 1$, it is the global minimizer, but when $p, q < 1$, it is a local maximum, and the global minimizer is found at either 0 or y; see Fig. 2, Panels A, B, and C.

When $p = 2$ and $q = 1$ (the Lasso case), we have $g_y'(\beta) > 0$ for all $\beta \in (0, y)$ (resulting in $\hat{\beta}_{\text{lplq}} = 0$) whenever $y < \lambda$, and $\hat{\beta}_{\text{lplq}} = y - \lambda$ whenever $y > \lambda$. In other words, $y \mapsto \hat{\beta}_{\text{lplq}}$ is the *soft thresholding function* with threshold λ, defined by $y \mapsto \max\{0, |y| - \lambda\} \cdot \text{sign}(y)$; see Fig. 3, Panel C.

The case $0 < q < p \le 1$ produces the following behavior: For $y < (\lambda p/q)^{1/(p-q)}$, we obtain $\hat{\beta}_{\text{lplq}} = 0$, while for $y > (\lambda p/q)^{1/(p-q)}$, we have $\hat{\beta}_{\text{lplq}} = y$. That is, $y \mapsto \hat{\beta}_{\text{lplq}}$ is the *hard thresholding function*, with threshold $(\lambda p/q)^{1/(p-q)}$ defined by $y \mapsto y \times 1_{\{y > \text{threshold}\}}$.

The situation when $p \le q \le 1$ does not seem to be useful, as the result is either a reversed hard thresholding (zero for large values of y), constant equal to either 0 or y (depending on the value of λ), or undefined because of multiple global

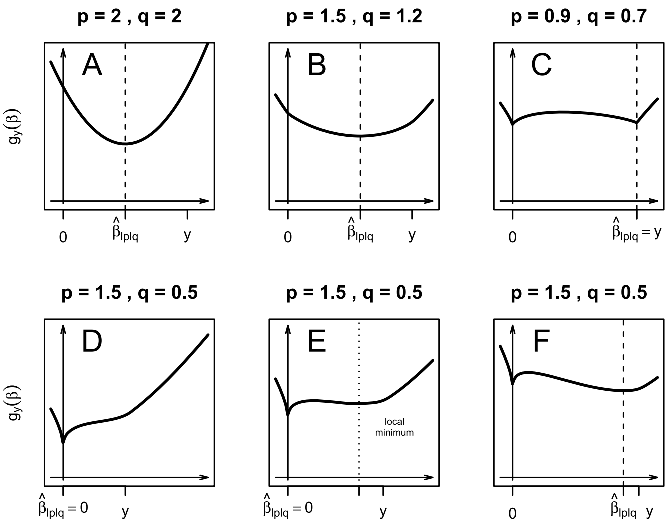

Fig. 2. Plots of the objective function $g_y(\beta)$ from the simple example in Section 2.5 for different values of p, q, and y. Vertical dotted lines mark local minima (other than at $\beta = 0$). Panels D, E, and F correspond to the situation of most interest when $0 < q < 1 < p \leq 2$: Small values of $|y|$ result in the minimum being attained at 0; larger values of $|y|$ lead to the appearance of another local minimum (Panel E). Finally, even larger values of $|y|$ make that local minimum become the global minimum.

minima. The case $1 < q \leq p \leq 2$ results in functions that "shrink" y, with more shrinkage for small values of y if $q < p$ (see Fig. 3, Panels A and B), but without inducing sparsity. In general, $p < q$ does not lead to useful results, as we obtain a function that expands small values of y, or the reverse hard thresholding function mentioned above.

The case of main interest is $0 < q < 1 < p \leq 2$. (See Fig. 2, Panels D, E, and F, and Fig. 3, Panels D, E, and F.) There is an inflection point for g_y between 0 and y; as y grows, the inflection point becomes a critical point, and immediately generates a local maximum (closer to 0) and a local minimum (closer to y). As y keeps growing, the value of g_y at that local minimum eventually becomes smaller than $g_y(0)$, becoming the global minimum. That is, $\hat{\beta}_{\text{lplq}}$ as a function of y has a jump discontinuity. The following proposition summarizes the results for this case.

Proposition 1. *Assume $0 < q < 1 < p \leq 2$, and consider the function $U : y \mapsto \hat{\beta}_{\text{lplq}}$, as above.*

(1) *There exists $t > 0$ (which depends on p, q, and λ), such that U is identically equal to zero for $|y| < t$.*
(2) *U has jump discontinuities at $-t$ and t.*
(3) *For $|y| > t$, $U(y)$ has the same sign as y, and $|U(y)| < |y|$.*
(4) *A lower bound for t is $\lambda^{1/(p-q)}$.*
(5) *An upper bound for t is $(\lambda p/q)^{1/(p-q)}$.*
(6) *As $|y| \to \infty$, $|y - U(y)| = |y| - |U(y)| \to 0$.*

Proof. Without loss of generality, we may assume that $y > 0$. Then g_y is decreasing on $(-\infty, 0]$ and increasing on $[y, \infty)$. Notice that $\lim_{\beta \to 0^+} g'_y(\beta) = +\infty$. Therefore any local minimum in $(0, y)$ is away from zero. For small values of y, g_y is increasing on $(0, y)$: If there is a $0 < \beta < y$ such that $g'_y(\beta) = 0$, that is, $(y - \beta)^{p-1} = \lambda \beta^{q-1}$, then $y^{p-1} > \lambda y^{q-1}$, from which we conclude that $y > \lambda^{1/(p-1)}$. Taking

$$t = \sup\{y : U(y') = 0, \forall y' \in (0, y)\}$$

shows (1) and (4). For (5), $y = (\lambda p/q)^{1/(p-q)}$ is the value for which $g_y(y) = g_y(0)$, which implies that the global minimum was achieved at a value $0 < \hat{\beta}_{\text{lplq}} < y$, since $g'_y(y) > 0$. This also shows (2), since $t < \infty$, and (3). For (6), notice first that g''_y has at most one zero and, therefore, there are at most two critical points. Now let $\epsilon > 0$. Since $q - 1 < 0$, if y is large enough, then $\lambda(y - \epsilon)^{q-1} < \epsilon^{p-1}$. This implies that $g'_y(y - \epsilon) < 0$, and it follows that the global minimum is achieved at a value $\hat{\beta}_{\text{lplq}} > y - \epsilon$. □

Fig. 3 shows a few thresholding and shrinkage functions for different combinations of values of p and q.

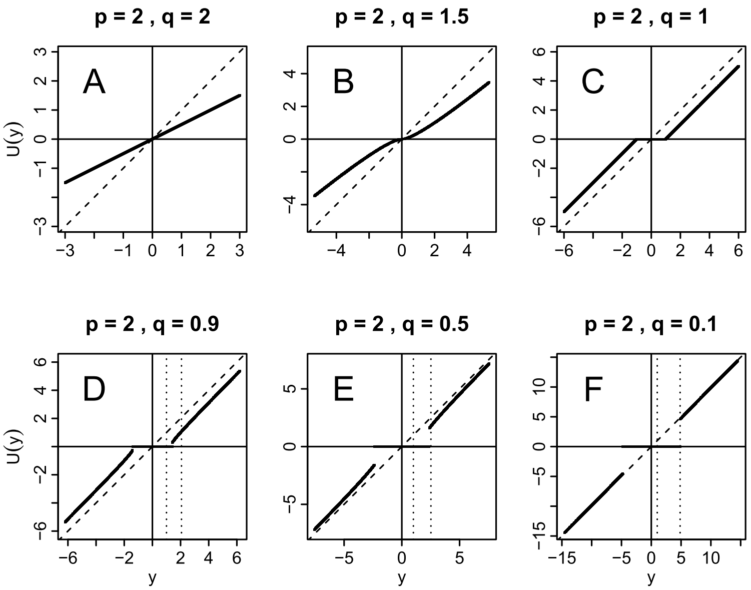

Fig. 3. Plots of the threshold functions $U(y)$ from Section 2.5 for $p = 2$ and different values of q. The diagonal dashed line is the identity. Panel A corresponds to ridge regression and gives shrinkage by a scalar factor. Panel C corresponds to lasso. The function U in this case gives soft thresholding. Panel B shows an intermediate function that shrinks small values more, but not all the way to zero. The Panels D, E, and F correspond to the situation of most interest when $0 < q < 1 < p \leq 2$: vertical dotted lines mark the theoretical bounds (lower and upper) for the threshold t from Proposition 1. As q gets closer to 0, the thresholding function approaches the hard thresholding function $h(y) = y1_{\{|y| \geq t\}}$.

3. Numerical methods

This section outlines the method for the solution of the minimization problem (5) that is used for the computed examples in Sections 4 and 5. Further details can be found in [13]. The minimization problem (5) can be written in more general form as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m} \mathcal{J}(\boldsymbol{\beta}), \qquad \mathcal{J}(\boldsymbol{\beta}) = \Phi_{\text{fid}}(\boldsymbol{\beta}) + \lambda\, \Phi_{\text{reg}}(\boldsymbol{\beta}), \tag{6}$$

where $\Phi_{\text{fid}}(\boldsymbol{\beta})$ and $\Phi_{\text{reg}}(\boldsymbol{\beta})$ are referred to as the *fidelity term* and the *regularization term*, respectively, and the parameter $\lambda > 0$ controls the trade-off between these terms; see [4,5] for a discussion on how to determine λ in the context of ill-posed inverse problems. For the subsequent analysis, it is convenient to write the fidelity and regularization terms in the form

$$\Phi_{\text{fid}}(\boldsymbol{\beta}) = \frac{1}{p}\sum_{i=1}^{n} \phi_p\big((X\boldsymbol{\beta} - \boldsymbol{y})_i\big), \quad \Phi_{\text{reg}}(x) = \frac{1}{q}\sum_{j=1}^{m} \phi_q\big((\boldsymbol{\beta})_j\big), \tag{7}$$

with the function $\phi_z : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ given by

$$\phi_z(t) = |t|^z, \quad z \in \mathbb{R}\,, \tag{8}$$

where $0 < p, q \leq 2$. We remark that the minimization problem discussed in [13] allows a regularization matrix L in the regularization term, but the application discussed in the present paper does not require this generality.

We note that the model (6)–(8) is convex and smooth when $1 < p, q \leq 2$, but nonconvex and nonsmooth when $0 < p < 1$ or $0 < q < 1$. Minimization problems of the type (6)-(8) appear in many applications in different areas, including numerical linear algebra [2,33], compressive sensing [6–9,19], and image restoration [8,13,17,18,25].

In the setting of linear regression, the matrix X in general is not square, and contains the covariate data. While in classical linear regression X is required to have linearly independent columns, in modern statistical applications that often fails, and often there are more columns than rows.

We remark that the structure of the matrices X that arise in linear regression is quite different from the structure encountered in image processing applications [4,5,13,17], where X is the discretization of a blurring operator. For instance, if the blur is space invariant, then the matrix X is square with a Toeplitz-type structure, which however may be modified due to boundary conditions. We found that models (6)–(8), that arise from linear regression, often are more difficult to solve than models from image restoration, in the sense that it is more important for the former models that a good initial approximate solution be available. We compute a suitable initial solution by solving a convex ℓ_2-ℓ_1 minimization problem.

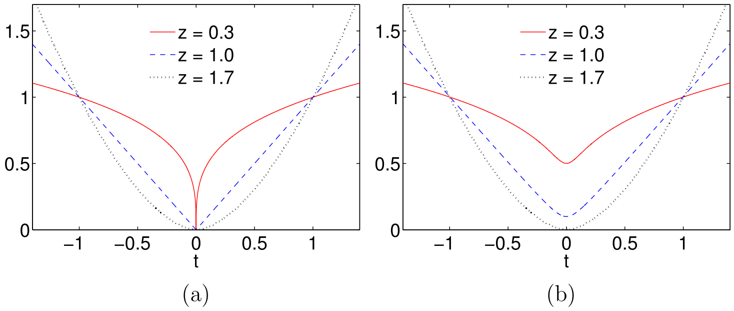

Fig. 4. (a) Plots of the penalty function $\phi_z(t)$ defined in (8) for some z-values, and (b) of the associated smoothed functions $\phi_{z,\varepsilon}(t)$ defined in (9) with $\varepsilon = 0.1$.

One of the most popular and effective methods for the solution of ℓ_p-ℓ_q minimization problems of the form (6)–(8), when $0 < p, q \le 2$ with $pq < 4$, is the iteratively reweighted norm (IRN) algorithm [23], also known as the iteratively reweighted least-squares (IRLS) algorithm [31]. This solution approach is equivalent to the (multiplicative) half-quadratic method [8] and to the gradient linearization iterative procedure [21]. The IRN method solves a sequence of penalized weighted least-squares problems that differ from each other by diagonal weighting matrices. Each one of these least-squares problems is solved by the conjugate gradient (CG) algorithm; see [23,24].

Based on the observation that the weighting matrices generated by the IRN method do not change very quickly during the iterations, two alternative approaches have recently been proposed in [13,17]. They extend the generalized Krylov subspace (GKS) method proposed in [16] for the case $p = q = 2$ to the situation when $0 < p, q \le 2$ and $pq < 4$. Instead of solving each least-squares problem generated independently by the conjugate gradient method, these methods use a generalized Krylov subspace determined by solving previously generated least-squares problems to solve the new least-squares problem at hand. Computed examples in [17] illustrate that this approach may require significantly fewer matrix-vector product evaluations than the IRN scheme described in [23,24].

The GKS methods in [13,17] are majorization-minimization (MM) methods. They replace the original, possibly nonconvex, ℓ_p-ℓ_q problem (6)-(8) by a sequence of simple convex weighted least-squares problems. The kth iteration of the MM approach consists of two steps: A majorization step, that generates a surrogate convex quadratic functional that majorizes the ℓ_p-ℓ_q functional, and a minimization step that finds a minimizer of the majorant. Two quadratic majorization techniques are described in [13], one of them is the “adaptive aperture” method, which we will outline. It is used in the computed examples of this paper.

Other solution methods for solving the problem (6)-(8) have been described in the literature; see, e.g., [7,12,22,32,33]. It is outside the scope of the present paper to compare these methods. Here we only note that the method used allows flexibility in the choice of p and q, which makes it attractive for many problems.

We describe the construction of quadratic majorants. Further details can be found in [13].

Definition 1. Let $\mathcal{G}(\boldsymbol{\beta}) : \mathbb{R}^m \to \mathbb{R}$ be a continuously differentiable function. Then the function $\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma}) : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is said to be a quadratic tangent majorant for $\mathcal{G}(\boldsymbol{\beta})$ if for any $\boldsymbol{\gamma} \in \mathbb{R}^m$ the following conditions hold:

c1) $\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is quadratic in $\boldsymbol{\beta}$,
c2) $\mathcal{Q}(\boldsymbol{\gamma}, \boldsymbol{\gamma}) = \mathcal{G}(\boldsymbol{\gamma})$,
c3) $\nabla_{\boldsymbol{\beta}} \mathcal{Q}(\boldsymbol{\gamma}, \boldsymbol{\gamma}) = \nabla_{\boldsymbol{\beta}} \mathcal{G}(\boldsymbol{\gamma})$,
c4) $\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \ge \mathcal{G}(\boldsymbol{\beta})$ for all $\boldsymbol{\beta} \in \mathbb{R}^m$,

where $\nabla_{\boldsymbol{\beta}}$ denotes the gradient with respect to the variable $\boldsymbol{\beta}$.

The determination of quadratic tangent majorants requires the function $\mathcal{G}$ to be continuously differentiable; cf. condition c3) above. We therefore smooth the functional $\mathcal{J}(\boldsymbol{\beta})$ when $0 < p \le 1$ or $0 < q \le 1$. A popular smoothed version of ϕ_z is given by

$$\phi_{z,\varepsilon}(t) = \left(\sqrt{t^2 + \varepsilon^2}\right)^z \quad \text{with} \quad \begin{cases} \varepsilon > 0 \text{ for } 0 < z \le 1, \\ \varepsilon = 0 \text{ for } 1 < z \le 2. \end{cases} \tag{9}$$

To allow quadratic majorization, we replace the possibly nonsmooth original ℓ_p-ℓ_q minimization problem (6)-(8) by the smoothed minimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^m} \mathcal{J}_\varepsilon(\boldsymbol{\beta}), \qquad \mathcal{J}_\varepsilon(\boldsymbol{\beta}) = \frac{1}{p} \sum_{i=1}^{n} \phi_{p,\varepsilon}\big((X\boldsymbol{\beta} - \mathbf{y})_i\big) + \frac{\lambda}{q} \sum_{j=1}^{s} \phi_{q,\varepsilon}\big((\boldsymbol{\beta})_j\big), \tag{10}$$

where a common value $\varepsilon > 0$ of the smoothing parameter is used for the penalty functions in the fidelity and regularization terms. Figs. 4(a)-(b) show the original and smoothed penalty functions ϕ_z and $\phi_{z,\varepsilon}$, respectively, for several values of the parameter z of interest in this paper. Note that the function ϕ_z is not smoothed when $1 < z \leq 2$.

One can show that the quadratic functional

$$\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}) = \frac{1}{2} \left\| (W_{\text{fid}}^{(k)})^{1/2} (X\boldsymbol{\beta} - \boldsymbol{y}) \right\|_2^2 + \frac{\lambda}{2} \left\| (W_{\text{reg}}^{(k)})^{1/2} L\boldsymbol{\beta} \right\|_2^2 + c,$$

is a tangent majorant for $\mathcal{J}_\varepsilon(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}$. All terms that are independent of $\boldsymbol{\beta}$ are collected in the term c, and the matrices $W_{\text{fid}}^{(k)}$ and $W_{\text{reg}}^{(k)}$ are defined by

$$W_{\text{fid}}^{(k)} = \text{diag}(\boldsymbol{w}_{\text{fid}}^{(k)}), \quad \boldsymbol{w}_{\text{fid}}^{(k)} = \left(\left(X\boldsymbol{\beta}^{(k)} - \boldsymbol{y} \right)^2 + \varepsilon^2 \right)^{p/2-1},$$

$$W_{\text{reg}}^{(k)} = \text{diag}(\boldsymbol{w}_{\text{reg}}^{(k)}), \quad \boldsymbol{w}_{\text{reg}}^{(k)} = \left(\left(L\boldsymbol{\beta}^{(k)} \right)^2 + \varepsilon^2 \right)^{q/2-1},$$

where all the operations are meant element-wise; see [13] for details.

The minimization step in the kth iteration of the MM-GKS method can be written as

$$\boldsymbol{\beta}^{(k+1)} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^m} \left[\left\| (W_{\text{fid}}^{(k)})^{1/2} (X\boldsymbol{\beta} - \boldsymbol{y}) \right\|_2^2 + \lambda \left\| (W_{\text{reg}}^{(k)})^{1/2} \boldsymbol{\beta} \right\|_2^2 \right].$$

This minimization problem has a unique solution if

$$\text{Ker}\left(X^T W_{\text{fid}}^{(k)} X \right) \cap \text{Ker}\left(W_{\text{reg}}^{(k)} \right) = \{\mathbf{0}\}, \tag{11}$$

where $\text{Ker}(M)$ denotes the null space of the matrix M. This requirement generally is satisfied. The MM method is shown in [13] to determine a stationary point of the functional (10) when condition (11) holds.

4. Simulations

This section describes a simulated numerical experiment to show the performance of the minimization method of Section 3. We apply the method to minimization problems with synthetic data and compare its performance to the Lasso method.

We first describe how we generated the data. The matrix $X \in \mathbb{R}^{200\times 250}$ is determined by using independent draws from the standard normal distribution. The vector $\boldsymbol{\beta} \in \mathbb{R}^{250}$ is such that 10 entries are chosen randomly with uniform distribution with standard deviation 1, while the remaining 240 entries are set to zero. Starting from X and $\boldsymbol{\beta}$, we computed $\boldsymbol{y}$ according to

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\text{GED}} + \boldsymbol{\varepsilon}_{\text{Gauss}},$$

where $\boldsymbol{\varepsilon}_{\text{GED}}$ is a GED variable with $p = 1.5$ and $\sigma = 1$, and $\boldsymbol{\varepsilon}_{\text{Gauss}}$ is such that each entry is the realization of a Gaussian variable with 0 mean and scaled such that $\|\boldsymbol{\varepsilon}_{\text{Gauss}}\|_2 = 0.01\|X\boldsymbol{\beta}\|_2$.

To generate random numbers from the GED, we exploit the following property: If $S \sim U(\{-1, 1\})$ and $Y \sim Gamma(1/\nu, 1/\nu)$ are independent, then $\sigma S Y^{1/\nu} + \mu$ has a GED distribution with parameters (μ, ν, σ). This is easy to prove; the result also can be found in [1, Eq. (4.2)].

Since the noise is a mixture of Gaussian and GED noise with $p = 1.5$, we set $p = 1.8$ and $q = 0.5$, and use cross validation to select the parameter λ; we refer to Stone [26] for a discussion on cross validation and to [5] for an application of cross validation to ℓ_p-ℓ_q minimization in the context of image reconstruction. The smoothing parameter in (10) is $\varepsilon = 10^{-3}$. We will set all entries of the computed solutions that are smaller than ε in magnitude to zero.

Since $q < 1$, the functional (10) is non-convex. Therefore, the computed solution may depend on the initial guess. To ensure that the initial guess is not too far from a desirable stationary point, we use the solution of the Lasso minimization as initial guess for the MM method.

Fig. 5 and Table 1 compare the method of Section 3 to the Lasso method and ℓ_2-ℓ_q minimization. Fig. 5 displays the computed vectors $\boldsymbol{\beta}$ and a scatter plot, i.e., the graphs obtained by plotting the magnitude of the computed coefficients against the magnitude of the exact ones. Visual comparison of the computed coefficients shows that Lasso fails to recognize as 0 most of the zero entries of β, while both the ℓ_p-ℓ_q and ℓ_2-ℓ_q minimization methods are able to better identify the zero elements of $\boldsymbol{\beta}$. However, the ℓ_2-ℓ_q method does not recognize as zeros 6 entries that are supposed to be nonvanishing and sets to 0 one nonzero entry of $\boldsymbol{\beta}$, while ℓ_p-ℓ_q only recognize 2 additional nonzero entries and does not fail to identify any nonzero entry of $\boldsymbol{\beta}$. The scatter plot shows that, due to the use of the 1-norm in Lasso, this method shrinks the nonzero coefficients. Shrinkage also can be observed in the solution computed by the ℓ_2-ℓ_q method, but to a lesser extent. The method of Section 3 is able to impose sparsity without shrinkage.

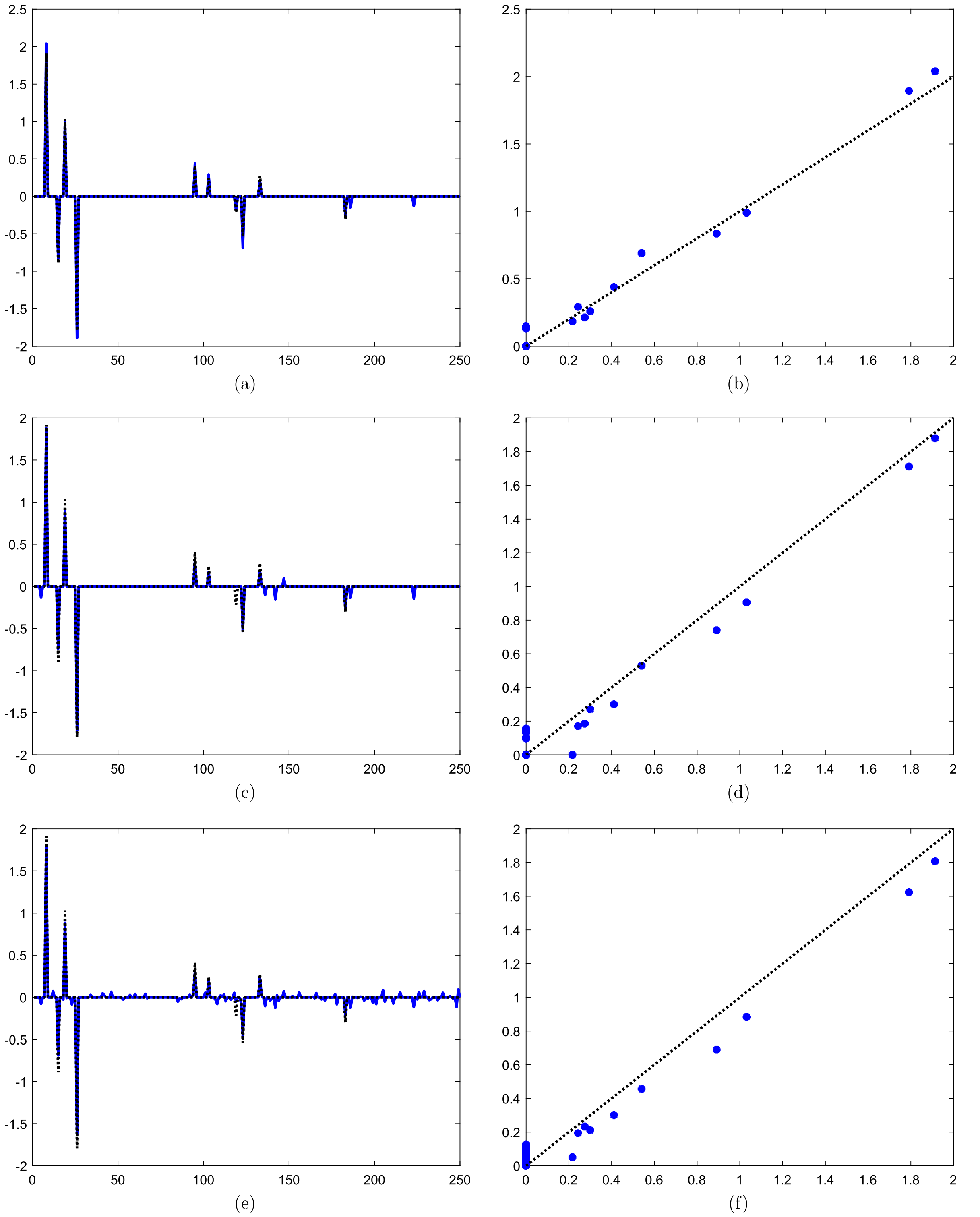

Fig. 5. Reconstructions and scatter plots. Panels (a) and (b) report the reconstructions obtained with the ℓ_p-ℓ_q method, and associated scatter plots. Panels (c) and (d) report reconstruction determined by the ℓ_2-ℓ_q method and the corresponding scatter plot. Panels (e) and (f) report the reconstruction and the scatter plot obtained with Lasso. The exact solution is displayed as a dotted line, while the solid line represents the computed approximate solution.

Table 1
Comparison of the results obtained with Lasso, ℓ_2-ℓ_q, and ℓ_p-ℓ_q in the computed example. For each method we report the number of non-zero entries (NNZ), the relative restoration error (RRE), and the residual error.

Method	NNZ	RRE	Residual error
ℓ_p-ℓ_q	12	0.10434	0.11604
ℓ_2-ℓ_q	15	0.15308	0.15538
Lasso	118	0.19842	0.20427

Table 1 shows the number of non-zero entries identified in each computed solution. Recall that we consider any entry of the computed solution that is of magnitude smaller than ε as zero. The table also shows the relative reconstruction error

$$\mathrm{RRE}(\hat{\boldsymbol{\beta}}) = \frac{\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_2}{\|\boldsymbol{\beta}\|_2},$$

where $\boldsymbol{\beta}$ denotes the exact solution of the problem and $\hat{\boldsymbol{\beta}}$ the computed one. Finally, Table 1 displays

$$\text{Residual error}(\hat{\boldsymbol{\beta}}) = \frac{\|X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}}\|_2}{\|X\boldsymbol{\beta}\|_2}.$$

Moreover, we consider the errors of type I and II, the first one is the percentage of nonzero entries of $\boldsymbol{\beta}$ that are set to 0 in the reconstruction and the latter is the percentage of nonzero entries in the reconstruction that are actually zero in the exact $\boldsymbol{\beta}$.

Table 1 shows the ℓ_p-ℓ_q method of Section 3 to give more accurate approximations $\hat{\boldsymbol{\beta}}$ of the desired solution $\boldsymbol{\beta}$ in terms of error $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$ and of residuals.

5. Example with real data: genomic association study for liver metabolites in mice

In this section, we use a real data set of fairly large size to illustrate the combined effect of using $q < 1$ for sparsity and $p < 2$ for dealing with outliers. The data are part of a genomic study in mice [15]. A total of 246 males, from 99 strains of inbred and recombinant inbred mice, were assayed using microarrays to measure the expression level of 22,716 genes/transcripts. Separately, the levels of 283 metabolites were also measured for those mice, with the goal of establishing associations between gene expression levels and metabolite levels.

For our example, we use only one of the metabolites, *pyridoxate*, which produced the strongest association result in the original study. Rather than performing 22,716 tests of association, as is customary, we use a single large-scale regression. Thus, our regression problem has 246 observations and 22,716 covariates, resulting in a design matrix X of size $246 \times 22,717$ (including the intercept).

As expected from the discussion in Section 2.5, the results for $p < 1$ are highly variable and often non-optimal. This difficulty does not arise when we set $q < 1$. In this example, we used values of p between 1.25 and 2.

We fitted 392 ℓ_p-ℓ_q models, using all combinations of values for p, q, and λ in Table 2. After fitting the model with $p = q = 2$ and $\lambda = 0.8$ (this corresponds to ridge regression), subsequent fits used available computed solutions with the closest p- q- and λ-values as starting points for the iterations. Note that, differently from the simulated case of Section 4, we do not use cross validation to determine the value of λ. In this way we are able to show the robustness of the proposed approach with respect to the choice of this parameter.

All the models identified *aldehyde oxidase 3* (*Aox3*) as the gene whose expression is most strongly associated to pyridoxate levels, in agreement with the original study. Results for the other genes vary.

Fig. 6 shows residuals-versus-fitted-values plots for four models that illustrate the effect of using $p < 2$. All models were fitted with $q = 0.75$ to encourage sparseness of the computed solution. The models A and C were fitted using $p = 2$, while the models B and D used $p = 1.25$. In order to make meaningful comparisons, the values of λ were chosen so that the computed solutions for models B and C have similar residuals of about the same ℓ_1-norm (expressed here as Mean Absolute Deviation, MAD), while at the same time model A has the same level of sparsity as model B (about 89 non-zero coefficients), and model D has the same level of sparsity as C (about 125 non-zero coefficients). Notice that for the same level of sparsity, $p = 1.25$ results in reduced MAD; similarly, for the same MAD, $p = 1.25$ results in greater sparseness. Table 3 summarizes the results.

Examination of the outliers in the plots suggests that the improvements obtained with $p = 1.25$ result from allowing the outliers to be farther from the predicted value, while fitting better the bulk of the data points. In this case, the bulk of the data have fitted values between -0.5 and 1.5, while a number of outliers are spread to the left. It is visible that, for the same sparsity, the models with $p = 1.25$ allow for the outliers to settle farther from the horizontal axis, while keeping the bulk of the points more tightly bound to the axis.

Table 2
Values for p, q, and λ used in the computations. The values of $p < 1$ led to highly variable and suboptimal results.

p	2.00	1.75	1.50	1.25	1.00	0.75	0.50	
q	2.00	1.75	1.50	1.25	1.00	0.75	0.50	
λ	0.80	0.69	0.59	0.48	0.37	0.26	0.16	0.05

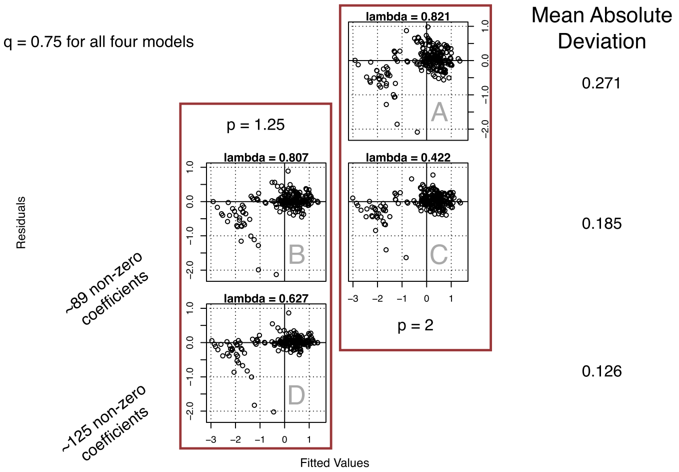

Fig. 6. Plots of residuals vs. fitted values, for four models (these are the same models in Table 3). All models were fitted with $q = 0.75$; models A and C have $p = 2$, while B and D use $p = 1.25$. The values of λ were chosen so that B and C have similar residual ℓ_1 norm (expressed here as Mean Absolute Deviation, MAD), A has the same level of sparsity as B (about 89 non-zero coefficients), and D the same level of sparsity as C (about 125 non-zero coefficients). Notice that for the same level of sparsity, $p = 1.25$ results in reduced MAD; similarly, for the same MAD, $p = 1.25$ results in greater sparseness. Examination of the outliers in the plots suggests that the improvements obtained with $p = 1.25$ result from allowing the outliers to be farther from the predicted value, while fitting better the bulk of the data points.

Table 3
Models in Fig. 6. All models were fitted with $q = 0.75$. Models A and C have $p = 2$, while B and D use $p = 1.25$. The values of λ were chosen so that B and C have similar residual ℓ_1-norm (expressed here as Mean Absolute Deviation, MAD), A has the same level of sparsity as B (about 89 non-zero coefficients), and D the same level of sparsity as C (about 125 non-zero coefficients). Notice that for the same level of sparsity, $p = 1.25$ results in reduced MAD; similarly, for the same MAD, $p = 1.25$ results in greater sparseness.

	p	q	λ	MAD	$\hat{\beta}$ non-zeros
Model A	2.00	0.75	0.82	0.271	88
Model B	1.25	0.75	0.81	0.185	89
Model C	2.00	0.75	0.42	0.185	122
Model D	1.25	0.75	0.63	0.126	127

6. Discussion

We considered the use of loss functions based on ℓ_p-(quasi)norms, in which, for fixed $p, q > 0$, we define the regression optimization problem by setting the fitness term to be the ℓ_p-(quasi)norm of the residual vector, and the penalty term to be the ℓ_q-(quasi)norm of the coefficient vector, possibly transformed by a linear transformation L, and scaled by a regularization parameter $\lambda > 0$.

The use of loss functions for the fitness term that penalize outliers to a lower degree than ℓ_2 has been studied for a long time in the robustness literature [14]. The use of ℓ_1 for the penalty term has become a standard tool in Statistics [28], while the use of non-convex penalties has been studied much less [20]. In this article we combined both approaches, together with a numerical algorithm for fitting the models.

Values of $q < 1$ lead to more sparsity in the vector $\hat{\boldsymbol{\beta}}$ of estimated coefficients, with similar residual ℓ_1-loss, compared with $q \geq 1$. Correspondingly, for similar levels of sparsity, $q < 1$ results in less shrinkage of the nonzero coefficients, and a correspondingly reduced residual ℓ_1-loss.

While we have found that $q < 1$ is useful and produces stable results, using $p < 1$ often results in unstable and/or suboptimal fits, likely due to local minima in the optimization landscape. One possible explanation is that using $p < 1$ encourages sparseness of the residuals, which is not a worthwhile goal since it means that the regression surface will interpolate a few of the data points, while paying little attention to the rest of the points; in this situation, it is not surprising to have many suboptimal local minima. The discussion in Section 2.5 suggests that choosing $p < q$ is not useful.

However, choosing $1 \leq p < 2$ proved helpful in reducing the effect of outliers, improving fit for the bulk of the data, and providing a better trade-off between sparseness and fit than when $p = 2$.

Part of the power of ℓ_p-ℓ_q minimization resides in its ability to deal with large problems by reducing them to smaller problems through the use of generalized Krylov subspaces. In practice, this means that one can fit regression models with many covariates (the example in Section 5 has over 22 thousand variables). This approach is useful for variable selection when the number of variables is large, and the data is contaminated by outliers.

The introduction of constraints in the problem (see, e.g., [3]) will be subject of future research.

Acknowledgement

The authors would like to thank the referee for valuable comments. The work of O.D.C. and L.R. is supported in part by NSF grants DMS-1720259 and DMS-1729509. A.B. and M.D. are members of INdAM-GNCS Gruppo Nazionale per il Calcolo Scientifico. A.B. work is partially founded by the Young Researcher Project “Reconstruction of sparse data” of the GNCS group of INdAM and by the Regione Autonoma della Sardegna research project “Algorithms and Models for Imaging Science [AMIS]” (RASSR57257, intervento finanziato con risorse FSC 2014-2020 - Patto per lo Sviluppo della Regione Sardegna).

References

[1] A. Azzalini, A. Capitanio, The Skew-Normal and Related Families, Cambridge University Press, Cambridge, 2014.
[2] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (2009) 183–202.
[3] A. Buccini, M. Pasha, L. Reichel, Modulus-based iterative methods for constrained ℓ_p-ℓ_q minimization, Inverse Probl. (2020), https://doi.org/10.1088/1361-6420/ab9f86, in press.
[4] A. Buccini, L. Reichel, An ℓ^2-ℓ^q regularization method for large discrete ill-posed problems, J. Sci. Comput. 78 (2019) 1526–1549.
[5] A. Buccini, L. Reichel, An ℓ_p-ℓ_q minimization method with cross-validation for the restoration of impulse noise contaminated images, J. Comput. Appl. Math. 375 (2020) 112824.
[6] E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principle: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inf. Theory 52 (2006) 489–509.
[7] E.J. Candès, M.B. Wakin, S.P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, J. Fourier Anal. Appl. 14 (2008) 877–905.
[8] R.H. Chan, H.X. Liang, Half-quadratic algorithm for ℓ_p-ℓ_q problems with applications to TV-ℓ_1 image restoration and compressive sensing, in: Efficient Algorithms for Global Optimization Methods in Computer Vision, in: Lecture Notes in Computer Science, vol. 8293, Springer, Berlin, 2014, pp. 78–103.
[9] D. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (2006) 1289–1306.
[10] H.W. Engl, M. Hanke, A. Neubauer, Regularization of Inverse Problems, Kluwer, Dordrecht, 1996.
[11] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.
[12] R. Horst, N.V. Thoai, DC programming: overview, J. Optim. Theory Appl. 103 (1999) 1–43.
[13] G. Huang, A. Lanza, S. Morigi, L. Reichel, F. Sgallari, Majorization-minimization generalized Krylov subspace methods for ℓ_p-ℓ_q optimization applied to image restoration, BIT Numer. Math. 57 (2017) 351–378.
[14] P.J. Huber, E.M. Ronchetti, Robust Statistics, 2nd ed., Wiley, Hoboken, 2009.
[15] A. Ghazalpour, B.J. Bennett, D. Shih, et al., Genetic regulation of mouse liver metabolite levels, Mol. Syst. Biol. 10 (5) (2014) 730.
[16] J. Lampe, L. Reichel, H. Voss, Large-scale Tikhonov regularization via reduction by orthogonal projection, Linear Algebra Appl. 436 (2012) 2845–2865.
[17] A. Lanza, S. Morigi, L. Reichel, F. Sgallari, A generalized Krylov subspace method for ℓ_p-ℓ_q minimization, SIAM J. Sci. Comput. 37 (2015) S30–S50.
[18] A. Lanza, S. Morigi, F. Sgallari, Constrained TV$_p$-ℓ_2 model for image restoration, J. Sci. Comput. 68 (2016) 64–91.
[19] Z. Liu, Z. Wei, W. Sun, An iteratively approximated gradient projection algorithm for sparse signal reconstruction, Appl. Math. Comput. 228 (2014) 454–462.
[20] R. Mazumder, J.H. Friedman, T. Hastie, Sparsenet: coordinate descent with nonconvex penalties, J. Am. Stat. Assoc. 106 (2011) 1125–1138.
[21] M. Nikolova, R.H. Chan, The equivalence of the half-quadratic minimization and the gradient linearization iteration, IEEE Trans. Image Process. 16 (2007) 5–18.
[22] R. Ramlau, C.A. Zarzer, On the minimization of a Tikhonov functional with a non-convex sparsity constraint, Electron. Trans. Numer. Anal. 39 (2012) 476–507.
[23] P. Rodríguez, B. Wohlberg, Efficient minimization method for a generalized total variation functional, IEEE Trans. Image Process. 18 (2009) 322–332.
[24] P. Rodríguez, B. Wohlberg, Numerical methods for inverse problems and adaptive decomposition (NUMIPAD), software library available from http://numipad.sourceforge.net/.
[25] L. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D 60 (1992) 259–268.
[26] M. Stone, Cross-validatory choice and assessment of statistical prediction, J. R. Stat. Soc. Ser. B 36 (1977) 111–147.
[27] M.Th. Subbotin, On the law of frequency of error, Mat. Sb. 31 (1923) 296–301.
[28] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[29] R. Tibshirani, T. Hastie, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, New York, 2009.
[30] S. Weisberg, Applied Linear Regression, 4th ed., Wiley, Hoboken, 2014.
[31] R. Wolke, H. Schwetlick, Iteratively reweighted least squares: algorithms, convergence analysis, and numerical comparisons, SIAM J. Sci. Stat. Comput. 9 (1988) 907–921.
[32] A.L. Yuille, A. Rangarajan, The convex-concave procedure, Neural Comput. 15 (2003) 915–936.
[33] Y. Zhao, D. Li, Reweighted ℓ_1-minimization for sparse solutions to undetermined linear systems, SIAM J. Optim. 22 (2012) 1065–1088.