manuscript No.
(will be inserted by the editor)

A comparison of parameter choice rules for ℓ^p-ℓ^q minimization

Alessandro Buccini · Monica Pragliola · Lothar Reichel · Fiorella Sgallari

Received: date / Accepted: date

Abstract Images that have been contaminated by various kinds of blur and noise can be restored by the minimization of an ℓ^p-ℓ^q functional. The quality of the reconstruction depends on the choice of a regularization parameter. Several approaches to determine this parameter have been described in the literature. This work presents a numerical comparison of known approaches as well as of a new one.

Keywords ℓ^p-ℓ^q minimization, inverse problems, parameter choice, regression, iterative method

Mathematics Subject Classification (2010) 65F10, 65R32, 90C26

A. Buccini
Department of Mathematics and Computer Science
University of Cagliari
09124 Cagliari, Italy
E-mail: alessandro.buccini@unica.it

M. Pragliola
Department of Mathematics and Applications
University of Naples Federico II
80126 Naples, Italy
E-mail: monica.pragliola@unina.it

L. Reichel
Department of Mathematical Sciences
Kent State University
Kent, OH 44242, USA
E-mail: reichel@math.kent.edu

F. Sgallari
Department of Mathematics
University of Bologna
40126 Bologna, Italy
E-mail: fiorella.sgallari@unibo.it

1 Introduction

In many areas of Science and Engineering, ranging from medical diagnostics to natural sciences [17, 29, 34], we are faced with the solution of linear systems of equations of the form

$$A\mathbf{x} + \boldsymbol{\eta} \approx \mathbf{b}^\delta, \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$ is a large matrix whose singular values decrease to zero with increasing index number without a significant gap. The matrix A then is severely ill-conditioned and may be rank-deficient. In many applications, the right-hand side $\mathbf{b}^\delta \in \mathbb{R}^m$ represents available data that is contaminated by error. The quantity $\boldsymbol{\eta} \in \mathbb{R}^m$ collects measurement and discretization errors; it is not explicitly known. The vector $\mathbf{x} \in \mathbb{R}^n$ represents the signal that we would like to determine. Problems of this kind often are referred to as *linear discrete ill-posed problems*; see, e.g., [19]. We remark that, in imaging applications, as the ones considered here, the unknown $\mathbf{x}$ as well as the observed data $\mathbf{b}^\delta$ and the additive noise $\boldsymbol{\eta}$ are vectorized forms of $n_1 \times n_2$ and $m_1 \times m_2$ 2D signals, respectively, with $n = n_1n_2$ and $m = m_1m_2$.

Let $\mathbf{b} = \mathbf{b}^\delta - \boldsymbol{\eta}$ denote the unknown error-free vector associated with $\mathbf{b}^\delta$. We are interested in determining the solution $\mathbf{x}^\dagger$ of the least squares problem $\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x}-\mathbf{b}\|_2$ of minimal Euclidean norm. This solution can be expressed as $\mathbf{x}^\dagger = A^\dagger\mathbf{b}$, where $A^\dagger$ denotes the Moore-Penrose pseudo-inverse of A. Since the vector $\mathbf{b}$ is not available, it is natural to try to determine the vector $\mathbf{x} = A^\dagger\mathbf{b}^\delta$. However, due to the ill-conditioning of A and the error $\boldsymbol{\eta}$ in $\mathbf{b}^\delta$, the latter vector often is a meaningless approximation of the desired vector $\mathbf{x}^\dagger$.

To compute a meaningful approximation of $\mathbf{x}^\dagger$ one may resort to regularization methods. These methods replace the ill-posed problem (1) by a nearby well-posed one that is less sensitive to the perturbation in $\mathbf{b}^\delta$ and whose solution is an accurate approximation of $\mathbf{x}^\dagger$. Among the various regularization methods described in the literature, the ℓ^p-ℓ^q minimization method has attracted considerable attention in recent years; see, e.g., [4, 12, 21, 23]. This method computes a regularized solution of (1) by solving the minimization problem

$$\arg\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{p}\left\|A\mathbf{x} - \mathbf{b}^\delta\right\|_p^p + \frac{\mu}{q}\left\|L\mathbf{x}\right\|_q^q, \tag{2}$$

where $0 < p, q \leq 2$, $\mu > 0$ is a regularization parameter, and the matrix $L \in \mathbb{R}^{s\times n}$ is chosen so that $\mathcal{N}(A) \cap \mathcal{N}(L) = \{\mathbf{0}\}$; here $\mathcal{N}(M)$ denotes the null space of the matrix M and

$$\|\mathbf{z}\|_s = \left(\sum_{j=1}^{n} |z_j|^s\right)^{1/s}, \quad \mathbf{z} = [z_1, z_2, \ldots, z_n]^T \in \mathbb{R}^n, \quad s > 0.$$

Note that $\mathbf{z} \to \|\mathbf{z}\|_s$ is not a norm for $0 < s < 1$, but for convenience, we nevertheless will refer to this function as a norm also for these values of s.

The first term in (2) ensures that the reconstructed signal fits the measured data, while the second term enforces some a priori information on the reconstruction. The parameter μ balances the two terms and determines the sensitivity of the solution of (2) to the noise in $\mathbf{b}^\delta$. An ill-suited choice of μ leads to a solution of (2) that is a poor approximation of $\mathbf{x}^\dagger$. It therefore is of vital importance to determining a suitable value of μ.

It is the purpose of this paper to review and compare a few popular parameter choice rules that already have been proposed for ℓ^p-ℓ^q minimization, and to apply, for the first time, a whiteness-based criterion described by Lanza et al. [27].

We briefly comment on the choices of p and q, for which several automatic selection strategies have been developed; see, e.g., [28] and references therein. However, since the focus of this work is the selection of the regularization parameter μ, the parameters p and q are considered to be fixed *a priori*; we assume that they have been set to suitable values.

The choice of p should be informed by the statistical properties of the noise. We consider two types of noise, namely noise that can be described by the Generalized Error Distribution (GED) and impulse noise. In the first case the entries of $\boldsymbol{\eta}$ are independent realization of a random variable with density function

$$\xi_{\theta,\nu\,\sigma}(t) = c_{\theta,\sigma}\text{exp}\left(-\frac{|t-\nu|^\theta}{\theta\sigma^\theta}\right), \tag{3}$$

where $c_{\theta,\sigma}$ is a constant such that $\int_{\mathbb{R}} \xi_{\theta,\nu\,\sigma}(t)dt = 1$, $\nu \in \mathbb{R}$, and $\theta, \sigma > 0$. For $\theta = 2$, (3) reduces to the Gaussian density function, while for $\theta = 1$ we obtain the Laplace distribution. It is shown in [6] that, for this kind of noise, the maximum a posteriori principle prescribes that $p = \theta$.

The data $\mathbf{b}^\delta$ are said to be affected by impulse noise of level σ if

$$\mathbf{b}_j^\delta = \begin{cases} r_j & \text{with probability } \sigma, \\ \mathbf{b}_j & \text{with probability } 1-\sigma, \end{cases}$$

where r_j is a uniformly distributed random variable in the dynamic range of $\mathbf{b}$. Numerical experience indicates that it is beneficial to let $p < 1$ when $\mathbf{b}^\delta$ is contaminated by impulse noise; see, e.g., [6, 21, 23] for illustrations.

The choice of the parameter q is determined by a priori knowledge of the desired solution that we would like to impose on the computed solution. In particular, it is often known that $L\mathbf{x}$ is sparse, i.e., $L\mathbf{x}$ has few nonvanishing entries. This is true, for instance, if L is a discretization of a differential operator or a framelet/wavelet operator. In this case, one ideally would want to let $q = 0$, where the ℓ^0-norm of a vector $\mathbf{x}$ measures the number of nonzero entries.

However, minimizing the ℓ^0-norm is an NP-hard problem. Therefore, it is prudent to approximate the ℓ^0-norm by an ℓ^q-norm with $0 < q \leq 1$. For smaller values of $q > 0$, the ℓ^q-norm approximates the ℓ^0-norm better, however, the minimizing algorithm requires more iterations the smaller $q > 0$ is, and may suffer from numerical instability for "tiny" $q > 0$. Therefore, it is usually a good practice to set q small enough, but not too small. We remark that the ℓ^q-norm does not satisfy all properties of a norm for $0 < q < 1$.

This paper is organized as follows: Section 2 outlines two algorithms for computing an approximate solution of (2). In Section 3 we describe the methods for determining μ that are compared in this paper. We report some numerical results in Section 4 and draw conclusions in Section 5.

2 Majorization-minimization in generalized Krylov subspaces

In [21] the authors proposed an effective iterative method for the solution of (2). At each iteration, a smoothed version of the ℓ^p-ℓ^q functional, denoted by $\mathcal{J}_\varepsilon$, is majorized by a quadratic functional that is tangent to $\mathcal{J}_\varepsilon$ at the current iterate $\mathbf{x}^{(k)}$. Then, the quadratic tangent majorant is minimized and the minimizer $\mathbf{x}^{(k+1)}$ is the new iterate; see below. Two approaches to determine quadratic majorants are described in [21]. We will outline both.

Majorization step. Consider the functional

$$\mathcal{J}(\mathbf{x}) = \frac{1}{p}\left\|A\mathbf{x}-\mathbf{b}^\delta\right\|_p^p + \frac{\mu}{q}\left\|L\mathbf{x}\right\|_q^q, \tag{4}$$

that is minimized in (2). It is shown in [8] that this functional has a global minimizer. When $0 < \min\{p,q\} < 1$, the functional (4) is neither convex nor differentiable. To construct a quadratic majorant, the functional has to be continuously differentiable. We, therefore, introduce a smoothed version

$$\mathcal{J}_\varepsilon(\mathbf{x}) = \frac{1}{p}\sum_{j=1}^{m}\Phi_{p,\varepsilon}\left((A\mathbf{x}-\mathbf{b}^\delta)_j\right) + \frac{\mu}{q}\sum_{j=1}^{s}\Phi_{q,\varepsilon}\left((L\mathbf{x})_j\right)$$

for some $\varepsilon > 0$, where

$$\Phi_{s,\varepsilon}(t) = \begin{cases} |t|^s & \text{for } s > 1, \\ \left(t^2+\varepsilon^2\right)^{s/2} & \text{for } 0 < s \leq 1. \end{cases}$$

Since $\Phi_{s,\varepsilon}$ is a differentiable function of t, $\mathcal{J}_\varepsilon(\mathbf{x})$ is everywhere differentiable. We will comment on the choice of ε in Section 4.

We would like to compute

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \mathcal{J}_\varepsilon(\mathbf{x}). \tag{5}$$

When $\min\{p,q\} > 1$, the functional $\mathcal{J}_\varepsilon(\mathbf{x})$ is strictly convex and therefore the minimization problem (5) has a unique minimum. On the other hand, if $\min\{p,q\} < 1$, which is the situation of interest to us, then the functional (5) generally is not convex. The methods we describe here therefore determine a stationary point.

Let $\mathbf{x}^{(k)}$ be the currently available approximation of $\mathbf{x}^*$ in (5). We first describe the majorant referred to as "adaptive" in [21]. This majorant is such that the quadratic approximation of each component of $\mathcal{J}_\varepsilon(\mathbf{x})$ at $\mathbf{x}^{(k)}$ has an as large as possible positive second order derivative. In general, each component is approximated by a different quadratic polynomial. We denote the adaptive quadratic tangent majorant of $\mathcal{J}_\varepsilon(\mathbf{x})$ at $\mathbf{x}^{(k)}$ by $\mathcal{Q}^A(\mathbf{x},\mathbf{x}^{(k)})$. It is characterized by

$$\begin{aligned}\mathcal{Q}^A(\mathbf{x}^{(k)},\mathbf{x}^{(k)}) &= \mathcal{J}_\varepsilon(\mathbf{x}^{(k)}),\\ \nabla_{\mathbf{x}}\mathcal{Q}^A(\mathbf{x}^{(k)},\mathbf{x}^{(k)}) &= \nabla_{\mathbf{x}}\mathcal{J}_\varepsilon(\mathbf{x}^{(k)}),\\ \mathcal{Q}^A(\mathbf{x},\mathbf{x}^{(k)}) &\geq \mathcal{J}_\varepsilon(\mathbf{x})\ \forall\mathbf{x}\in\mathbb{R}^n,\\ &\mathbf{x}\to\mathcal{Q}^A(\mathbf{x},\mathbf{x}^{(k)})\quad\text{is a quadratic functional,}\end{aligned}\tag{6}$$

where $\nabla_{\mathbf{x}}$ denotes the gradient with respect to $\mathbf{x}$.

The functional $\mathcal{Q}^A(\mathbf{x},\mathbf{x}^{(k)})$ can be constructed as follows: Evaluate the residual vectors

$$\mathbf{v}^{(k)} = A\mathbf{x}^{(k)} - \mathbf{b}^\delta,\quad \mathbf{u}^{(k)} = L\mathbf{x}^{(k)}$$

and compute the *weight vectors*

$$\boldsymbol{\omega}_{\text{fid}}^{A,(k)} = \left(\left(\mathbf{v}^{(k)}\right)^2+\varepsilon^2\right)^{p/2-1},\quad \boldsymbol{\omega}_{\text{reg}}^{A,(k)} = \left(\left(\mathbf{u}^{(k)}\right)^2+\varepsilon^2\right)^{q/2-1},$$

where all the operations are meant element-wise. The weight vectors determine the diagonal matrices

$$W_{\text{fid}}^{(k)} = \operatorname{diag}\left(\boldsymbol{\omega}_{\text{fid}}^{A,(k)}\right)\quad\text{and}\quad W_{\text{reg}}^{(k)} = \operatorname{diag}\left(\boldsymbol{\omega}_{\text{reg}}^{A,(k)}\right).$$

Then

$$\mathcal{Q}^A(\mathbf{x},\mathbf{x}^{(k)}) = \frac{1}{2}\left\|\left(W_{\text{fid}}^{(k)}\right)^{1/2}\left(A\mathbf{x}-\mathbf{b}^\delta\right)\right\|_2^2 + \frac{\mu}{2}\left\|\left(W_{\text{reg}}^{(k)}\right)^{1/2}L\mathbf{x}\right\|_2^2 + c,$$

where $c\in\mathbb{R}$ is a constant that is independent of $\mathbf{x}$; see [21] for details. The new approximation of $\mathbf{x}^*$ is given by the minimizer $\mathbf{x}^{(k+1)}$ of $\mathcal{Q}^A(\mathbf{x},\mathbf{x}^{(k)})$. We discuss the computation of an approximation or this minimizer below.

The second type of majorant considered in [21] is referred to as “fixed.” This majorant is constructed so that each component of the quadratic polynomial majorants have the same leading coefficient. As shown below, the evaluation of the fixed majorant at $\mathbf{x}$ requires less computational work than the computation of the adaptive majorant. However, in general, the method defined by fixed majorants requires more iterations than the method defined by adaptive majorants to reach numerical convergence.

For the fixed case the *weight vectors* are given by

$$\begin{aligned}\boldsymbol{\omega}_{\text{fid}}^{F,(k)} &= \mathbf{v}^{(k)}\left(1-\left(\frac{\left(\mathbf{v}^{(k)}\right)^2+\varepsilon^2}{\varepsilon^2}\right)^{p/2-1}\right),\\ \boldsymbol{\omega}_{\text{reg}}^{F,(k)} &= \mathbf{u}^{(k)}\left(1-\left(\frac{\left(\mathbf{u}^{(k)}\right)^2+\varepsilon^2}{\varepsilon^2}\right)^{q/2-1}\right),\end{aligned}\tag{7}$$

and determine the fixed quadratic tangent majorant

$$\begin{aligned}\mathcal{Q}^F(\mathbf{x},\mathbf{x}^{(k)}) = \tfrac{1}{2}\left(\left\|A\mathbf{x}-\mathbf{b}^\delta\right\|_2^2 - 2\left\langle \boldsymbol{\omega}_{\text{fid}}^{F,(k)}, A\mathbf{x}\right\rangle\right)\\ + \tfrac{\mu}{2}\varepsilon^{q-p}\left(\left\|L\mathbf{x}\right\|_2^2 - 2\left\langle \boldsymbol{\omega}_{\text{reg}}^{F,(k)}, L\mathbf{x}\right\rangle\right) + c,\end{aligned}$$

of $\mathcal{J}_\varepsilon$ at $\mathbf{x}^{(k)}$. Here $\langle\cdot,\cdot\rangle$ denotes the standard inner product and the constant $c \in \mathbb{R}$ is independent of $\mathbf{x}$. The functional $\mathcal{Q}^F(\mathbf{x},\mathbf{x}^{(k)})$ satisfies the properties (6) with $\mathcal{Q}^A(\mathbf{x},\mathbf{x}^{(k)})$ replaced by $\mathcal{Q}^F(\mathbf{x},\mathbf{x}^{(k)})$; see [21] for details.

Minimization step. We describe how to compute minimizers of $\mathcal{Q}^A$ and $\mathcal{Q}^F$ when the matrix $A \in \mathbb{R}^{m\times n}$ is large. To reduce the computational effort, we compute an approximate solution in a subspace $\mathcal{V}_k$ of fairly small dimension $\widehat{k} \ll \min\{m,n\}$. Let the columns of the matrix $V_k \in \mathbb{R}^{n\times\widehat{k}}$ form an orthonormal basis for $\mathcal{V}_k$. We determine approximations of the minima of $\mathcal{Q}^A$ and $\mathcal{Q}^F$ of the form

$$\mathbf{x}^{(k+1)} = V_k\mathbf{y}^{(k+1)}, \tag{8}$$

where $\mathbf{y}^{(k+1)} \in \mathbb{R}^{\widehat{k}}$.

We first discuss the adaptive case. We would like to solve

$$\min_{\mathbf{x}\in\mathcal{V}_k} \frac{1}{2}\left\|\left(W_{\text{fid}}^{(k)}\right)^{1/2}(A\mathbf{x}-\mathbf{b}^\delta)\right\|_2^2 + \frac{\mu}{2}\left\|\left(W_{\text{reg}}^{(k)}\right)^{1/2}L\mathbf{x}\right\|_2^2, \tag{9}$$

and denote the solution by $\mathbf{x}^{(k+1)}$. This minimization problem is equivalent to

$$\min_{\mathbf{y}\in\mathbb{R}^{\widehat{k}}} \frac{1}{2}\left\|\left(W_{\text{fid}}^{(k)}\right)^{1/2}(AV_k\mathbf{y}-\mathbf{b}^\delta)\right\|_2^2 + \frac{\mu}{2}\left\|\left(W_{\text{reg}}^{(k)}\right)^{1/2}LV_k\mathbf{y}\right\|_2^2. \tag{10}$$

Introduce the economic QR factorizations

$$\begin{aligned}\left(W_{\text{fid}}^{(k)}\right)^{1/2}AV_k &= Q_AR_A, \quad Q_A \in \mathbb{R}^{m\times\widehat{k}},\ R_A \in \mathbb{R}^{\widehat{k}\times\widehat{k}},\\ \left(W_{\text{reg}}^{(k)}\right)^{1/2}LV_k &= Q_LR_L, \quad Q_L \in \mathbb{R}^{s\times\widehat{k}},\ R_L \in \mathbb{R}^{\widehat{k}\times\widehat{k}},\end{aligned} \tag{11}$$

and compute

$$\mathbf{y}^{(k+1)} = \arg\min_{\mathbf{y}\in\mathbb{R}^{\widehat{k}}} \frac{1}{2}\left\|R_A\mathbf{y} - Q_A^T\left(W_{\text{fid}}^{(k)}\right)^{1/2}\mathbf{b}^\delta\right\|_2^2 + \frac{\mu}{2}\left\|R_L\mathbf{y}\right\|_2^2, \tag{12}$$

where the superscript ^{T} denotes transposition. Assume that

$$\mathcal{N}\left(\left(W_{\text{fid}}^{(k)}\right)^{1/2}AV_k\right) \cap \mathcal{N}\left(\left(W_{\text{reg}}^{(k)}\right)^{1/2}LV_k\right) = \{\mathbf{0}\}.$$

This condition typically holds in applications of interest to us. The solution $\mathbf{y}^{(k+1)}$ of (12) then is unique, and the approximate minimizer of $\mathcal{Q}^A(\mathbf{x},\mathbf{x}^{(k)})$ is given by (8).

We now enlarge the solution subspace $\mathcal{V}_k$ by including the normalized residual of the normal equations associated with (9). The residual is given by

$$\mathbf{r}^{(k+1)} = A^T W_{\text{fid}}^{(k)} \left(A\mathbf{x}^{(k+1)} - \mathbf{b}^\delta \right) + \mu L^T W_{\text{reg}}^{(k)} L\mathbf{x}^{(k+1)},$$

and the columns of the matrix

$$V_{k+1} = \left[V_k, \mathbf{r}^{(k+1)}/\|\mathbf{r}^{(k+1)}\|_2 \right]$$

form an orthonormal basis for the new solution subspace $\mathcal{V}_{k+1}$. We would like to point out that the vector $\mathbf{r}^{(k+1)}$ is proportional to the gradient of $\mathcal{Q}^A(\mathbf{x}, \mathbf{x}^{(k)})$ restricted to $\mathcal{V}_k$ at $\mathbf{x} = \mathbf{x}^{(k+1)}$. We refer to the solution subspace $\mathcal{V}_{k+1} = \text{range}(V_{k+1})$ as a *generalized Krylov subspace*. Note that the computation of $\mathbf{r}^{(k+1)}$ requires only one matrix-vector product with A^T and L^T, since we can exploit the QR factorizations (11) and the relation (8) to avoid computing any other matrix-vector products with the matrices A and L. Moreover, we store and update the “skinny” matrices AV_k and LV_k at each iteration to reduce the computational cost. The initial space $\mathcal{V}_1$ is usually chosen to contain a few selected vectors and to be of small dimension. A common choice is $\mathcal{V}_1 = \text{span}\{A^T\mathbf{b}^\delta\}$, which implies that $\widehat{k} = k$. We will use this choice in the computed examples reported in Section 4.

Summarizing, each iteration of the adaptive approach requires one matrix-vector product evaluation with each one of the matrices A, L, A^T, and L^T, as well as the computation of economic QR factorizations of two tall and skinny matrices, whose column numbers increase by one with each iteration. The latter computations can be quite demanding if the matrices A and L are large and many iterations are required. The algorithm requires the storage of the three matrices V_k, AV_k, and LV_k. In addition, storage of some representations of the matrices A and L is needed.

We turn to the fixed approach. The weight vectors are now given by (7), and we would like to solve the minimization problem

$$\min_{\mathbf{x}\in\mathcal{V}_k} \frac{1}{2} \left(\|A\mathbf{x} - \mathbf{b}^\delta\|_2^2 - 2\left\langle \boldsymbol{\omega}_{\text{fid}}^{F,(k)}, A\mathbf{x} \right\rangle \right) + \frac{\eta}{2} \left(\|L\mathbf{x}\|_2^2 - 2\left\langle \boldsymbol{\omega}_{\text{reg}}^{F,(k)}, L\mathbf{x} \right\rangle \right) \tag{13}$$

for $\mathbf{x}^{(k+1)}$, where $\eta = \mu\varepsilon^{q-p}$. This problem can be expressed as

$$\min_{\mathbf{y}\in\mathbb{R}^{\widehat{k}}} \left\| AV_k\mathbf{y} - \mathbf{b}^\delta - \boldsymbol{\omega}_{\text{fid}}^{F,(k)} \right\|_2^2 + \eta \left\| LV_k\mathbf{y} - \boldsymbol{\omega}_{\text{reg}}^{F,(k)} \right\|_2^2. \tag{14}$$

The solution $\mathbf{y}^{(k+1)}$ of (14) yields the solution $\mathbf{x}^{(k+1)} = V_k\mathbf{y}^{(k+1)}$ of (13).

Introduce the economic QR factorizations

$$\begin{aligned} AV_k &= Q_A R_A, \qquad Q_A \in \mathbb{R}^{m\times\widehat{k}},\ R_A \in \mathbb{R}^{\widehat{k}\times\widehat{k}} \\ LV_k &= Q_L R_L, \qquad Q_L \in \mathbb{R}^{s\times\widehat{k}},\ R_L \in \mathbb{R}^{\widehat{k}\times\widehat{k}}. \end{aligned}$$

Substituting these factorizations into (14), we obtain

$$\mathbf{y}^{(k+1)} = \arg\min_{\mathbf{y}\in\mathbb{R}^{\widehat{k}}} \left\| \begin{bmatrix} R_A \\ \sqrt{\eta} R_L \end{bmatrix} \mathbf{y} - \begin{bmatrix} Q_A^T \left(\mathbf{b}^\delta + \boldsymbol{\omega}_{\text{fid}}^{F,(k)}\right) \\ \sqrt{\eta} Q_L^T \boldsymbol{\omega}_{\text{reg}}^{F,(k)} \end{bmatrix} \right\|_2^2 .$$

Once we have computed $\mathbf{y}^{(k+1)}$ and $\mathbf{x}^{(k+1)}$, we enlarge the solution subspace by including the residual

$$\mathbf{r}^{(k+1)} = A^T \left(A\mathbf{x}^{(k+1)} - \left(\mathbf{b}^\delta + \boldsymbol{\omega}_{\text{fid}}^{F,(k)}\right)\right) + \eta L^T \left(L\mathbf{x}^{(k+1)} - \boldsymbol{\omega}_{\text{reg}}^{F,(k)}\right)$$

of the normal equations associated with (13). Thus, let $\mathbf{v}_{\text{new}} = \mathbf{r}^{(k+1)} / \left\|\mathbf{r}^{(k+1)}\right\|_2$. Then the columns of the matrix $V_{k+1} = [V_k, \mathbf{v}_{\text{new}}]$ form an orthonormal basis for the solution subspace $\mathcal{V}_{k+1}$. We remark that the residual is proportional to the gradient of $\mathcal{Q}^F(\mathbf{x}, \mathbf{x}^{(k)})$ restricted to $\mathcal{V}_k$ at $\mathbf{x} = \mathbf{x}^{(k+1)}$.

Note that differently from (10), the least-squares problem (14) does not have a diagonal scaling matrix. We therefore may compute the QR factorizations of AV_{k+1} and LV_{k+1} by updating the QR factorizations of AV_k and LV_k, respectively. This reduces the computational work and leads to that each new iteration with the fixed approach is cheaper than with the adaptive approach. Updating formulas for the QR factorization can be found in [13, 21].

Each iteration with the fixed approach requires one matrix-vector product evaluation with each one of the matrices A, L, A^T, and L^T, similarly as for the adaptive approach. Moreover the memory requirements of the fixed and adaptive approaches are essentially the same.

The memory requirement of both the adaptive and fixed approaches outlined grows linearly with the number of iterations. It follows that when the matrix A is large, the memory requirement may be substantial when many iterations are required to satisfy the stopping criterion. This could be a difficulty on computers with fairly little fast memory. Moreover, the arithmetic cost for computing QR factorizations in the adaptive approach and for updating QR factorizations in the fixed approach grows quadratically and linearly, respectively, with the number of iterations.

3 Parameter choice rules

This section discusses several approaches to determine a suitable value of the regularization parameter μ in (2). The strategies considered in the literature for the choice of the regularization parameter when variational models, including Tikhonov regularization, are employed can be divided into three main classes:

(i) Methods relying on the noise level, that either may be known or accurately estimated. A very popular approach belonging to this class is the *Discrepancy Principle* (DP).

(ii) Methods that rely on different statistical and non-deterministic properties of the noise, such as its *whiteness*.

(iii) Heuristic methods, which are typically only based on the knowledge of the data $\mathbf{b}^\delta$, among which we mention *Generalized Cross Validation* (GCV) and the L-curve criterion.

A general discussion on heuristic methods is provided by Kindermann [22]. Further references are provided below. In what follows, we are going to review some of the approaches mentioned and their modifications when applied to the solution of (2). Specifically, we are considering *stationary* and *non-stationary* scenarios: in the former, the methods are employed *a posteriori*, i.e., the minimization problem in (2) is solved for different μ-values, and the optimal value, μ^*, is selected based on a chosen strategy. In the latter scenario, the chosen strategy is applied during the iterations of the generalized Krylov method; therefore the optimization problem (2) is solved only once.

We remark that the formulation of the stationary strategies discussed in Section 3.1 does not depend on the selected algorithmic scheme. When considering the non-stationary rules presented in Section 3.2, several issues should be considered, such as the existence of a fast parameter update and the convergence of the overall iterative procedure. In this review, we focus on the robustness of the considered approaches when embedded in the iterations of the generalized Krylov method, and leave a rigorous analysis of the mentioned issues to a future study.

3.1 Stationary rules

We describe the stationary rules considered in this paper.

3.1.1 Discrepancy principle

Let the noise that corrupts the data be Gaussian. Then we set $p = 2$. Let $\mathbf{x}_\mu$ denote the solution of (2) with $p = 2$, i.e.,

$$\mathbf{x}_\mu = \arg\min_{\mathbf{x}} \frac{1}{2} \left\| A\mathbf{x} - \mathbf{b}^\delta \right\|_2^2 + \frac{\mu}{q} \left\| L\mathbf{x} \right\|_q^q .$$

Assume that a fairly accurate estimate $\delta > 0$ of the norm of the noise is available

$$\|\boldsymbol{\eta}\|_2 \leq \delta.$$

The discrepancy principle (DP) prescribes that the parameter μ be chosen such that

$$\mu_{\mathrm{DP}} = \sup\left\{\mu : \ \left\| A\mathbf{x}_\mu - \mathbf{b}^\delta \right\|_2 \leq \tau\delta \right\},$$

where $\tau > 1$ is a user-defined constant that is independent of δ.

A non-stationary strategy to estimate μ_{DP} is described in [9], and we will discuss it in Section 3.2.1. Here we propose a stationary way to determine an estimate of the DP parameter. Extensive numerical experience shows that the computation of $\mathbf{x}_\mu$ is fairly stable with respect to the choice of μ and, therefore, a rough estimate of μ_{DP} is usually enough to compute a satisfactory solution.

The analysis of the DP requires that $\mathbf{b} \in \mathcal{R}(A)$; see, e.g., [14]. Then μ_{DP} is well defined, since

$$0 \in \left\{\mu : \left\|A\mathbf{x}_\mu - \mathbf{b}^\delta\right\|_2 \leq \tau\delta\right\}.$$

Define the function $r(\mu) = \left\|A\mathbf{x}_\mu - \mathbf{b}^\delta\right\|_2 - \tau\delta$ and assume that $r(\mu)$ is continuous. This assumption is satisfied for $q = 2$, see, e.g., [14], but to the best of our knowledge a proof for general q is not currently available. We may employ a root-finder to determine μ_{DP}; see, e.g., [7, 31].

3.1.2 Residual whiteness principle

The whiteness property of the corrupting noise in linear inverse problems of the form (1) has been extensively explored in the context of variational methods. It is convenient to apply the whiteness property since it does not require knowledge of the standard deviation of the noise. Moreover, as it will be made clear in the following, it exploits more information of the data vector $\mathbf{b}^\delta$ than the DP.

The whiteness property has been incorporated in variational models for image denoising and deblurring problems in [3,24–26,32]. Despite the high-quality results achieved in these works, these approaches suffer from the strong nonconvexity of the variational models that have to be solved. This makes minimization a very hard task. Other approaches that exploit that the residual image is expected to model white Gaussian noise are described in [18,33] and [1], where the authors propose two statistically-motivated parameter choice procedures based on the maximization of the residual whiteness by the normalized cumulative periodogram and the normalized auto-correlation, respectively. The approach introduced in [1] and applied as an *a posteriori* parameter choice criterion for image deconvolution problems has been revisited in [27,30]. There the authors propose to automatically update the regularization parameter during the iterations of the algorithm used for the minimization of a wide class of convex variational models for image restoration and super-resolution problems.

In what follows, we recall the main steps of the a posteriori criterion described in [1, 27] and referred to as the Residual Whiteness Principle (RWP). We remark that, although the RWP has been originally designed for Gaussian noise corruption, it can been applied whenever the noise $\boldsymbol{\eta}$ in $\mathbf{b}^\delta$ has independent and identically distributed entries.

Consider the noise realization $\boldsymbol{\eta} \in \mathbb{R}^m$ in (1) represented in its original $m_1 \times m_2$ form. Thus,

$$\boldsymbol{\eta} \;=\; \left\{\eta_{i,j}\right\}_{(i,j)\in\Omega}, \quad \Omega := \{0, \ldots, m_1 - 1\} \times \{0, \ldots, m_2 - 1\}. \tag{15}$$

The *sample auto-correlation* of $\boldsymbol{\eta}$ is defined as

$$a(\boldsymbol{\eta}) \;=\; \left\{a_{l,k}(\boldsymbol{\eta})\right\}_{(l,k)\in\Theta},$$

with $\Theta := \{-(m_1-1), \ldots, m_1-1\} \times \{-(m_2-1), \ldots, m_2-1\}$. The scalar components $a_{l,k}(\boldsymbol{\eta})$ are given by

$$\begin{aligned} a_{l,k}(\boldsymbol{\eta}) &= \frac{1}{m}\left(\boldsymbol{\eta} \star \boldsymbol{\eta}\right)_{l,k} = \frac{1}{m}\left(\boldsymbol{\eta} * \boldsymbol{\eta}'\right)_{l,k} \\ &= \frac{1}{m}\sum_{(i,j)\in\Omega} \eta_{i,j}\,\eta_{i+l,j+k}\,, \qquad (l,k)\in\Theta\,, \end{aligned} \tag{16}$$

where the integer pairs (l,k) are referred to as *lags*, $\star$ and $*$ denote the 2D discrete correlation and convolution operators, respectively, and $\boldsymbol{\eta}'(i,j) = \boldsymbol{\eta}(-i,-j)$.

Clearly, for (16) to be defined for all lags $(l,k)\in\Theta$, the noise realization $\boldsymbol{\eta}$ must be padded with at least m_1-1 samples in the vertical direction and m_2-1 samples in the horizontal direction. We will assume periodic boundary conditions for $\boldsymbol{\eta}$, so that $\star$ and $*$ in (16) denote 2D circular correlation and convolution, respectively. Then the auto-correlation has some symmetries that allow us to only consider the lags

$$(l,k)\in\overline{\Theta} := \{0, \ldots, m_1-1\} \times \{0, \ldots, m_2-1\}.$$

If the error $\boldsymbol{\eta}$ in (1) is the realization of a white noise process, then it is well known that the sample auto-correlation $a(\boldsymbol{\eta})$ satisfies the asymptotic property:

$$\lim_{m\to+\infty} a_{l,k}(\boldsymbol{\eta}) = \begin{cases} \sigma^2 & \text{for } (l,k)=(0,0), \\ 0 & \text{for } (l,k)\in\overline{\Theta}_0 := \overline{\Theta}\setminus\{(0,0)\}\,. \end{cases} \tag{17}$$

We note that the discrepancy principle relies on exploiting only the lag $(0,0)$ – among the m asymptotic properties of the noise auto-correlation given in (17). Imposing whiteness of the residual image of the restoration by constraining the residual auto-correlation at non-zero lags to be small is a much stronger requirement.

The whiteness principle can be made independent of the noise level by considering the *normalized* sample auto-correlation of the noise realization $\boldsymbol{\eta}$ in (15), namely

$$\beta(\boldsymbol{\eta}) = \frac{1}{a_{0,0}(\boldsymbol{\eta})}\,a(\boldsymbol{\eta}) = \frac{1}{\|\boldsymbol{\eta}\|_F^2}\left(\boldsymbol{\eta}\star\boldsymbol{\eta}\right),$$

where $\|\boldsymbol{\eta}\|_F$ denotes the Frobenius norm of the matrix $\boldsymbol{\eta}$. It follows easily from (17) that

$$\lim_{m\to+\infty} \beta_{l,k}(\boldsymbol{\eta}) = \begin{cases} 1 & \text{for } (l,k)=(0,0), \\ 0 & \text{for } (l,k)\in\overline{\Theta}_0\,. \end{cases}$$

We introduce the following σ-independent non-negative scalar measure of whiteness $\mathcal{W}:\mathbb{R}^{m_1\times m_2}\to\mathbb{R}^+$ of the noise realization $\boldsymbol{\eta}$:

$$\mathcal{W}(\boldsymbol{\eta}) := \|\beta(\boldsymbol{\eta})\|_F^2 = \frac{\|\,\boldsymbol{\eta}\star\boldsymbol{\eta}\,\|_F^2}{\|\boldsymbol{\eta}\|_F^4}\,. \tag{18}$$

Clearly, the nearer the restored image $\mathbf{x}_\mu$ in (2) is to the target uncorrupted image $\mathbf{x}^\dagger$, the closer the associated $m_1 \times m_2$ residual image obtained from $\boldsymbol{d}_\mu = A\mathbf{x}_\mu - \mathbf{b}^\delta$ is to the white noise realization $\boldsymbol{\eta}$ in (1) and, hence, the whiter is the residual image according to the scalar measure in (18). The RWP for automatically selecting the regularization parameter μ in variational models of the general form (2) therefore can be formulated as

$$\mu^* \in \arg\min_{\mu>0} W(\mu), \quad W(\mu) := \mathcal{W}\left(\boldsymbol{d}_\mu\right), \quad \boldsymbol{d}_\mu = A\mathbf{x}_\mu - \mathbf{b}^\delta\,,$$

where the scalar cost function W is defined by

$$W(\mu) \;=\; \|\rho(\boldsymbol{d}_\mu)\|_2^2 \;=\; \frac{\|\,\boldsymbol{d}_\mu \,\star\, \boldsymbol{d}_\mu\,\|_2^2}{\|\boldsymbol{d}_\mu\|_2^4}\,. \tag{19}$$

We refer to W as the *residual whiteness function.*

3.1.3 Cross Validation

Two heuristic approaches to determine a suitable value of the regularization parameter μ in (2) for any positive values of p and q based on cross validation (CV) are described in [10]. This and the following subsections reviews these methods. For details on CV, we refer to [35].

Let $1 \leq d \ll m$ and choose d distinct random integers $i_1, \ldots, i_d$ in $\{1, \ldots, m\}$. Remove rows $i_1, \ldots, i_d$ from A and $\mathbf{b}^\delta$. This gives the "reduced" matrix and data vector $\widetilde{A} \in \mathbb{R}^{(m-d)\times n}$ and $\widetilde{\mathbf{b}^\delta} \in \mathbb{R}^{m-d}$, respectively. Consider a set $\{\mu_1, \ldots, \mu_l\}$ of positive parameter values. For each μ_j, we solve

$$\mathbf{x}_j = \arg\min_{\mathbf{x}} \frac{1}{p}\left\|\widetilde{A}\mathbf{x} - \widetilde{\mathbf{b}^\delta}\right\|_p^p + \frac{\mu_j}{q}\left\|L\mathbf{x}\right\|_q^q.$$

Our aim is to determine the parameter μ_j that yields a solution that predicts the values $\mathbf{b}^\delta_{i_1}, \ldots, \mathbf{b}^\delta_{i_d}$ well. Therefore, we compute for each j the residual error

$$r_j = \sqrt{\sum_{k=1}^{m}\left(A\mathbf{x}_j - \mathbf{b}^\delta\right)_{i_k}^2}$$

and let

$$j^* = \arg\min_j r_j.$$

Define $\mu^* = \mu_{j^*}$. To reduce variability, this process is repeated K times, producing K (possibly different) values $\mu_1^*, \ldots, \mu_K^*$. The computed approximation of $\mathbf{x}^\dagger$ is obtained as

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \frac{1}{p}\left\|A\mathbf{x} - \mathbf{b}^\delta\right\|_p^p + \frac{\widehat{\mu}}{q}\left\|L\mathbf{x}\right\|_q^q$$

with

$$\widehat{\mu} = \frac{1}{K}\sum_{k=1}^{K}\mu_j^*.$$

3.1.4 Modified Cross Validation

Cross validation can be applied to predict entries of the computed solution instead of entries of the data vector $\mathbf{b}^\delta$. This is described in [10] and there referred to as Modified Cross Validation (MCV). We outline this approach. The idea is to determine a value of μ such that the computed solution is stable with respect to loss of data. Let $1 \leq d \ll m$ and select two sets of d distinct indices between 1 and m referred to as $I_1 = \{i_1^{(1)}, i_2^{(1)}, \dots, i_d^{(1)}\}$ and $I_2 = \{i_1^{(2)}, i_2^{(2)}, \dots, i_d^{(2)}\}$. Let $\widetilde{A}_1$ and $\widetilde{\mathbf{b}^\delta}_1$, and $\widetilde{A}_2$ and $\widetilde{\mathbf{b}^\delta}_2$, denote the matrices and data vectors obtained by removing the rows with indexes I_1 and I_2 from A and $\mathbf{b}^\delta$, respectively. Let $\{\mu_1, \dots, \mu_l\}$ be a set of regularization parameters and solve

$$\mathbf{x}_j^{(i)} = \arg\min_{\mathbf{x}} \frac{1}{p} \left\| \widetilde{A}_i \mathbf{x} - \widetilde{\mathbf{b}}_i^\delta \right\|_p^p + \frac{\mu_j}{q} \|L\mathbf{x}\|_q^q, \quad j = 1, 2, \dots, l, \quad i = 1, 2.$$

We then compute the quantities

$$\mathit{\Delta}_j = \left\| \mathbf{x}_j^{(1)} - \mathbf{x}_j^{(2)} \right\|_2, \quad j = 1, 2, \dots, l,$$

evaluate

$$j^* = \arg\min_j \mathit{\Delta}_j,$$

and define $\mu^* = \mu_{j^*}$. To reduce variability, the computations are repeated K times. This results in K values of μ^*, which we denote by $\mu_1^*, \dots, \mu_K^*$. The computed approximation of $\mathbf{x}^\dagger$ is given by

$$\arg\min_{\mathbf{x}} \frac{1}{p} \left\| A\mathbf{x} - \mathbf{b}^\delta \right\|_p^p + \frac{\widehat{\mu}}{q} \|L\mathbf{x}\|_q^q$$

with

$$\widehat{\mu} = \frac{1}{K} \sum_{k=1}^{K} \mu_j^*.$$

3.2 Non-stationary rules

We now describe a few non-stationary rules for determining the regularization parameter.

3.2.1 Discrepancy principle

An iterative, nonstationary, approach to determine the regularization parameter μ for the minimization problem (2) when $p = 2$ and the error $\boldsymbol{\eta}$ in $\mathbf{b}^\delta$ is white Gaussian is described in [9]. We review this method for the fixed approach described at the end of Section 2. Our discussion easily can be extended to the adaptive case.

Formula (14) for $p = 2$ simplifies to

$$\mathbf{y}^{(k+1)} = \arg\min_{\mathbf{y}\in\mathbb{R}^{\widehat{k}}} \left\|AV_k\mathbf{y} - \mathbf{b}^\delta\right\|_2^2 + \rho \left\|LV_k\mathbf{y} - \boldsymbol{\omega}_{\text{reg}}^{F,(k)}\right\|_2^2.$$

Substituting QR factorizations of AV_k and LV_k into the above equation and allowing ρ to change in each iteration, we obtain

$$\mathbf{y}^{(k+1)} = \arg\min_{\mathbf{y}\in\mathbb{R}^{\widehat{k}}} \left\|R_A\mathbf{y} - Q_A^T\mathbf{b}^\delta\right\|_2^2 + \rho^{(k)} \left\|R_L\mathbf{y} - Q_L^T\boldsymbol{\omega}_{\text{reg}}^{F,(k)}\right\|_2^2. \tag{20}$$

We determine the parameter $\rho^{(k)}$ in each iteration so that

$$\left\|AV_k\mathbf{y}^{(k+1)} - \mathbf{b}^\delta\right\|_2 = \tau\delta. \tag{21}$$

The above is a nonlinear equation in $\rho^{(k)}$. After having computed the GSVD of the matrix pair $\{R_A, R_L\}$, this equation can be solved efficiently using a zero-finder; see [7, 31] for discussions. Here we just mention that the computation of the GSVD, even though it has to be done in each iteration, is not computationally expensive since the sizes of the matrices R_A an R_L are typically fairly small.

Since we compute the regularization parameter in each iteration, the method is non-stationary. We summarize the iterations as follows:

1. Compute the weight $\boldsymbol{\omega}_{\text{reg}}^{F,(k)}$;
2. Solve (21) for $\rho^{(k)}$;
3. Compute $\mathbf{y}^{(k+1)}$ in (20);
4. Expand the solution subspace by computing the residual of the normal equation and update the QR factorizations; see Section 2.

We carry out the iterations until two consecutive iterates are close enough, i.e., until

$$\left\|\mathbf{y}^{(k+1)} - \mathbf{y}^{(k)}\right\|_2 \leq \gamma \left\|\mathbf{y}^{(k)}\right\|_2$$

for some $\gamma > 0$.

We state the two main results of an analysis of this method provided in [9].

Theorem 1 *([9]) Let A be of full column rank and let $\mathbf{x}^{(k)}$ denote the iterates generated by the algorithm described above. Then there exists a convergent subsequence $\{\mathbf{x}^{(k_j)}\}$ whose limit $\mathbf{x}^*$ satisfies*

$$\left\|A\mathbf{x}^* - \mathbf{b}^\delta\right\|_2 = \tau\delta.$$

Theorem 2 *([9]) Let A be of full column rank and let $\mathbf{x}^\delta$ denote the limit (possibly of a subsequence) of the iterates generated by the method described above with data vector $\mathbf{b}^\delta$ such that*

$$\left\|\mathbf{b} - \mathbf{b}^\delta\right\|_2 \leq \delta.$$

Then

$$\limsup_{\delta\searrow 0} \left\|\mathbf{x}^\delta - \mathbf{x}^\dagger\right\|_2 = 0.$$

3.2.2 Residual whiteness principle

The iterated version of the RWP outlined in Section 3.1.2 was proposed in [27] for convex variational models, which can be solved by the Alternating Direction Method of Multipliers (ADMM). The approach described in [27] for the situation when the noise $\boldsymbol{\eta}$ is white Gaussian relies on the following steps:

1. Find an explicit expression of the residual whiteness function W defined in (19) in terms of μ for a general ℓ_2-ℓ_2 variational model.
2. Exploit this expression of W as a function of μ to automatically update the regularization parameter μ in the ℓ_2-ℓ_2 subproblem arising when employing ADMM using a suitable variable splitting.

The iterative procedure proposed in [27] can not be directly applied to the generalized Krylov method considered here. However, to explore the potential of the RWP when applied during the iterations of a generalized Krylov scheme, one can minimize the residual whiteness measure in (19) at each iteration in a similar fashion as for the DP.

For simplicity let us consider the adaptive approach, and note that the extension to the fixed case is straightforward. At each iteration, we solve

$$\mathbf{y}^{(k+1)} = \arg\min_{\mathbf{y}\in\mathbb{R}^{\widehat{k}}} \frac{1}{2}\left\|\left(W_{\text{fid}}^{(k)}\right)^{1/2}\left(AV_k\mathbf{y}-\mathbf{b}^{\delta}\right)\right\|_2^2 + \frac{\mu^{(k)}}{2}\left\|\left(W_{\text{reg}}^{(k)}\right)^{1/2} LV_k\mathbf{y}\right\|_2^2 .$$

Denote by

$$\mathbf{y}_{\mu} = \arg\min_{\mathbf{y}\in\mathbb{R}^{\widehat{k}}} \frac{1}{2}\left\|\left(W_{\text{fid}}^{(k)}\right)^{1/2}\left(AV_k\mathbf{y}-\mathbf{b}^{\delta}\right)\right\|_2^2 + \frac{\mu}{2}\left\|\left(W_{\text{reg}}^{(k)}\right)^{1/2} LV_k\mathbf{y}\right\|_2^2 .$$

The iterated RWP thus reduces to finding the μ-value by solving the minimization problem

$$\mu^{(k)} \in \arg\min_{\mu>0} W(\mu)\,, \quad W(\mu) = \frac{\|\mathbf{d}_{\mu}\star\mathbf{d}_{\mu}\|_2^2}{\|\mathbf{d}_{\mu}\|_2^4}\,, \tag{22}$$

with

$$\mathbf{d}_{\mu} = AV_k\mathbf{y}_{\mu} - \mathbf{b}^{\delta}\,.$$

The minimization problem (22) is solved by means of a Matlab optimization routine that relies on the existence of a closed form for $\mathbf{y}_{\mu}$. Alternatively, one could compute $\mathbf{y}_{\mu}$ on a grid of different μ-values and compute the corresponding whiteness measure function. Nonetheless, we believe that the design of an efficient algorithmic procedure for tackling problem (22) is worth further investigation.

We remark that the outlined strategy defines a new non-stationary method, for which a convergence proof is not yet available.

3.2.3 Generalized Cross Validation

This subsection summarizes the approach presented in [11]. Similarly as above, we consider a non-stationary method for determining μ. This method can be applied to any kind of noise.

Let us first assume that the noise is Gaussian and recall Generalized Cross Validation (GCV) for Tikhonov regularization, i.e., when $p = q = 2$ in (2):

$$\mathbf{x}_\mu = \arg\min_{\mathbf{x}} \left\| A\mathbf{x} - \mathbf{b}^\delta \right\|_2^2 + \mu \left\| L\mathbf{x} \right\|_2^2 . \tag{23}$$

Define the functional

$$G(\mu) = \frac{\left\| A\mathbf{x}_\mu - \mathbf{b}^\delta \right\|_2^2}{\operatorname{trace}\left(I - A(A^TA + \mu L^TL)^{-1}A^T\right)^2}.$$

The regularization parameter determined by the GCV method is given by

$$\mu_{\text{GCV}} = \arg\min_{\mu} G(\mu).$$

When the GSVD of the matrix pair $\{A, L\}$ is available, $G(\mu)$ can be evaluated inexpensively. If the matrices A and L are large, then they can be reduced to smaller matrices by a Krylov-type method and the GCV method can be applied to the reduced problem so obtained; see [11] for more details. This is how we apply the GCV method in the methods of the present paper. Specifically, in the adaptive solution method, we have to solve the minimization problem (12), which is of the same form as (23). Therefore, one can use GCV to compute an appropriate value of μ. Since the matrices R_A and R_L are fairly small, it is possible to compute the GSVD quite cheaply. It follows that the parameter μ_{GCV} can be determined fairly inexpensively. Moreover, the projection into a generalized Krylov subspace accentuates the convexity of G, making minimization easier; see [15, 16] discussions.

We iterate the above approach similarly as we iterated the discrepancy principle, and compute the parameter μ_{GCV} in each iteration. This furnishes a non-stationary algorithm. In detail, consider the kth iteration with regularization parameter $\mu^{(k)}$,

$$\mathbf{y}^{(k+1)} = \arg\min_{\mathbf{y}\in\mathbb{R}^{\hat{k}}} \frac{1}{2} \left\| R_A\mathbf{y} - Q_A^T \left(W_{\text{fid}}^{(k)} \right)^{1/2} \mathbf{b}^\delta \right\|_2^2 + \frac{\mu^{(k)}}{2} \left\| R_L\mathbf{y} \right\|_2^2 .$$

Compute the GSVD of the matrix pair $\{R_A, R_L\}$. It gives the factorizations

$$R_A = U\Sigma_AX^T, \quad R_L = V\Sigma_LX^T,$$

where Σ_A and Σ_L are diagonal matrices, the matrix X is nonsingular, and the matrices U and V have orthonormal columns. We would like to compute the minimizer of

$$G(\mu) = \frac{\left\| R_A\mathbf{y}_\mu - \widehat{\mathbf{b}} \right\|_2^2}{\operatorname{trace}\left(I - R_A(R_A^TR_A + \mu R_L^TR_L)^{-1}R_A^T\right)^2}, \tag{24}$$

where $\widehat{\mathbf{b}} = Q_A^T \left(W_{\text{fid}}^{(k)}\right)^{1/2} \mathbf{b}^\delta$ and $\mathbf{y}_\mu = (R_A^T R_A + \mu R_L^T R_L)^{-1} R_A^T \widehat{\mathbf{b}}$. Substituting the GSVD of $\{R_A, R_L\}$ into (24), we get

$$G(\mu) = \frac{\left\|(\Sigma_A(\Sigma_A^T\Sigma_A + \mu\Sigma_L^T\Sigma_L)^{-1}\Sigma_A^T - I)U^T\widehat{\mathbf{b}}\right\|_2^2}{\operatorname{trace}\left(I - \Sigma_A(\Sigma_A^T\Sigma_A + \mu\Sigma_L^T\Sigma_L)^{-1}\Sigma_A^T\right)^2}.$$

Since Σ_A and Σ_L are diagonal matrices, the value of $G(\mu)$ can be computed cheaply for any value of μ; see, e.g., [5] for a derivation. We let at each iteration

$$\mu^{(k)} = \arg\min_\mu G(\mu).$$

Note that, even though it does not appear explicitly in the formulas, the function G varies with k.

We now consider the case in which the noise is not Gaussian. Since, in the continuous setting, the value of $\left\|A\mathbf{x}_\mu - \mathbf{b}^\delta\right\|_2$ is infinite when $\mathbf{b}^\delta$ is corrupted by impulse noise, a smoothed version of GCV was proposed in [11]. Let $\mathbf{b}^\delta_{\text{smooth}}$ denote a smoothed version of $\mathbf{b}^\delta$ obtained by convolving $\mathbf{b}^\delta$ with a Gaussian kernel; see [11] for details. Consider the smoothed function

$$G_{\text{smooth}}(\mu) = \frac{\left\|A\mathbf{x}_\mu - \mathbf{b}^\delta_{\text{smooth}}\right\|_2^2}{\operatorname{trace}\left(I - A(A^TA + \mu L^TL)^{-1}A^T\right)^2}$$

and the parameter μ is determined by minimizing $G_{\text{smooth}}(\mu)$. The non-stationary ℓ_p-ℓ_q method in this case is obtained in a similar fashion as above. We therefore do not dwell on the details. In the following, we will refer to this method as the GCV method regardless of the type of noise, and use G for Gaussian noise and G_{smooth} for non-Gaussian noise.

4 Numerical experiments

We compare the parameter choice rules described above when applied to some image deblurring problems. These problems can be modeled by a Fredholm integral equation of the first kind

$$g(s,t) = \int_{\mathbb{R}^2} k(s,u,t,v) f(u,v) du dv, \tag{25}$$

where the function g represents the blurred image, k is a possibly smooth integral kernel with compact support, and f is the sharp image that we would like to recover. Since k has compact support and is smooth, the solution of (25) is an ill-posed problem. When the blur is spatially invariant, the integral equation (25) reduces to a convolution

$$g(s,t) = \int_{\mathbb{R}^2} k(s-u, t-v) f(u,v) du dv.$$

The kernel k is often referred to as the Point Spread Function (PSF). After discretization we obtain a linear system of equation of the form (1). Since we only have access to a limited field of view (FOV), it is customary to make an assumption about the behavior of f outside the FOV, i.e., we impose boundary conditions to the problem; see [20] for more details on image deblurring.

We consider three different images and kinds of blur, and three different types of noise. Specifically, we regard the cameraman image (242×242) and blur it with a motion PSF, the boat image (222×222), which we blur with a PSF that simulates the effect of taking a picture while one's hands are shaking, and the clock image (226×226) with a Gaussian PSF; see Figure 1. We consider Gaussian noise, Laplace noise, and a mixture of impulse and Gaussian noise. In the first case, we scale the noise so that $\|\boldsymbol{\eta}\|_2 = 0.02\,\|\mathbf{b}\|_2$, the second case is obtained by setting $\theta = 1$ and $\sigma = 5$ in (3), and in the third case we first modify randomly 20% of the pixels of $\mathbf{b}$ and then add white Gaussian noise so that $\|\boldsymbol{\eta}\|_2 = 0.01\,\|\mathbf{b}\|_2$. Following [6], we set $p = 2$ when the data is corrupted by Gaussian noise, $p = 1$ when the data is contaminated by Laplace noise, and $p = 0.8$ in the mixed noise case. We let L be a discretization of the gradient operator. Assume, for simplicity, that $n = n_1^2$ and define the matrix

$$L_1 = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \\ 1 & & & -1 \end{bmatrix} \in \mathbb{R}^{n_1 \times n_1}.$$

Then

$$L = \begin{bmatrix} L_1 \otimes I_{n_1} \\ I_{n_1} \otimes L_1 \end{bmatrix} \in \mathbb{R}^{2n \times n},$$

where I_{n_1} denotes the $n_1 \times n_1$ identity matrix and $\otimes$ is the Kronecker product. Natural images typically have a sparse gradient. Therefore, we set $q = 0.1$.

We now briefly discuss the computational effort required by each parameter choice rule. The cost of the stationary methods depends on how many values of μ and, for CV and MCV, on how many training and testing sets are considered. For DP and RWP we sample 15 values of μ, while for CV and MCV we consider 10 values of μ and 10 different training sets. Therefore, DP and RWP require the MM algorithm to be run 15 time, while CV and MCV run the MM algorithm 101 and 201 times, respectively. The non-stationary methods require a single run, however, the regularization parameter μ_k has to be tuned at each iteration. Therefore, the cost of a single run of a non-stationary method is, in general, more expensive that the cost of single run of a stationary method. Nevertheless, since the computation of the parameter μ_k can be performed cheaply, thanks to the projection into the generalized Krylov subspace, the non-stationary methods are overall more computationally efficient than their stationary counterparts.

Table 1: RRE obtained with the different parameter choice rules in the computed examples.

	Noise	Parameter Selection Rule							
		Stationary					Non-stationary		
		Optimal	DP	RWP	CV	MCV	DP	RWP	GCV
Cameraman	Gaussian	0.081373	0.083719	**0.083260**	0.089404	0.089404	0.093004	0.085529	0.121020
	Laplace	0.107120	−−	**0.108342**	0.115830	0.667481	−−	0.108974	0.621421
	Impulse + Gaussian	0.096814	−−	−−	0.203390	**0.145490**	−−	−−	0.396891
Boat	Gaussian	0.089190	0.094686	0.094041	**0.089190**	0.262010	0.091353	0.093658	0.328060
	Laplace	0.118621	−−	**0.119582**	0.126140	0.979573	−−	0.120721	0.925001
	Impulse + Gaussian	0.095678	−−	−−	**0.197083**	0.983831	−−	−−	0.232542
Clock	Gaussian	0.035170	0.040890	0.038767	**0.035170**	0.035588	0.049876	0.037852	0.686450
	Laplace	0.040524	−−	0.044341	**0.040524**	0.043585	−−	0.043287	0.721820
	Impulse + Gaussian	0.043197	−−	−−	**0.129171**	0.722730	−−	−−	0.196423

We compare the performances of the considered parameter choice rules in term of accuracy using the Relative Restoration Error (RRE), defined as

$$\mathrm{RRE}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_{\mathrm{true}}\|_2}{\|\mathbf{x}_{\mathrm{true}}\|_2}.$$

Table 1 displays values of the RRE. We report the optimal RRE as well, i.e., the RRE obtained by hand-tuning μ to minimize the RRE. For each test image and noise type, the RRE value that is closest to the optimal one is reported in bold.

Since the DP only can be used for Gaussian noise, we cannot determine the DP parameter for Laplace and mixed noise, while the RWP can be used only for white noise. Table 1 shows the DP and RWP to usually provide very accurate reconstructions, that is the achieved RRE is very close to the optimal one. The limitation of the DP is that it requires a fairly accurate estimate of $\|\boldsymbol{\eta}\|_2$ to be known. However, due to the Bakushinskii veto [2], the DP method is the only one for which a complete theoretical analysis is possible.

The CV, MCV, and GCV methods can be applied to any type of noise; however, being so-called heuristic methods, they may fail in certain situations. Table 1 indicates that the MCV tends to select a μ-value that in some cases leads to significantly large RREs. The same behavior can be observed for the other CV-based strategies in the mixed noise case. Nonetheless, in general the CV algorithm provides the most consistent results.

For the strategies that admit both stationary and non-stationary formulations, the DP for Gaussian noise and the RWP for Gaussian and Laplace noises, we observe that for the two scenarios the corresponding RRE are very close, suggesting that the computed restorations are very similar. This behavior, which can be rigorously justified when the DP is used, provides empirical evidences of the robustness of the RWP.

We report in Figure 2 the reconstructions obtained with the optimal choice of the parameter μ in all considered cases. Note that, if the parameter μ is chosen properly, then ℓ^p-ℓ^q minimization is able to determine very accurate reconstructions.

5 Conclusions

This paper compares a few parameter choice rules for the ℓ^p-ℓ^q minimization method. Their pros and cons are discussed and their performances are illustrated. We have shown that, if the regularization parameter is tuned carefully, then the ℓ_p-ℓ_q model, solved by means of the generalized Krylov method, can provide very accurate reconstructions. The RWP, which here has been applied with the ℓ_p-ℓ_q model for the first time, can be seen to be particularly robust and is able to determine restorations of high quality.

Acknowledgment

A.B., M.P., and F.S. are members of the GNCS-INdAM group.

References

1. Almeida, M., Figueiredo, M.: Parameter estimation for blind and non-blind deblurring using residual whiteness measures. IEEE Transactions on Image Processing **22**, 2751–2763 (2013)
2. Bakushinskii, A.: Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion. USSR Computational Mathematics and Mathematical Physics **24**(4), 181–182 (1984)
3. Baloch, G., Ozkaramanli, H., Yu, R.: Residual correlation regularization based image denoising. IEEE Signal Processing Letters **25**, 298–302 (2018)
4. Bianchi, D., Buccini, A., Donatelli, M., Randazzo, E.: Graph Laplacian for image deblurring. Electronic Transactions on Numerical Analysis **55**, 169–186 (2021)
5. Buccini, A.: Generalized Cross Validation stopping rule for Iterated Tikhonov regularization. In: 2021 21st International Conference on Computational Science and Its Applications (ICCSA), pp. 1–9. IEEE (2021)
6. Buccini, A., De la Cruz Cabrera, O., Donatelli, M., Martinelli, A., Reichel, L.: Large-scale regression with non-convex loss and penalty. Applied Numerical Mathematics **157**, 590–601 (2020)
7. Buccini, A., Pasha, M., Reichel, L.: Generalized singular value decomposition with iterated Tikhonov regularization. Journal of Computational and Applied Mathematics **373**, 112,276 (2020)
8. Buccini, A., Pasha, M., Reichel, L.: Modulus-based iterative methods for constrained ℓ^p-ℓ^q minimization. Inverse Problems **36**(8), 084,001 (2020)
9. Buccini, A., Reichel, L.: An ℓ^2-ℓ^q regularization method for large discrete ill-posed problems. Journal of Scientific Computing **78**, 1526–1549 (2019)
10. Buccini, A., Reichel, L.: An ℓ^p-ℓ^q minimization method with cross-validation for the restoration of impulse noise contaminated images. Journal of Computational and Applied Mathematics **375**, 112,824 (2020)

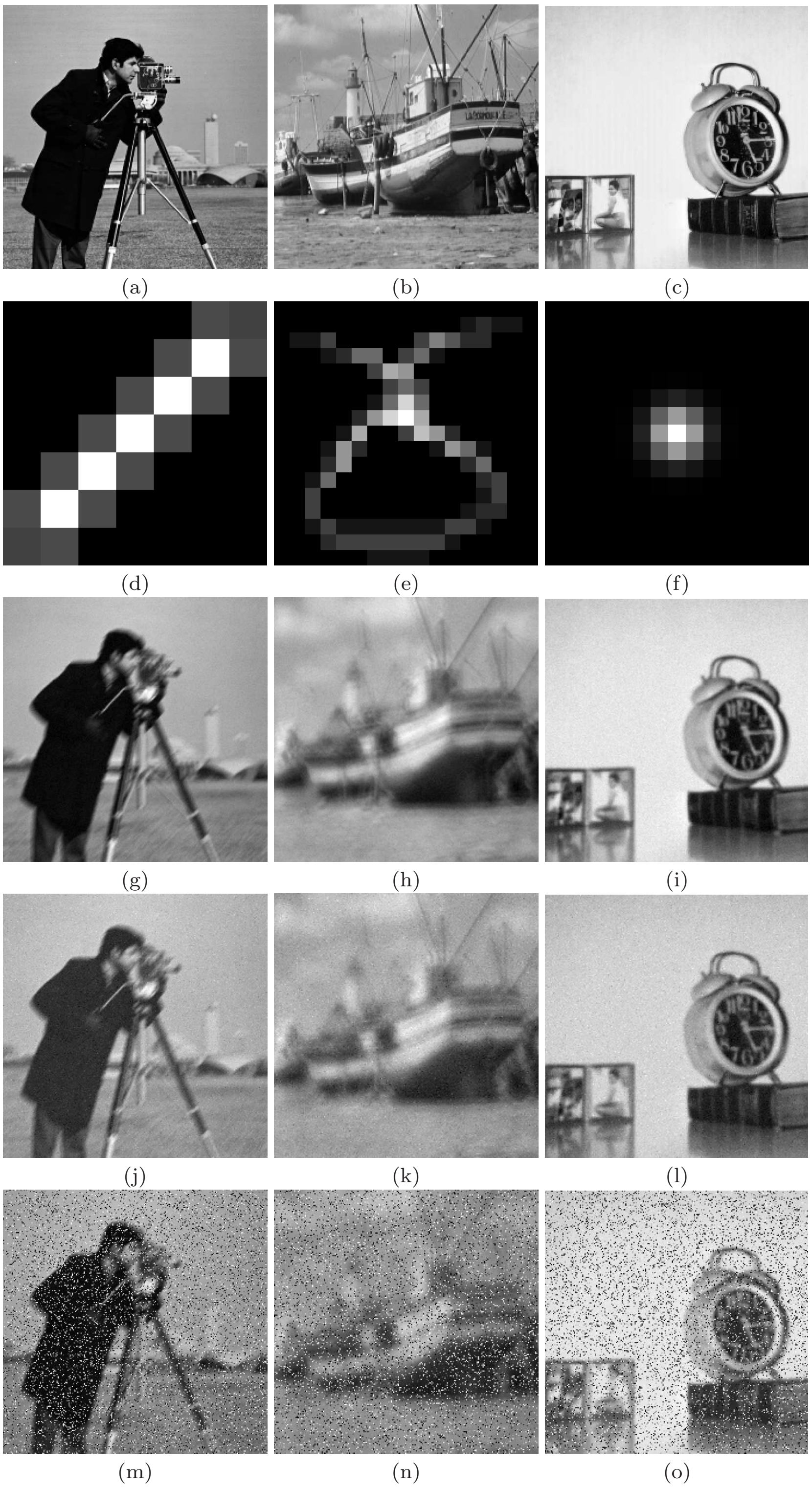

Fig. 1: Test images considered. Cameraman test example: (a) true image (242×242 pixels), (d) PSF (7×7 pixels), (g) blurred image with 2% of white Gaussian noise, (j) blurred image with Laplace noise with $\sigma = 5$, (m) blurred image with 20% of impulse noise and 1% of white Gaussian noise. Boat test example: (b) true image (222×222 pixels), (e) PSF (17×17 pixels), (h) blurred image with 2% of white Gaussian noise, (k) blurred image with Laplace noise with $\sigma = 5$, (n) blurred image with 20% of impulse noise and 1% of white Gaussian noise. Clock test example: (c) true image (226×226 pixels), (f) PSF (15×15 pixels), (i) blurred image with 2% of white Gaussian noise, (k) blurred image with Laplace noise with $\sigma = 5$, (o) blurred image with 20% of impulse noise and 1% of white Gaussian noise.

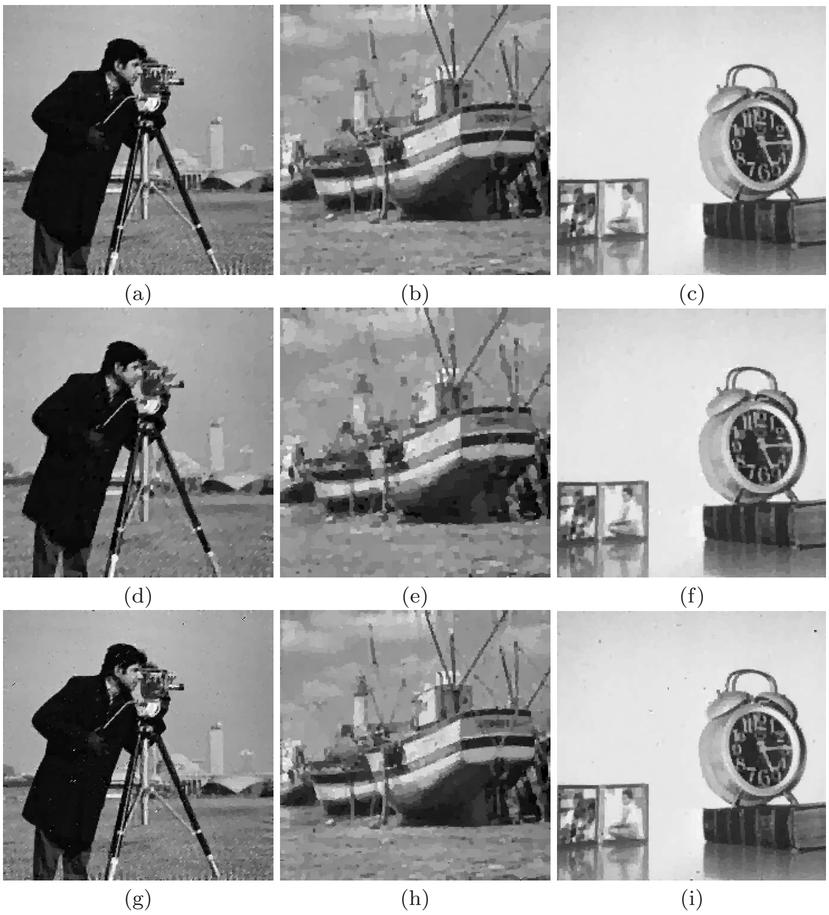

Fig. 2: Recovered images with optimal choice of the parameter μ. Cameraman test example: (a) 2% of white Gaussian noise, (d) Laplace noise with $\sigma = 5$, (g) 20% of impulse noise and 1% of white Gaussian noise. Boat test example: (b) 2% of white Gaussian noise, (e) Laplace noise with $\sigma = 5$, (h) 20% of impulse noise and 1% of white Gaussian noise. Clock test example: (c) 2% of white Gaussian noise, (f) Laplace noise with $\sigma = 5$, (i) 20% of impulse noise and 1% of white Gaussian noise. All the images are projected into $[0, 255]^n$.

11. Buccini, A., Reichel, L.: Generalized Cross Validation for ℓ^p-ℓ^q minimization. Numerical Algorithms **88**, 1595–1616 (2021)
12. Chan, R.H., Liang, H.X.: 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, pp. 78–103. Springer, New York (2014)
13. Daniel, J.W., Gragg, W.B., Kaufman, L., Stewart, G.W.: Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization. Mathematics of Computation **30**(136), 772–795 (1976)
14. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Doordrecht (1996)

15. Fenu, C., Reichel, L., Rodriguez, G.: GCV for Tikhonov regularization via global Golub-Kahan decomposition. Numerical Linear Algebra with Applications **23**(3), 467–484 (2016)
16. Fenu, C., Reichel, L., Rodriguez, G., Sadok, H.: GCV for Tikhonov regularization by partial SVD. BIT Numerical Mathematics **57**(4), 1019–1039 (2017)
17. Galligani, I.: Optimal numerical methods for direct and inverse problems in hydrology. SIMULATION **38**(1), 20–22 (1982)
18. Hansen, P., Kilmer, M., Kjeldsen, R.: Exploiting residual information in the parameter choice for discrete ill-posed problems. BIT Numerical Mathematics **46**, 41–59 (2006)
19. Hansen, P.C.: Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion. SIAM, Philadelphia (1998)
20. Hansen, P.C., Nagy, J.G., O'Leary, D.P.: Deblurring Images: Matrices, Spectra, and Filtering. SIAM, Philadelphia (2006)
21. Huang, G., Lanza, A., Morigi, S., Reichel, L., Sgallari, F.: Majorization-minimization generalized Krylov subspace methods for ℓ_p-ℓ_q optimization applied to image restoration. BIT Numerical Mathematics **57**, 351–378 (2017)
22. Kindermann, S.: Convergence analysis of minimization-based noise level-free parameter choice rules for linear ill-posed problems. Electronic Transactions on Numerical Analysis **38**, 233–257 (2011)
23. Lanza, A., Morigi, S., Reichel, L., Sgallari, F.: A generalized Krylov subspace method for ℓ_p-ℓ_q minimization. SIAM Journal on Scientific Computing **37**, S30–S50 (2015)
24. Lanza, A., Morigi, S., Sciacchitano, F., Sgallari, F.: Whiteness constraints in a unified variational framework for image restoration. Journal of Mathematical Imaging and Vision **60**, 1503–1526 (2018)
25. Lanza, A., Morigi, S., Sgallari, F.: Variational image restoration with constraints on noise whiteness. Journal of Mathematical Imaging and Vision **53**, 61–67 (2015)
26. Lanza, A., Morigi, S., Sgallari, F., Yezzi, A.: Variational image denoising based on autocorrelation whiteness. SIAM Journal on Imaging Sciences **6**, 1931–1955 (2013)
27. Lanza, A., Pragliola, M., Sgallari, F.: Residual whiteness principle for parameter-free image restoration. Electronic Transactions on Numerical Analysis **53**, 329–351 (2020)
28. Lanza, A., Pragliola, M., Sgallari, F.: Automatic fidelity and regularization terms selection in variational image restoration. BIT Numerical Mathematics (2021)
29. Louis, A.K.: Medical imaging: State of the art and future development. Inverse Problems **8**(5), 709–738 (1992)
30. Pragliola, M., Calatroni, L., Lanza, A., Sgallari, F.: ADMM-based residual whiteness principle for automatic parameter selection in super-resolution problems (2021). URL `arxiv.org/abs/2108.13091`
31. Reichel, L., Shyshkov, A.: A new zero-finder for Tikhonov regularization. BIT Numer. Math. **48**, 627–643 (2008)
32. Riot, P., Almansa, A., Gousseau, Y., Tupin, F.: Penalizing local correlations in the residual improves image denoising performance. In: 24th European Signal Processing Conference (EUSIPCO 2016), pp. 1867–1871 (2016)
33. Rust, B.W., O'Leary, D.P.: Residual periodograms for choosing regularization parameters for ill-posed problems. Inverse Problems **24**, 034,005 (2008)
34. Snieder, R., Trampert, J.: Inverse problems in geophysics. In: A. Wirgin (ed.) Wavefield Inversion, pp. 119–190. Springer, Vienna (1999)
35. Stone, M.: Cross-validatory choice and assessment of statistical prediction. Journal of the Royal Statistical Society, Series B **36**, 111–147 (1977)