CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration

Charles-Alban Deledalle, Nicolas Papadakis, Joseph Salmon, Samuel Vaiter

CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration

Charles-Alban Deledalle*, Nicolas Papadakis*, Joseph Salmon†, AND Samuel Vaiter‡

Abstract. In this paper, we propose a new framework to remove parts of the systematic errors affecting popular restoration algorithms, with a special focus for image processing tasks. Extending ideas that emerged for ℓ_1 regularization, we develop an approach that can help re-fitting the results of standard methods towards the input data. Total variation regularizations and non-local means are special cases of interest. We identify important covariant information that should be preserved by the re-fitting method, and emphasize the importance of preserving the Jacobian (w.r.t to the observed signal) of the original estimator. Then, we provide an approach that has a "twicing" flavor and allows re-fitting the restored signal by adding back a local affine transformation of the residual term. We illustrate the benefits of our method on numerical simulations for image restoration tasks.

Key words. inverse problems, image restoration, variational methods, re-fitting, twicing, boosting, debiasing

AMS subject classifications. 49N45, 65K10, 68U10.

1. Introduction.

Restoring an image from its single noisy and incomplete observation necessarily requires enforcing regularity or a model *prior* on the targeted properties of the sought solution. Regularity properties such as sparsity or gradient sparsity of an image, in general, are difficult to enforce, and notably lead to combinatorial and non-convex problems. When one is willing to guarantee such kinds of features for the recovered signal, convex relaxation is a popular path. This is typically done using the ℓ_1 norm instead of the ℓ_0 pseudo-norm, as for the Lasso [43] or the total variation regularization [37]. Nevertheless, such relaxations are well known to create solution with a potentially large bias.

Typically, for the Lasso, using the ℓ_1 convex relaxation of the ℓ_0 pseudo-norm leads large coefficients to be shrunk towards zero.

For the total variation, the same relaxation on the jumps of the signal induces a loss of contrast in the recovered image; see Figure 1.(a) for an illustration in this case.

In the Lasso case, a well known re-fitting scheme consists in performing *a posteriori* a least-square re-estimation of the non-zero coefficients of the solution. This post re-fitting technique became popular under various names in the literature: Hybrid Lasso [20], Lasso-Gauss [35], OLS post-Lasso [2], Debiased Lasso (see [27, 2] for extensive details on the subject). For the anisotropic total-variation (aniso-TV), the same post re-fitting approach can be performed to re-estimate the amplitudes of the jumps, provided their locations have been correctly identified.

In this paper, we introduce a generalization of this re-fitting technique that aims at re-enhancing the estimation towards the data without altering the desired properties imposed by the model prior. To that end, we define a set of properties that need to be preserved. In particular, we introduce the notion of *covariant re-fitting* of the solution. Though this method was originally elaborated with ℓ_1 analysis problems in mind, it has the ability to generalize to a wider family, while in simple cases such as the Lasso or the aniso-TV, it recovers the classical post re-fitting solution

*IMB, CNRS, Université de Bordeaux, Bordeaux INP, F-33405 Talence, France ({charles-alban.deledalle, nicolas.papadakis}@math.u-bordeaux.fr).

†LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013 Paris, France (joseph.salmon@telecom-paristech.fr).

‡IMB, CNRS, Université de Bourgogne, 21078 Dijon, France (samuel.vaiter@u-bourgogne.fr).

described earlier. For instance, our methodology successfully applies to the Tikhonov regularization [45], the isotropic total-variation (iso-TV) and the non-local means [4]. In common variational contexts, *e.g.,* $\ell_1 - \ell_2$ analysis [21] (encompassing the Lasso, the group Lasso [28, 52, 31], the aniso- and iso-TV), we show that our re-fitting technique can be performed with a complexity overload of about twice that of the original algorithm. In other cases, *e.g.,* for the Tikhonov regularization or the non-local means, we introduce a scheme requiring about three times the complexity of the original algorithm.

Moreover, while our covariant re-fitting technique recovers the classical post re-fitting solution in specific cases, the proposed algorithm offers more stable solutions. Indeed, unlike the Lasso post re-fitting technique, ours does not require identifying *a posteriori* the support of the solution, *i.e.,* the set of non-zero coefficients. In the same vein, it does not require identifying the locations of the jumps of the aniso-TV solution. Since the Lasso or the aniso-TV solutions are usually obtained through an iterative algorithm, stopped at a prescribed convergence accuracy, the support or jump numerical identification might be imprecise (all the more as the problem is ill-posed). Such erroneous support identifications can lead to results that strongly deviates from the sought re-fitted solution. Our covariant re-fitting algorithm jointly estimate the re-enhanced solution during the iterations of the original algorithm and, as a by product, produces more robust solutions in practice.

This work follows a preliminary study [15] that attempted to suppress the bias emerging from the choice of the method (*e.g.,* ℓ_1 relaxation), while leaving unchanged the bias due to the unavoidable choice of the model (*e.g.,* sparsity). While the approach from [15] – hereafter referred to as *the invariant re-fitting* – provides interesting results, it is however limited to a class of restoration algorithms that satisfy restrictive local properties.

In particular, the invariant re-fitting cannot handle the classical iso-TV regularizer. As we will see, if such local properties are not satisfied, this technique leads to an unsatisfactory re-fitting that removes some desired aspects enforced by the prior, such as smoothness, and suffer from a significant increase of variance. A simple illustration of this phenomenon for iso-TV is provided in Figure 2.(d) where artificial oscillations are wrongly amplified near the boundary.

While the covariant and the invariant re-fitting both correspond to the least-square post re-fitting step in the case of aniso-TV, the two techniques do not match for iso-TV. Indeed, our Covariant LEAst square Re-fitting (CLEAR), to be precisely described later, outputs a more relevant solution than the one from the invariant re-fitting. Figure 2.(e) shows the benefit of our proposed solution *w.r.t.* the (naive) invariant re-fitting displayed in Figure 2.(d).

It is worth mentioning that the covariant re-fitting is also strongly related to boosting methods re-injecting useful information remaining in the residual (*i.e.,* the map of the point-wise difference between the original signal and its prediction). Such approaches can be traced back to *twicing* [46] and have recently been thoroughly investigated: boosting [5], Bregman iterations and nonlinear inverse scale spaces [32, 6, 50, 33], ideal spectral filtering in the analysis sense [24], SAIF-boosting [29, 42] and SOS-boosting [36] being some of the most popular ones. Note that most of these methods can be performed iteratively, leading to a difficult choice for the number of steps to consider in practice. Our method has the noticeable advantage that it is by construction a two-step one. Iterating more would not be beneficial. Unlike our covariant re-fitting, these later approaches aim at improving the overall image quality by authorizing the re-enhanced result to deviate strongly from the original

biased solution. In particular, they do not recover the aforementioned post re-fitting technique in the Lasso case. Our objective is not to guarantee the image quality to be improved but rather to generalize the re-fitting approach with the ultimate goal of reducing the bias while preserving the structure and the regularity of the original biased solution.

Interestingly, we have also realized that our scheme presents some similarities with the classical shrinking estimators introduced in [40], especially as presented in [23]. Indeed the step performed by CLEAR, is similar to a shrinkage step with a data-driven residual correction weight (later referred to as ρ in our approach, see Definition 21) when performing shrinkage as in [23, Section 3.1].

Last but not least, it is well known that bias reduction is not always favorable in terms of mean square error (MSE) because of the so-called bias-variance trade-off. It is important to highlight that a re-fitting procedure is expected to re-inject part of the variance, therefore it could lead to an increase of residual noise. Hence, the MSE is not always expected to be improved by such re-fitting techniques (unlike the aforementioned boosting-like methods that attempt to improve the MSE). We will show in our numerical experiments, that re-fitting is in practice beneficial when the signal of interest fits well the model imposed by the prior. In other scenarios, when the model mismatches the sought signal, the original biased estimator remains favorable in terms of MSE. Re-fitting is nevertheless essential in this latter case for applications where the image intensities have a physical sense and critical decisions are taken from their values.

2. Background models and notation. We tackle the problem of estimating an unknown vector $x_0 \in \mathbb{R}^p$ from noisy and incomplete observations

$$y = \Phi x_0 + w \ , \tag{1}$$

where Φ is a linear operator from $\mathbb{R}^p$ to $\mathbb{R}^n$ and $w \in \mathbb{R}^n$ is the realization of a noisy random vector. This linear model is widely used in imagery for encoding degradations such as entry-wise masking, convolution or in statistics under the name of linear regression. Typically, the inverse problem associated to (1) is ill-posed, and one should add additional information in order to recover at least an approximation of x_0.

Such additional information leads to the definition of an estimator of x_0 which is represented by an estimation procedure $\hat{x} : y \mapsto \hat{x}(y)$. A popular way to define an estimator is to consider a variational problem in order to make a trade-off between a data fidelity term $F(x, y)$ and a regularizing term $G(x)$ as

$$\hat{x}(y) \in \operatorname*{argmin}_{x \in \mathbb{R}^p} F(x, y) + \lambda G(x) \ . \tag{2}$$

Often, the solution of Problem (2) is non unique, but for simplicity we only consider in this paper a selection of such solutions, and we assume that the selected path $\hat{x} : y \mapsto \hat{x}(y)$ is differentiable almost everywhere.

Another kind of estimator can be defined as the output of an iterative algorithm $(k, y) \mapsto x^k$, *e.g.,* solving an optimization problem. In this context, we define the final estimator $\hat{x}(y) = x^k$ for some chosen k. Such a framework includes proximal splitting methods, *e.g.,* [10, 1], as well as discretization of partial differential equations, thought we do not investigate this latter road in details.

Notation. For a matrix M, M^+ is its Moore-Penrose pseudo-inverse. For a (closed convex) set C, Π_C is the Euclidean projection over C and ι_C is its indicator

function defined by

$$\iota_C(u) = \begin{cases} 0 & \text{if} \quad u \in C \ , \\ +\infty & \text{otherwise} \ . \end{cases} \tag{3}$$

For any vector v, $v_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ is the sub-vector whose elements are indexed by $\mathcal{I} \subset \mathbb{N}$ and $|\mathcal{I}|$ is its cardinality. For any matrix M, $M_{\mathcal{I}}$ is the sub-matrix whose columns are indexed by $\mathcal{I}$. We denote respectively by $\mathrm{Im}[A]$ and $\mathrm{Ker}[A]$ the image space and the kernel space of an operator A.

3. Main estimators investigated. In this section, we provide several examples of $\hat{x}$ that will be of interest to illustrate the different notions and results given all along this paper. We detail in Appendix A how to retrieve closed-form expressions of some of these estimators.

3.1. Affine constrained least-squares. The least-square estimator constrained to the affine subspace $C = b + \mathrm{Im}[A]$, $b \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p \times n}$, is a particular instance of (2) where

$$F(x, y) = \frac{1}{2}\|y - \Phi x\|_2^2 \quad \text{and} \quad G(x) = \iota_C(x) \ . \tag{4}$$

The solution of minimum Euclidean norm is unique and given by (see Appendix A.1)

$$\hat{x}(y) = b + A(\Phi A)^{+}(y - \Phi b) \ . \tag{5}$$

3.2. Tikhonov regularization. The Tikhonov regularization [45] (or Ridge regression [26]) is another instance of (2) where, for some parameter $\lambda > 0$ and matrix $\Gamma \in \mathbb{R}^{m \times p}$,

$$F(x, y) = \frac{1}{2}\|y - \Phi x\|_2^2 \quad \text{and} \quad G(x) = \frac{\lambda}{2}\|\Gamma x\|^2 \ . \tag{6}$$

Provided $\mathrm{Ker}\,\Phi \cap \mathrm{Ker}\,\Gamma = \{0\}$, $\hat{x}(y)$ is uniquely defined as (see Appendix A.2)

$$\hat{x}(y) = (\Phi^\top \Phi + \lambda \Gamma^\top \Gamma)^{-1} \Phi^\top y \ . \tag{7}$$

3.3. Thresholding estimators. The hard-thresholding [17], used when $\Phi = \mathrm{Id}$ and x_0 is supposed to be sparse, is a solution of (2) where, for some parameter $\lambda > 0$,

$$F(x, y) = \frac{1}{2}\|y - x\|_2^2 \quad \text{and} \quad G(x) = \frac{\lambda^2}{2}\|x\|_0 \ , \tag{8}$$

where $\|x\|_0 = |\{i \in [p] \ : \ x_i \neq 0\}|$ counts the number of non-zero entries of x and $[p] = \{1, \dots, p\}$. The hard-thresholding operation $\hat{x}(y) = \mathrm{HT}_\lambda(y)$ writes (see Appendix A.3)

$$(\mathrm{HT}_\lambda(y))_{\mathcal{I}} = y_{\mathcal{I}} \quad \text{and} \quad (\mathrm{HT}_\lambda(y))_{\mathcal{I}^c} = 0 \ , \tag{9}$$

where $\mathcal{I} = \{i \in [p] \ : \ |y_i| > \lambda\}$ and $\mathcal{I}^c$ is the complement of $\mathcal{I}$ in $[p]$.

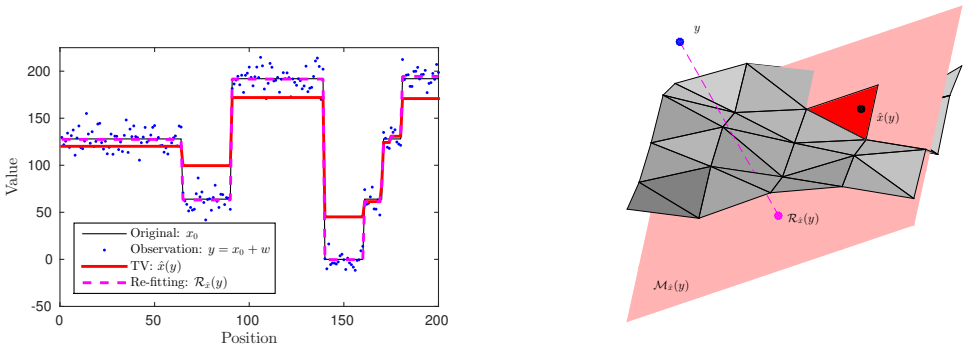
In contrast, the soft-thresholding [17], used when $\Phi = \mathrm{Id}$ and x_0 is supposed to be sparse, is another particular solution of (2) where

$$F(x, y) = \frac{1}{2}\|y - x\|_2^2 \quad \text{and} \quad G(x) = \lambda\|x\|_1 \ , \tag{10}$$

with $\|x\|_1 = \sum_i |x_i|$ being the ℓ_1 norm of x. The soft-thresholding operation $\hat{x}(y) = \mathrm{ST}_\lambda(y)$ writes (see Appendix A.4)

$$(\mathrm{ST}_\lambda(y))_{\mathcal{I}} = y_{\mathcal{I}} - \lambda\,\mathrm{sign}(y_{\mathcal{I}}) \quad \text{and} \quad (\mathrm{ST}_\lambda(y))_{\mathcal{I}^c} = 0 \ , \tag{11}$$

where $\mathcal{I}$ is defined as above. Compared to the hard-thresholding (9), the soft-thresholding suffers form a systematic contraction towards 0 given by $\lambda\,\mathrm{sign}(y_{\mathcal{I}})$.



(a) (b)

FIG. 1. *(a) Solutions of 1D-TV and our re-fitting on a noisy piece-wise constant signal. (b) Geometrical illustration of the invariant re-fitting in a denoising problem of dimension $p = 3$. The gray surface is a piece-wise affine mapping that models the evolution of $\hat{x}$ in an extended neighborhood of y. The light red affine plane is the model subspace, i.e., the set of images sharing the same jumps as those of the solution $\hat{x}(y)$. The red triangle is the restriction of the model subspace to images that can be produced by TV. Finally, the pink dot represents the re-fitting $\mathcal{R}^{inv}_{\hat{x}}(y)$ as the orthogonal projection of y on $\mathcal{M}_{\hat{x}}(y)$.*

3.4. The ℓ_1 synthesis. The ℓ_1 synthesis, also referred to as the Lasso [43, 17], is a particular solution of (2) where

$$F(x,y) = \frac{1}{2}\|\Phi y - x\|_2^2 \quad \text{and} \quad G(x) = \lambda\|x\|_1 \ , \tag{12}$$

with $\|x\|_1 = \sum_{i=1}^p |x_i|$ being the ℓ_1 norm of x. In particular, the soft-thresholding is a Lasso estimator in the special case where $\Phi = \mathrm{Id}$. Provided that the solution is unique (see for instance [44]), the Lasso estimator reads as

$$\hat{x}(y)_{\mathcal{I}} = (\Phi_{\mathcal{I}})^+ y_{\mathcal{I}} - \lambda((\Phi_{\mathcal{I}})^\top \Phi_{\mathcal{I}})^{-1} s_{\mathcal{I}} \quad \text{and} \quad \hat{x}(y)_{\mathcal{I}^c} = 0 \ , \tag{13}$$

where $\mathcal{I} = \mathrm{supp}(\hat{x}(y)) = \{i \in [p] \ : \ \hat{x}(y)_i \neq 0\}$ is the support of $\hat{x}(y)$, $\mathcal{I}^c$ is the complement of $\mathcal{I}$ on $[p]$, and $s_{\mathcal{I}} = \mathrm{sign}(\hat{x}(y)_{\mathcal{I}})$. It then follows that this estimator suffers form a systematic contraction towards 0 given by $\lambda((\Phi_{\mathcal{I}})^\top \Phi_{\mathcal{I}})^{-1} s_{\mathcal{I}}$.

3.5. The ℓ_1 analysis. Given a linear operator $\Gamma \in \mathbb{R}^{m \times p}$, the ℓ_1 analysis minimization reads, for $\lambda > 0$, as

$$F(x,y) = \frac{1}{2}\|\Phi x - y\|^2 \quad \text{and} \quad G(x) = \lambda\|\Gamma x\|_1 \ . \tag{14}$$

Provided $\mathrm{Ker}\,\Phi \cap \mathrm{Ker}\,\Gamma = \{0\}$, there exists a solution given implicitly, see [47], as

$$\hat{x}(y) = U(\Phi U)^+ y - \lambda U(U^\top \Phi^\top \Phi U)^{-1} U^\top (\Gamma^\top)_{\mathcal{I}} s_{\mathcal{I}} \ , \tag{15}$$

for almost all y and where $\mathcal{I} = \mathrm{supp}(\Gamma\hat{x}(y)) = \{i \ : \ (\Gamma\hat{x}(y))_i \neq 0\} \subseteq [m]$ is called the Γ-support of the solution, $s_{\mathcal{I}} = \mathrm{sign}((\Gamma\hat{x}(y))_{\mathcal{I}})$, U is a matrix whose columns form a basis of $\mathrm{Ker}[\Gamma_{\mathcal{I}^c}]$ and ΦU has full column rank. Note that $s_{\mathcal{I}}$ and U are locally constant almost everywhere since the Γ-support is stable with respect to small perturbations [47]. It then follows that this estimator suffers form a systematic contraction towards $\mathrm{Ker}[\Gamma]$ given by $\lambda U(U^\top \Phi^\top \Phi U)^{-1} U^\top (\Gamma^\top)_{\mathcal{I}} s_{\mathcal{I}}$.

The anisotropic Total-Variation (aniso-TV) [37] is a particular instance of (14) where $x_0 \in \mathbb{R}^p$ can be identified to a b-dimensional discrete signal, for which $\Gamma = \nabla : \mathbb{R}^p \to \mathbb{R}^{p\times b}$ is the discrete gradient operator and $\|\nabla x\|_1 = \sum_{i=1}^p \|(\nabla x)_i\|_1$. Aniso-TV promotes piece-wise constant solutions with large constant regions and few sharp transitions. In this case $\mathcal{I}$ is the set of indexes where the solution has discontinuities (non-null gradients). The ℓ_1 norm of the gradient field induces an anisotropic effect by favoring the jumps to be aligned with the canonical directions (in other words, it favors squared like structures rather than rounded ones). It is well known that TV suffers from a systematic loss of contrast: a shift of intensity on each piece depending on its surrounding and the ratio between its perimeter and its area (see [41] for a thorough study for the 1D case).

Figure 1 gives an illustration of TV used for denoising a 1D piece-wise constant signal in $[0, 192]$ and damaged by additive white Gaussian noise (AWGN) with a standard deviation $\sigma = 10$. Even though TV has perfectly retrieved the support of ∇x_0 with one more extra jump, the intensities of some regions are systematically under- or overestimated.

3.6. The $\ell_1 - \ell_2$ analysis. Given a linear operator $\Gamma \in \mathbb{R}^p \mapsto \mathbb{R}^{m\times b}$, the $\ell_1 - \ell_2$ analysis minimization (*a.k.a.,* generalized group Lasso [28, 52], structured or block sparsity [31]) reads, for $\lambda > 0$, as

$$F(x, y) = \frac{1}{2}\|\Phi x - y\|^2 \quad \text{and} \quad G(x) = \lambda\|\Gamma x\|_{1,2} \ , \tag{16}$$

where $\|\Gamma x\|_{1,2} = \sum_{i=1}^m \|(\Gamma x)_i\|_2$. Unlike in the ℓ_1 analysis case, there is no known implicit closed-form expression for the solution of this problem. However, its behavior have been intensively studied. In particular, this estimator is known to suffer form a systematic contraction in the same vein as the one of the ℓ_1 analysis.

The isotropic Total-Variation (iso-TV) [37] is a particular instance of (16) where $x_0 \in \mathbb{R}^p$ can be identified to a b-dimensional discrete signal, for which $\Gamma = \nabla : \mathbb{R}^p \to \mathbb{R}^{p\times b}$ and $\|\nabla x\|_{1,2} = \sum_{i=1}^p \|(\nabla x)_i\|_2$. Like aniso-TV, it promotes solution with large constant regions, but some transition regions can be smooth (see, *e.g.,* [7]), typically those with high curvature in the input image y, see Figure 2.(a)-(c). A major difference is that the $\ell_1 - \ell_2$ norm induces an isotropic effect by favoring rounded like structures rather than squared ones. It also suffers from a systematic loss of contrast.

3.7. Non-local means. The (blockwise) non-local means estimators [4], used when $\Phi = \mathrm{Id}$ and the image $x_0 \in \mathbb{R}^p$ (with $p = p_1 \times p_2$) is composed of many redundant patterns, is the solution of the minimization problem (2) where

$$F(x, y) = \frac{1}{2}\sum_{i,j} w_{i,j}\|\mathcal{P}_i x - \mathcal{P}_j y\|^2 \quad \text{and} \quad G(x) = 0 \ , \tag{17}$$

where $\mathcal{P}_i$ is the linear operator extracting a patch (*i.e.,* a small window) at pixel i of size $(2b + 1) \times (2b + 1)$ where $i \in [p_1] \times [p_2]$ spans the whole image domain and $j - i \in [-s, s] \times [-s, s]$ spans a limited search window. The weights $w_{i,j}$ are usually defined as $w_{i,j} = \varphi\left(\|\mathcal{P}_i y - \mathcal{P}_j y\|_2^2\right)$ where the kernel $\varphi : \mathbb{R}^+ \to [0, 1]$ is a decreasing function which is typically a decay exponential function. Its solution is given in closed form as the following non-local weighted average (see Appendix A.5)

$$\hat{x}(y)_i = \frac{\sum_j \bar{w}_{i,j} y_j}{\sum_j \bar{w}_{i,j}} \quad \text{with} \quad \bar{w}_{i,j} = \sum_k w_{i-k,j-k} \ , \tag{18}$$

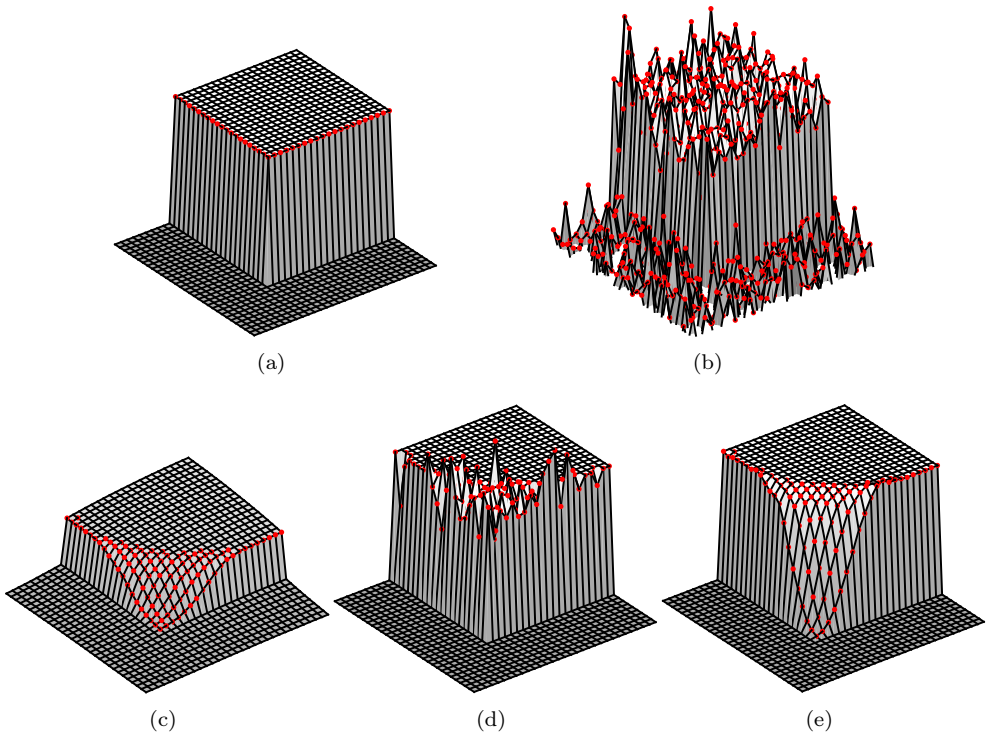


FIG. 2. *(a) A piece-wise constant signal. (b) Its noisy version. (c) Solution of iso-TV on the noisy signal. (d) Solution of the invariant re-fitting of iso-TV. (e) Solution of the covariant re-fitting of iso-TV. Red points indicate locations where the discrete gradient is non-zero.*

where $k \in [-d, d] \times [-d, d]$ spans the patch domain. Note that we assume periodical conditions such that all quantities remain inside the image domain, *i.e.,* all quantities q indexed by $i = (i_1, i_2) \in \mathbb{Z} \times \mathbb{Z}$ are such that $q_i = q_{(i_1, i_2)} = q_{(i_1 + k_1 p_1, i_2 + k_2 p_2)}$ for all $(k_1, k_2) \in \mathbb{Z} \times \mathbb{Z}$.

4. Invariant least-square re-fitting. Practitioners have realized that a systemic contraction affect estimators like the Lasso and the anisotropic total-variation. In the Lasso case, a simple remedy (presented in the introduction) is to perform *a posteriori* a least-square re-fitting step of the non-zero coefficients, *i.e.,* constrained to the support $\mathcal{I}$ of the Lasso solution $\hat{x}(y)$, and given by

$$\operatorname*{argmin}_{x \,;\, \mathrm{supp}(x) \subseteq \mathcal{I}} \tfrac{1}{2}\|\Phi x - y\|^2 \quad . \tag{19}$$

In this section, we present a re-fitting procedure, discussed in [15] that generalizes this approach to a broad family of estimators.

4.1. Re-fitting through model subspace least-squares. We investigate a re-fitting procedure well suited for estimators differentiable almost everywhere. It relies on the notion of model subspace, which requires Jacobian matrix computations. From now on, we consider only estimators $y \mapsto \hat{x}(y)$ from $\mathbb{R}^n$ to $\mathbb{R}^p$ and differentiable almost everywhere (*a.e.* differentiable).

In many estimation procedures, there is a (possibly implicit) prior on the structure of the data. Such structures include smoothness, sparsity, auto-similarity or the fact

that the signal is piece-wise constant. Note that the introduced priors can be captured by the following notion of model subspace.

Definition 1. The *model subspace* associated to an *a.e.* differentiable estimator $\hat{x}$ is defined at almost all points $y \in \mathbb{R}^n$ by the affine subspace of $\mathbb{R}^p$

$$\mathcal{M}_{\hat{x}}(y) = \hat{x}(y) + \mathrm{Im}\left[J_{\hat{x}}(y)\right] \quad , \tag{20}$$

where $J_{\hat{x}}(y)$ is the Jacobian matrix of $\hat{x}$ taken at y.

The model subspace captures what is linearly invariant through $\hat{x}$ with respect to small perturbations of y, typically, for the Lasso, it will encode the set of signals sharing the same support. In order to generalize the re-fitting step, it is thus natural to cast it as a constrained optimization procedure preserving the model subspace.

Definition 2. The *invariant re-fitting* associated to an *a.e.* differentiable estimator $y \mapsto \hat{x}(y)$ is given for almost all $y \in \mathbb{R}^n$ by

$$\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y) = \hat{x}(y) + J(\Phi J)^{+}(y - \Phi\hat{x}(y)) \in \operatorname*{argmin}_{x \in \mathcal{M}_{\hat{x}}(y)} \tfrac{1}{2}\|\Phi x - y\|^2 \quad , \tag{21}$$

where $J = J_{\hat{x}}(y)$ is the Jacobian matrix of $\hat{x}$ at the point y. In the following we use the notation J when no ambiguity is possible.

Remark 3. Though we only consider the case $F(x,y) = \frac{1}{2}\|\Phi x - y\|^2$, the extension to more general F (*e.g.,* for logistic regression) reads $\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y) \in \operatorname*{argmin}_{x \in \mathcal{M}_{\hat{x}}(y)} F(x,y)$.

Remark 4. When $\hat{x}(y) \in \mathrm{Im}[J]$, then $\mathcal{M}_{\hat{x}}(y) = \mathrm{Im}[J]$ and $\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y) = J(\Phi J)^{+}(y)$.

We first exemplify the previous definitions for the various variational estimators introduced in Section 3.

Example 5. The affine constrained least-squares defined in Eq. (5) admits everywhere the same Jacobian matrix $J = A(\Phi A)^{+}$ and its affine model subspace is $\mathcal{M}_{\hat{x}}(y) = b + \mathrm{Im}[A(\Phi A)^{\top}]$ (as $\mathrm{Im}[M^{+}] = \mathrm{Im}[M^{\top}]$). Taking $C = \mathbb{R}^p$ with for instance $n = p$, $A = \mathrm{Id}$ and $b = 0$, leads to an unconstrained solution $\hat{x}(y) = \Phi^{+}y$ whose model subspace is $\mathcal{M}_{\hat{x}}(y) = \mathrm{Im}[\Phi^{\top}]$ reducing to $\mathbb{R}^p$ when Φ has full column rank. In this case, the invariant re-fitting is trivial $\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y) = \hat{x}(y)$.

Example 6. The Tikhonov regularization has everywhere the same Jacobian matrix $J = (\Phi^{\top}\Phi + \lambda\Gamma^{\top}\Gamma)^{-1}\Phi^{\top}$ and everywhere the same affine model subspace $\mathcal{M}_{\hat{x}}(y) = \mathrm{Im}[J]$. If follows that $\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y) = J(\Phi J)^{+}y$. In particular, when Φ has full column rank, $\mathcal{M}_{\hat{x}}(y) = \mathbb{R}^p$ and $\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y) = \Phi^{+}y$.

Example 7. The soft-thresholding as well as the hard-thresholding share the same Jacobian matrix given by $J = \mathrm{Id}_{\mathcal{I}} \in \mathbb{R}^{p \times |\mathcal{I}|}$. Their model subspace reads as $\mathcal{M}_{\hat{x}}(y) = \mathrm{Im}[\mathrm{Id}_{\mathcal{I}}] = \mathrm{Im}[J]$ and the invariant re-fitting is for both the hard-thresholding.

Example 8. The Lasso has for Jacobian matrix $J = \mathrm{Id}_{\mathcal{I}}(\Phi_{\mathcal{I}})^{+}$ and since $\Phi_{\mathcal{I}}$ has full column rank, it shares the same model subspace $\mathcal{M}_{\hat{x}}(y) = \mathrm{Im}[\mathrm{Id}_{\mathcal{I}}] = \mathrm{Im}[J]$ as the hard- and soft-thresholding. Its invariant re-fitting is in this case $\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y) = \mathrm{Id}_{\mathcal{I}}(\Phi_{\mathcal{I}})^{+}y$. While the Lasso systematically underestimates the amplitude of the signal by a shift $\lambda \mathrm{Id}_{\mathcal{I}}((\Phi_{\mathcal{I}})^{t}\Phi_{\mathcal{I}})^{-1}s_{\mathcal{I}}$, the re-fitting $\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y)$ is free of such a contraction.

Example 9. For the ℓ_1 analysis, the Jacobian matrix of the solution in (15) reads as $J = U(\Phi U)^{+}$ where ΦU has full column rank, U is a matrix whose columns form a basis of $\mathrm{Ker}[\Gamma_{\mathcal{I}^c}]$ where the Γ-support is $\mathcal{I} = \{i \,:\, (\Gamma\hat{x}(y))_i \neq 0\} \subseteq [m]$. It results

that the model subspace reads as $\mathcal{M}_{\hat{x}}(y) = \mathrm{Ker}[\Gamma_{\mathcal{I}^c}] = \mathrm{Im}[J]$. The extension of the Lasso re-fitting leads to consider the following least-square estimator, constrained on the Γ-support of $\hat{x}(y)$, given by

$$\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y) = U(\Phi U)^{+} y \ . \tag{22}$$

In the case of the aniso-TV denoising, *i.e.,* with $\Phi = \mathrm{Id}$ and $\Gamma = \nabla$, the model subspace is the space of images whose jumps are included in those of the solution $\hat{x}(y)$. The re-fitting procedure $\mathcal{R}^{\mathrm{inv}}_{\hat{x}}$ is thus the projector $\Pi_{\mathrm{Im}[J]} = UU^{+}$ that performs a piece-wise average of its input on each plateau of the solution. A visualization of this re-fitting is provided for a 1D signal in Figure 1.(a).

Example 10. For the $\ell_1 - \ell_2$ analysis, while there is no implicit closed-form expression for $\hat{x}(y)$, its Jacobian has been derived in closed form in [48] and reads as

$$J = U(U^\top \Phi^\top \Phi U + \lambda U^\top \Gamma^\top \Omega \Gamma U)^{-1} U^\top \Phi^\top y \tag{23}$$

$$\text{where} \quad \Omega : z \in \mathbb{R}^{m \times b} \mapsto \begin{cases} \frac{1}{\|(\Gamma \hat{x}(y))_i\|_2} \left(z_i - \left\langle z_i, \frac{(\Gamma \hat{x}(y))_i}{\|(\Gamma \hat{x}(y))_i\|_2} \right\rangle \frac{(\Gamma \hat{x}(y))_i}{\|(\Gamma \hat{x}(y))_i\|_2} \right) & \text{if} \quad i \in \mathcal{I} \ , \\ 0 & \text{otherwise} \ , \end{cases}$$

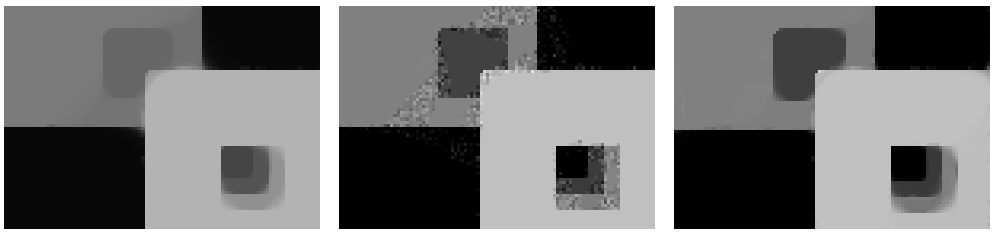
and U and $\mathcal{I}$ are defined exactly as for the ℓ_1 analysis case. Note that Eq. (23) is well founded as soon as $\mathrm{Ker}[\Gamma_{\mathcal{I}^c}] \cap \mathrm{Ker}\, \Phi = \{0\}$. For weaker conditions, see [48, example 26]. In the particular case where ΦU has full column rank, the model subspace matches with the one of the ℓ_1 analysis, *i.e.,* $\mathcal{M}_{\hat{x}}(y) = \mathrm{Ker}[\Gamma_{\mathcal{I}^c}] = \mathrm{Im}[J]$, and the re-fitting is also given by Eq. (22), hence $\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y) = U(\Phi U)^{+} y$. As a consequence, for the iso-TV denoising, *i.e.,* when $\Phi = \mathrm{Id}$ and $\Gamma = \nabla$, the re-fitting step consists again in performing a piece-wise average on each plateau of the solution.

Example 11. For the non-local means, the Jacobian has a complex structure, and we refer the interested reader to [49, 19]. In particular, computing the projection on the model subspace is challenging in this case and so is the computation of the invariant re-fitting. Note that a greedy procedure was proposed in [15] to compute the invariant re-fitting.

4.2. Results and limitations. Figure 1.(a) gives an illustration of the invariant re-fitting in the case of a 1D total-variation denoising example (ℓ_1 analysis estimator). It recovers ∇x_0, the support of the underlying signal, with one extra jump, but systematically under-estimates the amplitude of the jumps. As expected, our re-fitting re-enhances the amplitudes of all plateaus towards the data. Figure 1.(b) gives a geometrical interpretation in dimension $p = 3$ of the model subspace and the invariant re-fitting. The model subspace is represented as the tangent plane of $\hat{x}$ at y and its re-fitting is the projection of y on this tangent plane. Every element of this plane shares the same jumps as those of the solution $\hat{x}(y)$. Figure 3.(a)-(c) gives a similar illustration in the case of a 2D aniso-TV denoising example.

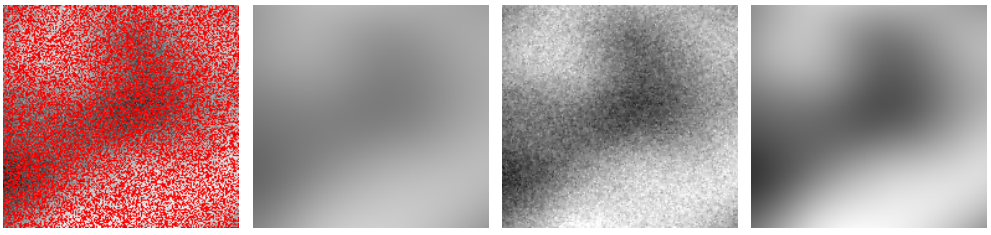
While the invariant re-fitting acts properly for the ℓ_1 analysis estimator, it is however less appealing in other scenarios. Figure 2.(c),(d) and Figure 3.(d),(e) give two illustrations of the invariant re-fitting of a 2D iso-TV denoising example. As for aniso-TV, the invariant re-fitting is the projection of y on the space of signals whose jumps are located at the same position as those of $\hat{x}(y)$. But unlike the anisotropic case, $\hat{x}(y)$ is not piece-wise constant. Instead of being composed of distinct flat regions, it reveals smoothed transitions with dense supports (referred to as extended supports in [7]), see Figure 2.(c). As a consequence, the invariant re-fitting re-introduces a large

(a) Noisy obs. y (b) aniso-TV $\hat{x}(y)$ (c) Ref. $\mathcal{R}^{\text{inv}}_{\hat{x}}(y)$ $(=\mathcal{R}_{\hat{x}}(y))$



(d) iso-TV $\hat{x}(y)$ (e) Inv. ref. $\mathcal{R}^{\text{inv}}_{\hat{x}}(y)$ (f) Cov. ref. $\mathcal{R}_{\hat{x}}(y)$

FIG. 3. *(a) A noisy image $y = x_0 + w$ where x_0 contains six overlapping squares. (b) The solution $\hat{x}(y)$ of aniso-TV, and(c) its invariant re-fitting $\mathcal{R}^{inv}_{\hat{x}}(y)$ (which coincides in this case with the covariant one $\mathcal{R}_{\hat{x}}(y)$). (d) The solution $\hat{x}(y)$ of iso-TV, (e) its invariant re-fitting $\mathcal{R}^{inv}_{\hat{x}}(y)$ and (f) its covariant one $\mathcal{R}_{\hat{x}}(y)$.*



(a) Masked obs. y (b) Tikhonov $\hat{x}(y)$ (c) Inv. ref. $\mathcal{R}^{\text{inv}}_{\hat{x}}(y)$ (d) Cov. ref. $\mathcal{R}_{\hat{x}}(y)$

FIG. 4. *(a) A noisy and incomplete image $y = \Phi x_0 + w$ where Φ is a masking operator encoding missing pixel values (in red) and x_0 is a smooth signal. (b) The solution $\hat{x}(y)$ of the Tikhonov regularization with $\Gamma = \nabla$, (c) its invariant re-fitting $\mathcal{R}^{inv}_{\hat{x}}(y)$ and (d) its covariant one $\mathcal{R}_{\hat{x}}(y)$.*

amount of the original noisy signal in these smooth but non-constant areas, creating the artifacts observed on Figure 2.(d) and Figure 3.(e).

Figure 4,(a)-(c) gives another illustration of the invariant re-fitting of a 2D Tikhonov masking example (with Φ a diagonal matrix with 0 or 1 elements on the diagonal and $\Gamma = \nabla$). While the dynamic of the Tikhonov solution $\hat{x}(y)$ has been strongly reduced, the re-fitting $\mathcal{R}^{\text{inv}}_{\hat{x}}(y)$ re-fits clearly the solution towards the original intensities. However, such a re-fitting is not satisfactory as it does not preserve the smoothness of the solution $\hat{x}(y)$.

In fact, the model subspace captures only what is linearly invariant through $\hat{x}$ with respect to small perturbations of y. This includes the support of the solution for the iso-TV, and the absence of variations inside $\text{Im}[J]^{\perp}$ for the Tikhonov regularization. In particular, it fails at capturing some of the desirable relationships between the

entries of y and the entries of $\hat{x}(y)$, what we call the *covariants*. These relationships typically encode some of the local smoothness and non-local interactions between the entries of the solution $\hat{x}(y)$. Such crucial information is not encoded in the linear model subspace, but interestingly the Jacobian matrix captures by definition how much the entries of $\hat{x}$ linearly varies with respect to all the entries of y. This is at the heart of *the covariant re-fitting* defined in the next section and, for comparison, it produces the results given in Figure 3.(e), Figure 3.(f) and Figure 4.(d).

5. Covariant LEast-square Re-fitting (CLEAR). The objective of this section is to present our main contribution, the introduction of the covariant re-fitting procedure. We particularly aim at solving the issues raised in Subsection 4.2. Toward this goal, we put a stronger emphasis on the first-order behavior of our original estimator by imposing the conservation of its Jacobian, at least locally. The construction of this re-fitting procedure is a bit more involved than for the invariant re-fitting one. Indeed, we first need to define a procedure which, loosely speaking, takes as input the original estimator and a guess of Φx_0 and outputs a new estimator with desirable properties. We next define our covariant re-fitting by choosing a naive guess for Φx_0, namely y.

5.1. Local approach and desired properties for a suitable re-fitting. In this subsection, our objective is to define, from the original estimator $\hat{x}$ and a guess $z \in \mathbb{R}^n$ of Φx_0, a new estimator $\mathcal{D}_{\hat{x},z} : \mathbb{R}^n \to \mathbb{R}^p$ that satisfies several desirable properties and shares with $\hat{x}$ some first-order properties. After-wise, we will consider the choice $z = y$, and the resulting estimator is going to be our covariant re-fitting $\mathcal{R}_{\hat{x}}$. We are now equipped to introduce such a guess based re-fitting.

Definition 12. Let $\hat{x}$ be an estimator from $\mathbb{R}^n$ to $\mathbb{R}^p$ differentiable at $z \in \mathbb{R}^n$. An estimator $\mathcal{D}_{\hat{x},z} : \mathbb{R}^n \to \mathbb{R}^p$ is called a guess based covariant re-fitting of $\hat{x}$ at z, if

$$\mathcal{D}_{\hat{x},z} \in \operatorname*{argmin}_{h \in \mathcal{H}} \|\Phi h(z) - z\|_2^2 \ , \tag{24}$$

where $\mathcal{H}$ is the set of maps $h : \mathbb{R}^n \to \mathbb{R}^p$ satisfying, for all $y \in \mathbb{R}^n$,

1. **Affine map:** $h(y) = Ay + b$ for some $A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$,
2. **Covariant preserving:** $J_h(z) = \rho J_{\hat{x}}(z)$ for some $\rho \in \mathbb{R}$,
3. **Coherent map:** $h(\Phi \hat{x}(z)) = \hat{x}(z)$.

Definition 12 is natural as it states that a guess based re-fitting of $\hat{x}$ for z should be, in prediction, as close as possible to z. Of course, it should satisfy some extra conditions. First, the estimator should be easy to compute, and we achieve this by choosing first order approximation. This means that locally the estimator is affine. Second, as highlighted in Subsection 4.2, the relative variation of the original estimator with respect to the input should be preserved in order to capture not only the invariant features of the estimator but also its first-order behavior, capturing both its singularities and smoothness. Third, applying a re-fitting step to the prediction obtained by the original estimator at z should not modify this estimate. Indeed, the purpose of such a re-fitting is to be as closed as possible of its input y, while preserving the structure of $\hat{x}(z)$, hence if the input is $y = \Phi \hat{x}(z)$ itself, the result should be unaltered.

The next theorem provides a unique closed form expression for $\mathcal{D}_{\hat{x},z}(y)$.

THEOREM 13. *Let $\hat{x}$ be an estimator from $\mathbb{R}^n$ to $\mathbb{R}^p$ differentiable at $z \in \mathbb{R}^p$. Then, for $\delta = z - \Phi \hat{x}(z)$, the guess based covariant re-fitting, defined in Definition 12,*

exists, is unique if $\Phi J\delta \neq 0$*, and is given by*

$$\mathcal{D}_{\hat{x},z}(y) = \hat{x}(z) + \rho J(y - \Phi\hat{x}(z)) \quad \textit{where} \quad \rho = \begin{cases} \dfrac{\langle \Phi J\delta,\, \delta\rangle}{\|\Phi J\delta\|_2^2} & \textit{if} \quad \Phi J\delta \neq 0\ , \\ 1 & \textit{otherwise}\ , \end{cases} \tag{25}$$

where $J = J_{\hat{x}}(z)$ *is the Jacobian matrix of* $\hat{x}$ *at the point* z*.*

Proof. Let h be a mapping satisfying properties 1., 2. and 3. in the previous definition. Observe that properties 1. and 2. of the set $\mathcal{H}$ together ensures that the estimator is of the form $h(y) = \rho Jy + b$ for some $\rho \in \mathbb{R}$ and $b \in \mathbb{R}^p$. Plugging condition 3. gives that $b = (\mathrm{Id} - \rho J\Phi)\hat{x}(z)$, hence $h(y) = \hat{x}(z) + \rho J(y - \Phi\hat{x}(z))$. Reciprocally, it is easy to see that any estimator of the form $h(y) = \hat{x}(z) + \rho J(y - \Phi\hat{x}(z))$ satisfies properties 1., 2. and 3. It thus remains to find ρ.

Remark that Problem (24) can be recast as a one-dimensional problem

$$\min_{\rho\in\mathbb{R}} \left\{ \|\Phi(\hat{x}(z) + \rho J(z - \Phi\hat{x}(z))) - z\|_2^2 = \|(\mathrm{Id} - \rho\Phi J)(\Phi\hat{x}(z) - z)\|_2^2 \right\} \tag{26}$$

whose unique solution, when $\Phi J(\Phi\hat{x}(z) - z) \neq 0$ is given by (25), and a candidate solution is $\rho = 1$, otherwise. ∎

The case where ΦJ is an orthogonal projector leads to interesting properties for instance when $\hat{x}$ is associated to the constrained least-square, the Lasso or aniso-TV.

Proposition 14. *Assume* ΦJ *is an orthogonal projector. Then, the scaling parameter is* $\rho = 1$*.*

Proof. As ΦJ is an orthogonal projector, we have $\Phi J = (\Phi J)^2 = (\Phi J)^\top$. It follows that $\|\Phi J\delta\|^2 = \langle \Phi J\delta,\, \Phi J\delta\rangle = \langle (\Phi J)^\top \Phi J\delta,\, \delta\rangle = \langle \Phi J\delta,\, \delta\rangle$. Injecting this equality in (25) gives $\rho = 1$. ∎

Statistical interpretation. When y becomes a random vector with expectation Φx_0 and with finite second order moment, the bias and covariances of $\mathcal{D}_{\hat{x},z}$ can be obtained in closed form.

Proposition 15. *Let* Y *be a random vector in* $\mathbb{R}^n$ *such that* $\mathbb{E}[Y] = \Phi x_0$*, and* $\mathrm{Cov}[Y] = \Sigma \in \mathbb{R}^{n\times n}$*. Then* $y \mapsto \mathcal{D}_{\hat{x},z}(y)$ *satisfies*

$$\mathbb{E}[\mathcal{D}_{\hat{x},z}(Y)] - x_0 = (\mathrm{Id} - \rho J\Phi)(\hat{x}(z) - x_0)\ , \tag{27}$$

$$\mathrm{Cov}[\mathcal{D}_{\hat{x},z}(Y), Y] = \rho J\Sigma\ , \tag{28}$$

$$\mathrm{Cov}[\mathcal{D}_{\hat{x},z}(Y)] = \rho^2 J\Sigma J^\top, \tag{29}$$

where the cross covariance matrix is defined as $\mathrm{Cov}[X, Y] = \mathbb{E}[XY^\top] - \mathbb{E}[X]\mathbb{E}[Y]^\top$*, for any random column vectors* X *and* Y *(not necessarily of the same size), and* $\mathrm{Cov}[Y] = \mathrm{Cov}[Y, Y]$*.*

Proof. The first equality is a direct consequence of the linearity of the expectation operator. The second equality arises from the following

$$\begin{aligned} \mathbb{E}[(\hat{x}(z) + \rho J(Y - \Phi\hat{x}(z)))Y^\top] - \mathbb{E}[(\hat{x}(z) + \rho J(Y - \Phi\hat{x}(z)))]\mathbb{E}[Y]^\top \\ = \rho J\left(\mathbb{E}[YY^\top] - \mathbb{E}[Y]\mathbb{E}[Y]^\top\right) \end{aligned} \tag{30}$$

where we use the fact that J and ρ are constant with respect to y since they depend only on the guess z. The third equation follows the same sketch by expanding the expression of $\mathrm{Cov}[\mathcal{D}_{\hat{x},z}(Y)]$. ∎

Proposition 15 provides a closed form expression for the bias, the cross-covariance and the covariance of $\mathcal{D}_{\hat{x},z}$. In the general case, these quantities are much more intricate to derive for a non-linear estimator $\hat{x}$. Nevertheless, the next corollary shows how these quantities relate to those of the first order Taylor expansion of the original estimator $\hat{x}$.

COROLLARY 16. *Let $\mathcal{T}_{\hat{x},z}(y)$ be the tangent estimator of $\hat{x}$ at $z \in \mathbb{R}^n$ defined as*

$$\mathcal{T}_{\hat{x},z}(y) = \hat{x}(z) + J(y - z) \ . \tag{31}$$

Let Y be a random vector in $\mathbb{R}^n$ such that $\mathbb{E}[Y] = \Phi x_0$ and $\mathrm{Cov}[Y] = \Sigma$. Then $y \mapsto \mathcal{T}_{\hat{x},z}(y)$ and $y \mapsto \mathcal{D}_{\hat{x},z}(y)$ satisfy

$$\mathbb{E}[\mathcal{T}_{\hat{x},z}(Y)] - x_0 = (\hat{x}(z) - x_0) + J(\Phi x_0 - z) \ , \tag{32}$$

$$\mathrm{Cov}[\mathcal{D}_{\hat{x},z}(Y), Y] = \rho \mathrm{Cov}[\mathcal{T}_{\hat{x},z}(Y), Y] \ , \tag{33}$$

$$\mathrm{Cov}[\mathcal{D}_{\hat{x},z}(Y)] = \rho^2 \mathrm{Cov}[\mathcal{T}_{\hat{x},z}(Y)] \ , \tag{34}$$

$$\mathrm{Corr}[\mathcal{D}_{\hat{x},z}(Y), Y] = \mathrm{Corr}[\mathcal{T}_{\hat{x},z}(Y), Y] \ , \tag{35}$$

$$\mathrm{Corr}[\mathcal{D}_{\hat{x},z}(Y)] = \mathrm{Corr}[\mathcal{T}_{\hat{x},z}(Y)] \ , \tag{36}$$

where the cross correlation matrix is defined as $\mathrm{Corr}[X, Y]_{i,j} = \mathrm{Cov}[X, Y]_{i,j} / \sqrt{\mathrm{Cov}[X]_{i,i} \mathrm{Cov}[Y]_{j,j}}$, for any random column vectors X and Y (not necessarily of the same size), and $\mathrm{Corr}[Y] = \mathrm{Corr}[Y, Y]$.

Proof. The first relation holds form the expression of $\mathcal{T}_{\hat{x},z}$ and that J does not depend on y. It follows that $\mathrm{Cov}[\mathcal{T}_{\hat{x},z}(Y), Y] = J\Sigma$ and $\mathrm{Cov}[\mathcal{T}_{\hat{x},z}(Y)] = J\Sigma J^\top$. These, jointly with Proposition 15, conclude the proof. ∎

Corollary 16 is essential in this work as it states that, by preserving the Jacobian structure, $\mathcal{D}_{\hat{x},z}(Y)$ cannot depart from the tangent estimator of $\hat{x}$ at z in terms of (cross-)correlations. As a consequence, one can expect that they only differ in terms of expectation, *i.e.,* in terms of bias. The next propositions state that when ΦJ is a projector, the bias in prediction is guaranteed to be reduced by our re-fitting.

PROPOSITION 17. *Let Y be a random vector of $\mathbb{R}^n$ such that $\mathbb{E}[Y] = \Phi x_0$. Assume ΦJ is an orthogonal projector, then $y \mapsto \mathcal{D}_{\hat{x},z}(y)$ satisfies*

$$\|\Phi(\mathbb{E}[\mathcal{D}_{\hat{x},z}(Y)] - x_0)\| \leqslant \|\Phi(\mathbb{E}[\mathcal{T}_{\hat{x},z}(Y)] - x_0)\| \ . \tag{37}$$

Proof. As ΦJ is an orthogonal projector, by virtue of Proposition 14, $\rho = 1$, then

$$\begin{aligned} \|\Phi(\mathbb{E}[\mathcal{T}_{\hat{x},z}(Y)] - x_0)\|^2 &= \|\Phi(\hat{x}(z) + J(\Phi x_0 - z) - x_0)\|^2 \\ &= \|(\mathrm{Id} - \Phi J)\Phi(\hat{x}(z) - x_0) + \Phi J(\Phi \hat{x}(z) - z)\|^2 \\ &= \|(\mathrm{Id} - \Phi J)\Phi(\hat{x}(z) - x_0)\|^2 + \|\Phi J(\Phi \hat{x}(z) - z)\|^2 \\ &= \|\Phi(\hat{x}(z) + J\Phi(x_0 - \hat{x}(z)) - x_0)\|^2 + \|\Phi J(\Phi \hat{x}(z) - z)\|^2 \\ &= \|\Phi(\mathbb{E}[\mathcal{D}_{\hat{x},z}(Y)] - x_0)\|^2 + \|\Phi J \delta\|^2 \end{aligned} \tag{38}$$

which concludes the proof. ∎

Proposition 17 is a bit restrictive as it requires ΦJ to be a projector. Nevertheless, this assumption can be relaxed when z satisfies a more technical assumption as shown in the next proposition.

PROPOSITION 18. *Let Y be a random vector of $\mathbb{R}^n$ such that $\mathbb{E}[Y] = \Phi x_0$. Let $\rho_0 = \frac{\langle \delta_0,\, \Phi J \delta_0 \rangle}{\|\Phi J \delta_0\|^2}$ and $\delta_0 = \Phi(x_0 - \hat{x}(z))$. Assume there exists $\alpha \in [0,1]$ such that*

$$\left| \frac{\rho - \rho_0}{\rho_0} \right| \leqslant \sqrt{1-\alpha} \;, \tag{39}$$

$$\textit{and} \quad \|\Phi J(\delta - \delta_0)\|^2 + 2\langle \delta_0,\, \Phi J(\delta - \delta_0)\rangle \geqslant -\alpha \frac{\langle \delta_0,\, \Phi J \delta_0 \rangle^2}{\|\Phi J \delta_0\|^2} \;. \tag{40}$$

Then, $y \mapsto \mathcal{D}_{\hat{x},z}(y)$ satisfies

$$\|\Phi(\mathbb{E}[\mathcal{D}_{\hat{x},z}(Y)] - x_0)\| \leqslant \|\Phi(\mathbb{E}[\mathcal{T}_{\hat{x},z}(Y)] - x_0)\| \;. \tag{41}$$

Proof. It follows from Proposition 15 and Corollary 16 that $\|\Phi(\mathbb{E}[\mathcal{D}_{\hat{x},z}(Y)] - x_0)\| = \|(\mathrm{Id} - \rho\Phi J)\delta_0\|$ and $\|\Phi(\mathbb{E}[\mathcal{T}_{\hat{x},z}(Y)] - x_0)\| = \|\delta_0 + \Phi J(\delta - \delta_0)\|$. Subsequently, we get that Equation (41) holds true if

$$\|(\mathrm{Id} - \rho\Phi J)\delta_0\|^2 \leqslant \|\delta_0 + \Phi J(\delta - \delta_0)\|^2 \;, \tag{42}$$

$$\textit{i.e.,} \quad \rho^2\|\Phi J\delta_0\|^2 - 2\rho\langle \delta_0,\, \Phi J\delta_0\rangle \leqslant \|\Phi J(\delta - \delta_0)\|^2 + 2\langle \delta_0,\, \Phi J(\delta - \delta_0)\rangle \;. \tag{43}$$

Using Assumption (40), a sufficient condition for Equation (41) to hold is

$$\rho^2\|\Phi J\delta_0\|^2 - 2\rho\langle \delta_0,\, \Phi J\delta_0\rangle + \alpha \frac{\langle \delta_0,\, \Phi J\delta_0\rangle^2}{\|\Phi J\delta_0\|^2} \leqslant 0 \;. \tag{44}$$

The roots of this second order polynomial are given by $(1 \pm \sqrt{1-\alpha})\rho_0$, which concludes the proof. ∎

Remark 19. Remark that requiring (39) is quite natural as it states that ρ should be close enough to the optimal ρ_0 minimizing the discrepancy with regards to Φx_0 (*i.e.,* minimizing $\|\Phi h(z) - \Phi x_0\|_2^2$ for $h \in \mathcal{H}$ defined as in Definition 12). While the condition (40) sounds more technical, it however holds true in several interesting cases. For instance, when $z = \Phi x_0$, Assumption (40) holds true as it would read $0 \geqslant -\alpha \frac{\langle \delta_0,\, \Phi J\delta_0\rangle^2}{\|\Phi J\delta_0\|^2}$ (since $\delta = \delta_0$ and $\rho = \rho_0$). Another case of interest is when ΦJ is an orthogonal projector for which (40) holds true as it would read $\|\Phi J\delta\|^2 - \|\Phi J\delta_0\|^2 \geqslant -\|\Phi J\delta_0\|^2$ (using that $\rho = \rho_0 = 1$, $\langle \cdot,\, \Phi J \cdot \rangle = \|\Phi J \cdot\|^2$, and choosing $\alpha = 1$). Hence, Proposition 18 recovers Proposition 17.

Remark 20. Using the same sketch of proof as Proposition 18, the condition $\left|\frac{\rho - \rho_0}{\rho_0}\right| \leqslant 1$ is sufficient to get that $\|\Phi(\mathbb{E}[\mathcal{D}_{\hat{x},z}(Y)] - x_0)\| \leqslant \|\Phi(\hat{x}(z) - x_0)\|$. In other words, even though ρ as a relative error of 100% with respect to ρ_0, the estimator $y \mapsto \mathcal{D}_{\hat{x},z}(y)$ still reduces the bias of the constant estimator $y \mapsto \hat{x}(z)$. This result remains valid when comparing $y \mapsto \mathcal{D}_{\hat{x},z}(y)$ to the pseudo-oracle estimator $y \mapsto \hat{x}(z) + J(y - \Phi x_0)$, with the notable difference that they moreover share the same correlation structure.

While it is difficult to make a general statement, we can reasonably claim from Proposition 17, Proposition 18 and Remark 19 that the bias in prediction will be reduced by our re-fitting providing ΦJ behaves closely as a projector (*i.e.,* its eigenvalues are concentrated around 0 and 1) and/or z is not too far from Φx_0. In such cases, the estimator $\mathcal{D}_{\hat{x},z}$ can be considered as a debiasing procedure of $\hat{x}$, in the sense that it reduces the bias of $\mathcal{T}_{\hat{x},z}$ while preserving its correlation structure (according to Corollary 16).

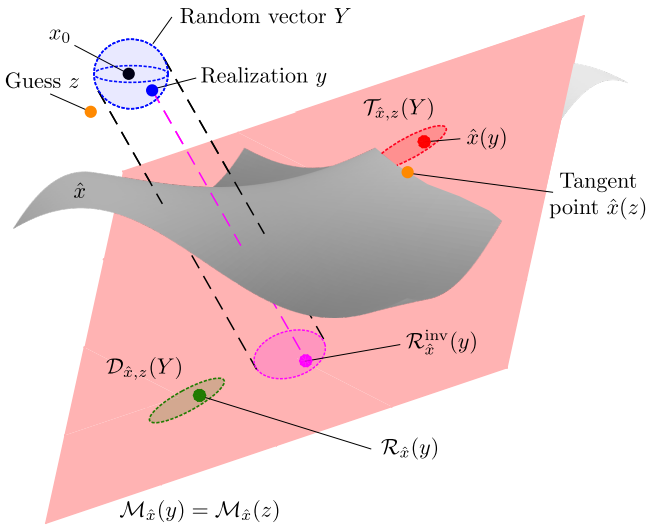


FIG. 5. *Geometrical illustration of the covariant re-fitting in a denoising problem of dimension $p = 3$. We assume that $\mathcal{M}_{\hat{x}}(z) = \mathcal{M}_{\hat{x}}(y)$ for the sake of clarity. The gray surface is the manifold modeling the evolution of $\hat{x}$ in an extended neighborhood of y. The light red affine plane is the model subspace tangent at z. The ellipses represent the positive-definite symmetric covariance matrices of some random vectors, as defined in Proposition 15 and Corollary 16.*

5.2. The covariant least-square re-fitting: definition and properties. Using the procedure $\mathcal{D}_{\hat{x},z}$ defined in Theorem 13, we can now give an explicit definition of our proposed covariant re-fitting as $\mathcal{R}_{\hat{x}}(y) = \mathcal{D}_{\hat{x},y}(y)$.

Definition 21. The *covariant re-fitting* associated to an *a.e.* differentiable estimator $y \mapsto \hat{x}(y)$ is given for almost all $y \in \mathbb{R}^n$ by the relationship

$$\mathcal{R}_{\hat{x}}(y) = \hat{x}(y) + \rho J(y - \Phi\hat{x}(y)) \quad \text{with} \quad \rho = \begin{cases} \dfrac{\langle \Phi J\delta,\, \delta \rangle}{\|\Phi J\delta\|_2^2} & \text{if} \quad \Phi J\delta \neq 0\ , \\ 1 & \text{otherwise}\ , \end{cases} \tag{45}$$

where $\delta = y - \Phi\hat{x}(y)$ and $J = J_{\hat{x}}(y)$ is the Jacobian matrix of $\hat{x}$ at the point y.

Figure 5 gives a geometrical interpretation of CLEAR for a denoising task in dimension $p = 3$. One can observe that if Y varies isotropically, its projection will also vary isotropically on the model subspace. Contrarily, the tangent estimator at a guess z can present an anisotropic behavior along the model subspace, and the guess based re-fitting, which is closer to z, will respect this anisotropy in order to capture the local regularity of $\hat{x}$. Finally, the covariant re-fitting is obtained from the guess based re-fitting at $z = y$. For the sake of clarity, it was assumed on this illustration that $\mathcal{M}_{\hat{x}}(z) = \mathcal{M}_{\hat{x}}(y)$.

Remark 22. The covariant re-fitting performs an additive correction of $\hat{x}(y)$ with a fraction of the directional derivative $J\delta$ in the direction of the residual δ.

Remark 23. Observe that in Definition 21, the value of ρ varies when y varies, contrary to the map $y \mapsto \mathcal{D}_{\hat{x},z}(y)$ for which ρ is constant. Note that the mapping $y \mapsto \mathcal{D}_{\hat{x},z}(y)$ is affine, but *not* the map $y \mapsto \mathcal{R}_{\hat{x}}(y)$. Note that, as a consequence, the statistical interpretations given in the previous section do not hold for $\mathcal{R}_{\hat{x}}(y)$ even though they shed some light on its behavior.

Remark 24. The re-fitting procedure computes $\tilde{x}^2 = \tilde{x}^1 + \rho J(y - \Phi\tilde{x}^1)$, with $\tilde{x}^1 = \hat{x}(y)$. One may wonder if it is beneficial to iterate the process as $\tilde{x}^{k+1} = \tilde{x}^k + \rho J(y - \Phi\tilde{x}^k)$ (in the same vein as [46, 5, 32, 42, 36]). Consider a denoising problem $\Phi = \mathrm{Id}$ with Tikhonov or iso-TV, for which J is symmetrical and $\hat{x}(y) \in \mathrm{Im}[J]$ (see, Example 6 and Example 10). The sequence will converges if and only if $J(y - \tilde{x}^k)$ vanishes, *i.e.,* $\tilde{x}^k$ must converge to $J^+Jy + \zeta$ with $\zeta \in \mathrm{Ker}[J]$. By construction, $\tilde{x}^k \in \mathrm{Im}[J] = \mathrm{Ker}[J^\top]^\perp = \mathrm{Ker}[J]^\perp$, hence $\zeta = 0$. Moreover, as J is symmetrical and $\hat{x}(y) \in \mathrm{Im}[J]$, the quantity J^+Jy coincides with $JJ^+y = \mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y)$ (by virtue of Remark 4), *i.e.,* the invariant re-fitting. Reminding the artifacts illustrated in Figure 3.(e), this is not satisfying.

An interesting property of $\mathcal{R}_{\hat{x}}$ is the fact that it belongs to the model subspace of $\hat{x}$ as stated in the following proposition.

Proposition 25. *Let $y \mapsto \hat{x}(y)$ be an* a.e. *differentiable estimator. Then for almost all $y \in \mathbb{R}^n$, one has $\mathcal{R}_{\hat{x}}(y) \in \mathcal{M}_{\hat{x}}(y)$.*

Proof. Simply recall that $\mathcal{M}_{\hat{x}}(y) = \hat{x}(y) + \mathrm{Im}\, J$, since $\rho J(y - \Phi\hat{x}(y)) \in \mathrm{Im}\, J$ the proposition follows. ∎

Again, the case where ΦJ is an orthogonal projector, leads to interesting properties that will be of main interest regarding several of the estimators considered in Section 3.

Proposition 26. *Suppose that ΦJ is an orthogonal projector. Then, $\mathcal{R}_{\hat{x}}(y) = \hat{x}(y) + J(y - \Phi\hat{x}(y))$, and, $\Phi\mathcal{R}_{\hat{x}}(y) = \Phi\mathcal{R}^{inv}_{\hat{x}}(y)$.*

Proof. By virtue of Proposition 14, $\rho = 1$ and then $\mathcal{R}_{\hat{x}}(y) = \hat{x}(y) + J(y - \Phi\hat{x}(y))$. The fact that $\Phi\mathcal{R}_{\hat{x}}(y) = \Phi\mathcal{R}^{\mathrm{inv}}_{\hat{x}}(y)$ comes from the fact that $\Phi J(\Phi J)^+ = \Phi J$. ∎

The next proposition provides, when $J\Phi$ satisfies a fixed point formulation, an expression of $\mathcal{R}_{\hat{x}}(y)$ as a combination of the estimator with the application of the Jacobian to the noisy signal. This expression will be of main interest regarding the efficient computation of the re-fitting as discussed in Section 6, with a notable example being the iso-TV regularization.

Proposition 27. *Assume that $J\Phi\hat{x}(y) = \hat{x}(y)$. Then, the covariant re-fitting reads $\mathcal{R}_{\hat{x}}(y) = (1 - \rho)\hat{x}(y) + \rho Jy$.*

Proof. We have $\mathcal{R}_{\hat{x}}(y) = \hat{x}(y) + \rho J(y - \Phi\hat{x}(y)) = \hat{x}(y) + \rho Jy - \rho J\Phi\hat{x}(y)$, and since $J\Phi\hat{x}(y) = \hat{x}(y)$ by assumption, this concludes the proof. ∎

Interestingly, the next theorem shows that the condition $J\Phi\hat{x}(y) = \hat{x}(y)$ can be met provided $\hat{x}(y)$ is solution of a variational problem with a 1-homogeneous regularizer.

Theorem 28. *Let $\hat{x}(y)$ be the unique* a.e. *differentiable solution of*

$$\hat{x}(y) = \operatorname*{argmin}_x F(y - \Phi x) + G(x) \quad , \tag{46}$$

with F, G two convex functions and G being 1-homogeneous. Then, for almost all y, $J\Phi\hat{x}(y) = \hat{x}(y)$.

The proof of Theorem 28 is postponed to Appendix B.

The affine constrained least-squares, the ℓ_1 synthesis, the $\ell_1 - \ell_2$ analysis, aniso-TV and iso-TV, are solutions of a variational problem with F being differentiable and

G being 1-homogeneous. As a consequence, Theorem 28 shows that the aforementioned methods satisfy $J\Phi\hat{x}(y) = \hat{x}(y)$, and hence $\mathcal{R}_{\hat{x}}(y) = (1-\rho)\hat{x}(y) + \rho Jy$.

5.3. Examples of re-fitting procedures. In this section, we exemplify the previous definitions for the wide class of variational estimators introduced in Section 3.

Example 29. For the affine constrained least-squares defined in Eq. (5), the Jacobian reads as $J = A(\Phi A)^+$. In this case, $\Phi J = \Phi A(\Phi A)^+$ is an orthogonal projector, $\rho = 1$ and the covariant re-fitting coincides with the invariant one and reads $\mathcal{R}_{\hat{x}}(y) = \mathcal{R}^{\text{inv}}_{\hat{x}}(y) = \hat{x}(y)$.

Example 30. The Tikhonov regularization has for Jacobian $J = (\Phi^\top\Phi + \lambda\Gamma^\top\Gamma)^{-1}\Phi^\top$ and in this case ρ depends on the residual δ and the re-fitting reads as the weighted sum $\mathcal{R}_{\hat{x}}(y) = (1+\rho)\hat{x}(y) - \rho J\Phi\hat{x}(y)$.

Example 31. For the soft-thresholding and the hard-thresholding used when $\Phi = \mathrm{Id}$, the Jacobian reads $J = \mathrm{Id}_{\mathcal{I}} \in \mathbb{R}^{p\times|\mathcal{I}|}$. As a consequence $\Phi J = \mathrm{Id}_{\mathcal{I}}$ is an orthogonal projection and the covariant re-fitting coincides with the invariant one, namely the hard-thresholding itself.

Example 32. For the Lasso, the Jacobian reads $J = \mathrm{Id}_{\mathcal{I}}(\Phi_{\mathcal{I}})^+$ where $\Phi_{\mathcal{I}}$ has full column rank. As for the thresholding, $\Phi J = \Phi U(\Phi U)^+$ is an orthogonal projection and the covariant re-fitting reads $\mathcal{R}_{\hat{x}}(y) = \mathcal{R}^{\text{inv}}_{\hat{x}}(y)$.

Example 33. For the ℓ_1 analysis, the Jacobian reads $J = U(\Phi U)^+$ where ΦU has full column rank and U is a matrix whose columns form a basis of $\mathrm{Ker}[\Gamma_{\mathcal{I}^c}]$ where the Γ-support is $\mathcal{I} = \{i \;:\; (\Gamma\hat{x}(y))_i \neq 0\} \subseteq [m]$. Again, $\Phi J = \Phi U(\Phi U)^+$ is an orthogonal projection and the covariant re-fitting reads $\mathcal{R}_{\hat{x}}(y) = \mathcal{R}^{\text{inv}}_{\hat{x}}(y)$.

Example 34. For the $\ell_1 - \ell_2$ analysis, by virtue of Theorem 28, $J\Phi\hat{x}(y) = \hat{x}(y)$ and hence $\mathcal{R}_{\hat{x}}(y) = (1-\rho)\hat{x}(y) + \rho Jy$. Moreover, we can show that its Jacobian given in Eq. (23), applied to a vector $d \in \mathbb{R}^n$ is a solution of the following problem

$$Jd \in \operatorname*{argmin}_{x\,;\,\mathrm{supp}(\Gamma x)\subseteq\mathcal{I}} \tfrac{1}{2}\|\Phi x - d\|^2 + \tfrac{\lambda}{2}\omega(\Gamma x) \;, \tag{47}$$

$$\text{where } \omega : z \in \mathbb{R}^{m\times b} \mapsto \frac{1}{\|(\Gamma\hat{x}(y))_i\|_2}\left(\|z_i\|_2^2 - \left\langle z_i, \frac{(\Gamma\hat{x}(y))_i}{\|(\Gamma\hat{x}(y))_i\|_2}\right\rangle^2\right) \;.$$

Remark that $\omega(\Gamma Jd) = 0$ if and only if $(\Gamma Jd)_i$ is co-linear to $(\Gamma\hat{x}(y))_i$, for all $i \in \mathcal{I}$. For iso-TV, it means that the level lines of Jd must be included in the ones of $\hat{x}(y)$. As a consequence, it becomes clear that unlike the invariant re-fitting of $\hat{x}(y)$, the covariant re-fitting is constrained to be faithful to the regularity of $\hat{x}(y)$, since it enforces the discontinuities of Jd to be co-linear to $(\Gamma\hat{x}(y))_{\mathcal{I}}$. This is especially important where the iso-TV solution presents transitions with high curvature. Such an appealing behavior of the covariant re-fitting explains the results observed in Figure 2.(e) and Figure 3.(f).

Example 35. For the non-local means, the Jacobian has a more intricate structure. Nevertheless, its directional derivative has a simpler expression given for any direction $d \in \mathbb{R}^n$, by

$$Jd = \frac{\sum_j \bar{w}'_{i,j}y_j + \sum \bar{w}_{i,j}d_j - \hat{x}(y)_i\sum_j \bar{w}'_{i,j}}{\sum_j \bar{w}_{i,j}} \quad \text{with} \quad \bar{w}'_{i,j} = \sum_k w'_{i-k,j-k} \;, \tag{48}$$

where $\hat{x}(y)$ is defined in Eq. (18) and

$$w'_{i,j} = 2\langle \mathcal{P}_i y - \mathcal{P}_j y,\, \mathcal{P}_i d - \mathcal{P}_j d\rangle \varphi'\left(\|\mathcal{P}_i y - \mathcal{P}_j y\|^2\right) \;, \tag{49}$$

with φ' the *a.e.* derivative of the kernel function φ. Subsequently, the covariant re-fitting is obtained from its general form with two steps, by computing first $\hat{x}(y)$, and applying next the Jacobian to the direction $d = y - \hat{x}(y)$.

6. Covariant re-fitting in practice. We now detail how to compute CLEAR from estimators given by standard algorithms. To that end, we first recall some interesting properties of two different differentiation techniques that allow computing some image of J jointly with $\hat{x}(y)$.

6.1. Algorithmic differentiation. Following [16], we consider restoration algorithms whose solutions $\hat{x}(y) = x^k$ are obtained via an iterative scheme of the form

$$\begin{cases} x^k & = \gamma(a^k) \ , \\ a^{k+1} & = \psi(a^k, y) \ , \end{cases} \tag{50}$$

where $a^k \in \mathcal{A}$ is a sequence of auxiliary variables and $\psi : \mathcal{A} \times \mathbb{R}^n \to \mathcal{A}$ is a fixed point in the sense that a^k converges to a fixed point $a^\star$, and $\gamma : \mathcal{A} \to \mathbb{R}^p$ is non-expansive (*i.e.,* $\|\gamma(a_1) - \gamma(a_2)\| \leqslant \|a_1 - a_2\|$ for any $a_1, a_2 \in \mathcal{A}$) entailing that x^k will converge to $x^\star = \gamma(a^\star)$.

As a result, for almost all y and for any direction $d \in \mathbb{R}^n$, the directional derivatives $\mathcal{D}_x^k = J_{\hat{x}^k}(y)d$ and $\mathcal{D}_a^k = J_{a^k}(y)d$ can be jointly obtained with x^k and a^k as

$$\begin{cases} x^k & = \gamma(a^k) \ , \\ a^{k+1} & = \psi(a^k, y) \ , \\ \mathcal{D}_x^k & = \Gamma_a \mathcal{D}_a^k \ , \\ \mathcal{D}_a^{k+1} & = \Psi_a \mathcal{D}_a^k + \Psi_y d \ , \end{cases} \tag{51}$$

where $\Gamma_a = \frac{\partial \gamma(a)}{\partial a}\Big|_{a^k}$, $\Psi_a = \frac{\partial \psi(a,y)}{\partial a}\Big|_{a^k}$ and $\Psi_y = \frac{\partial \psi(a^k,y)}{\partial y}\Big|_{y}$. Interestingly, in all considered cases, the cost of evaluating Γ_a, Ψ_a and Ψ_y is of about the same as the cost of evaluating γ and ψ. As a result, the complexity of (51) is of about twice the complexity of (50). Note that in practice, Γ_a, Ψ_a and Ψ_y can be implemented either by investigating their closed form expression or in a black box manner using automatic differentiation tools. The later strategy has been well studied and we refer to [25, 30] for a comprehensive study.

6.2. Finite difference based differentiation. Another strategy is to approximate the directional derivative by finite differences, for any direction $d \in \mathbb{R}^n$ and $\varepsilon > 0$, as

$$J_{\hat{x}}(y)d \approx \frac{\hat{x}(y + \varepsilon d) - \hat{x}(y)}{\varepsilon} \ . \tag{52}$$

As a result, the complexity of evaluating (52) is also twice the complexity of (50) since $\hat{x}$ must be evaluated at both y and $y + \varepsilon d$. The main advantage of this method is that $\hat{x}$ can be used as a black box, *i.e.,* without any knowledge on the underlying algorithm that provides $\hat{x}(y)$. For ε small enough, it performs as well as the approach described in (51) (with $\hat{x}(y) = x^k$) that requires the knowledge of the derivatives. Indeed, if $y \mapsto \hat{x}(y)$ is Lipschitz-continuous, then (52) converges to (51) when $\varepsilon \to 0$ (by virtue of Rademacher's theorem and [22, Theorem 1-2, Section 6.2]). This implies that the value ε can so be chosen as small as possible (up to machine precision) yielding to an accurate approximation of $J_{\hat{x}}(y)d$. This finite difference approach has been used in many fields, and notably for risk estimation, see, *e.g.,* [51, 39, 34].

6.3. Two-step computation for the general case. In the most general case, the computation of the covariant re-fitting, given by

$$\mathcal{R}_{\hat{x}}(y) = \hat{x}(y) + \rho J(y - \Phi\hat{x}(y)) \quad \text{with} \quad \rho = \frac{\langle \Phi J\delta,\, \delta\rangle}{\|\Phi J\delta\|^2} \quad \text{and} \quad \delta = y - \Phi\hat{x}(y)\ , \tag{53}$$

requires to evaluate sequentially $\hat{x}(y)$ and $J(y - \Phi\hat{x}(y))$.

In the case of finite difference differentiation, two steps are required. First $\hat{x}(y)$ must be computed with the original algorithm and next $J(y - \Phi\hat{x}(y))$ is obtained by finite difference (52) on the direction of the residual $d = y - \Phi\hat{x}(y)$. Once $J(y - \Phi\hat{x}(y))$ is computed, ρ can be evaluated and subsequently (53). The overall complexity is about twice that of the original algorithm producing $\hat{x}(y)$.

In the case of algorithmic differentiation, as $J(y - \Phi\hat{x}(y))$ depends on $\hat{x}(y)$, the original iterative scheme (50) must be run first. In the second step, $J(y - \Phi\hat{x}(y))$ is obtained with the differentiated version (51) on the direction of the residual $d = y - \Phi\hat{x}(y)$. As a result, $\hat{x}(y)$ is unfortunately computed twice, first by (50) and next by (51) leading to an overall complexity of about three times the one of the original algorithm. Nevertheless, in several cases, one can avoid the first step by running only Algorithm (51). This is the topic of the next section.

6.4. One-step computation for specific cases. When $\hat{x}(y)$ fulfills the assumption $J\Phi\hat{x}(y) = \hat{x}(y)$ of Proposition 27, the covariant re-fitting reads as

$$\mathcal{R}_{\hat{x}}(y) = (1 - \rho)\hat{x}(y) + \rho Jy \quad \text{with} \quad \rho = \frac{\langle \Phi(Jy - \hat{x}(y)),\, y - \Phi\hat{x}(y)\rangle}{\|\Phi(Jy - \hat{x}(y))\|_2^2}\ . \tag{54}$$

The computations of $\hat{x}(y)$ and Jy are then sufficient to directly compute the re-fitting $\mathcal{R}_{\hat{x}}(y)$. As a result, in the case of algorithmic differentiation, only Algorithm (51) can be run to get $\mathcal{R}_{\hat{x}}(y)$ since using the direction $d = y$ provides directly $\hat{x}(y)$, Jy and subsequently ρ. Compared to the two step approach, the complexity of the re-fitting reduces to about twice the one of the original step from (50). Recall, that the condition $J\Phi\hat{x}(y) = \hat{x}(y)$ is met for several cases of interest including the Lasso, the Generalize Lasso, aniso-TV and iso-TV. Hence, all of them can be re-enhanced with a complexity being twice the one of the original algorithm.

6.5. Example for a primal dual optimization of the ℓ_1 analysis problem. In this section we instantiate Algorithm (51) to the case of the primal-dual sequence of [8]. To that end, let us first recall some of the properties of primal-dual techniques. By dualizing the ℓ_1 analysis norm $x \mapsto \lambda\|\Gamma x\|_1$, the primal problem (14) can be reformulated, with $x^\star = \hat{x}(y)$, as the following saddle-point problem

$$(z^\star, x^\star) \in \arg\max_{z\in\mathbb{R}^m} \min_{x\in\mathbb{R}^p} \frac{1}{2}\|\Phi x - y\|^2 + \langle \Gamma x, z\rangle - \iota_{\mathcal{B}_\lambda}(z)\ , \tag{55}$$

where $z^\star \in \mathbb{R}^m$ is the dual variable, and $\mathcal{B}_\lambda = \{z \in \mathbb{R}^m \ :\ \|z\|_\infty \leqslant \lambda\}$ is the ℓ_∞ ball.

Problem (55) can be efficiently solved using the primal-dual algorithm of [8]. By taking $\sigma\tau < \frac{1}{\|\Gamma\|_2^2}$, $\theta \in [0, 1]$ and initializing (for instance,) $x^0 = v^0 = 0 \in \mathbb{R}^p$, $z^0 = 0 \in \mathbb{R}^m$, the algorithm reads

$$\left\{\begin{array}{ll} z^{k+1} & = \Pi_{\mathcal{B}_\lambda}(z^k + \sigma\Gamma v^k)\ , \\ x^{k+1} & = (\mathrm{Id} + \tau\Phi^\top\Phi)^{-1}\left(x^k + \tau(\Phi^\top y - \Gamma^\top z^{k+1})\right)\ , \\ v^{k+1} & = x^{k+1} + \theta(x^{k+1} - x^k)\ , \end{array}\right. \tag{56}$$

where the projection of z over $\mathcal{B}_\lambda$ is done component-wise as

$$\Pi_{\mathcal{B}_\lambda}(z)_i = \begin{cases} z_i & \text{if} \quad |z_i| \leqslant \lambda \ , \\ \lambda \ \mathrm{sign}(z_i) & \text{otherwise} \ . \end{cases} \tag{57}$$

The primal-dual sequence x^k converges to a solution $x^\star$ of the ℓ_1 analysis problem (14) as proved in [8].

It is easy to check that the primal dual sequence defined in (56) can be written in the general form considered in (50), see for instance [16]. As a result, we can use the algorithmic differentiation based strategy described by (51) and that reads, for initializations $\tilde{x}^0 = \tilde{v}^0 = 0 \in \mathbb{R}^p$, $\tilde{z}^0 = 0 \in \mathbb{R}^m$, and for $\beta = 0$, as

$$\left\{ \begin{array}{ll} z^{k+1} & = \Pi_{\mathcal{B}_\lambda}(z^k + \sigma\Gamma v^k) \ , \\ x^{k+1} & = (\mathrm{Id} + \tau\Phi^\top\Phi)^{-1}\left(x^k + \tau(\Phi^\top y - \Gamma^\top z^{k+1})\right) \ , \\ v^{k+1} & = x^{k+1} + \theta(x^{k+1} - x^k) \ , \\ \tilde{z}^{k+1} & = \Pi_{z^k + \sigma\Gamma v^k}(\tilde{z}^k + \sigma\Gamma\tilde{v}^k) \ , \\ \tilde{x}^{k+1} & = (\mathrm{Id} + \tau\Phi^\top\Phi)^{-1}\left(\tilde{x}^k + \tau(\Phi^\top y - \Gamma^\top \tilde{z}^{k+1})\right) \ , \\ \tilde{v}^{k+1} & = \tilde{x}^{k+1} + \theta(\tilde{x}^{k+1} - \tilde{x}^k) \ , \end{array} \right. \tag{58}$$

$$\text{where} \quad \Pi_z(\tilde{z})_i = \begin{cases} \tilde{z}_i & \text{if} \quad |z_i| \leqslant \lambda + \beta \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

Recall that the re-fitting is directly $\mathcal{R}_{x^k}(y) = \tilde{x}^k$, since $J\Phi$ is an orthogonal projector for the ℓ_1 analysis case.

Remark that the algorithmic differentiation of (56) is exactly (58) for $\beta = 0$, hence, $\tilde{x}^k = \mathcal{R}_{x^k}(y)$. However, if one wants to guarantee the convergence of the sequence $\tilde{x}^k$ towards $\mathcal{R}_{\hat{x}}(y)$, one needs a small $\beta > 0$ as shown in the next theorem. In practice, β can be chosen as the smallest available positive floating number.

THEOREM 36. *Assume that $x^\star$ satisfies (15) with ΦU full-column rank[1]. Let $\alpha > 0$ be the minimum non zero value[2] of $|\Gamma x^\star|_i$ for all $i \in [m]$. Choose β such that $\alpha\sigma > \beta > 0$. Then, the sequence $\tilde{x}^k = \mathcal{R}_{x^k}(y)$ defined in (58) converges to the re-fitting $\mathcal{R}_{\hat{x}}(y)$ of $\hat{x}(y) = x^\star$.*

The proof of this theorem is postponed to Appendix C.

A similar result was obtained in [14] when solving the ℓ_1 analysis problem (14) with the Douglas-Rachford splitting algorithm described in [18, 9].

6.6. Example for a primal dual optimization of the $\ell_1 - \ell_2$ analysis problem. The algorithm for the $\ell_1 - \ell_2$ analysis regularization can be derived with the exact same considerations as for the ℓ_1 analysis case. The only difference in the application of the primal dual algorithm comes from the non linear operation (57) that now reads, for $z \in \mathbb{R}^{m \times b}$, as

$$\Pi^{\mathrm{iso}}_{\mathcal{B}_\lambda}(z)_i = \begin{cases} z_i & \text{if} \quad \|z_i\|_2 \leqslant \lambda \ , \\ \lambda \frac{z_i}{\|z_i\|_2} & \text{otherwise} \ . \end{cases} \tag{59}$$

[1] This could be enforced as shown in [47].

[2] If $|\Gamma x^\star|_i = 0$ for all $i \in [m]$, the result remains true for any $\alpha > 0$.

Algorithm Non-local means [4] and its directional derivative

Inputs: noisy image y, direction d, noise standard-deviation σ
Parameters: half search window width s, half patch width b, kernel function φ
Outputs: $x^\star = \hat{x}(y)$ and $\tilde{x} = J_{\hat{x}}(y)d$

Initialize $W \leftarrow \varphi(2\sigma^2(2b+1)^2)\, 1_{p_1\times p_2}$ (add weights for the central pixels [38])
Initialize $W_y \leftarrow \varphi(2\sigma^2(2b+1)^2)\, y$ (accumulators for the weighted sum)
Initialize $W' \leftarrow 0_{p_1\times p_2}$
Initialize $W'_y \leftarrow \varphi(2\sigma^2(2b+1)^2)\, d$
for $k \in [-s,s]\times[-s,s]\setminus\{0,0\}$ **do**
  Compute $e \leftarrow [\,(y - S_k(y))^2\,] \star \kappa$ (error between each k shifted patches [12, 13])
  Compute $w \leftarrow \varphi(e) \star \kappa$ (contribution for each patch of its k shift)
  Update $W \leftarrow W + w$ (add weights at each position)
  Update $W_y \leftarrow W_y + wS_k(y)$ (add contribution of each k shifted patches)

  Compute $e' \leftarrow [\,2(y - S_k(y))(d - S_k(d))\,] \star \kappa$
  Compute $w' \leftarrow [\,e'\varphi'(e)\,] \star \kappa$
  Update $W' \leftarrow W' + w'$
  Update $W'_y \leftarrow W'_y + w'S_k(f) + wS_k(d)$
end for
Compute $x^\star \leftarrow W_y/W$ (weighted mean)
Compute $\tilde{x} \leftarrow (W'_y - W'x^\star)/W$

FIG. 6. *Pseudo-algorithm for the computation of the block-wise non-local means and its Jacobian in a direction d. All arithmetic operations are element wise, S_k is the operator that shift all pixels in the direction k, $\star$ is the discrete convolution operator, and $\kappa \in \mathbb{R}^{p_1\times p_2}$ is such that $\kappa_{i,j} = 1$ if $(i,j) \in [-b,b]\times[-b,b]$, 0 otherwise.*

It follows that the algorithmic differentiation strategy reads as

$$(60)\qquad \begin{cases} z^{k+1} &= \Pi^{\mathrm{iso}}_{\mathcal{B}_\lambda}(z^k + \sigma\Gamma v^k)\ , \\ x^{k+1} &= (\mathrm{Id} + \tau\Phi^\top\Phi)^{-1}\left(x^k + \tau(\Phi^\top y - \Gamma^\top(z^{k+1}))\right)\ , \\ v^{k+1} &= x^{k+1} + \theta(x^{k+1} - x^k)\ , \\ \tilde{z}^{k+1} &= \Pi^{\mathrm{iso}}_{z^k+\sigma\Gamma v^k}(\tilde{z}^k + \sigma\Gamma\tilde{v}^k)\ , \\ \tilde{x}^{k+1} &= (\mathrm{Id} + \tau\Phi^\top\Phi)^{-1}\left(\tilde{x}^k + \tau(\Phi^\top y - \Gamma^\top\tilde{z}^{k+1})\right), \\ \tilde{v}^{k+1} &= \tilde{x}^{k+1} + \theta(\tilde{x}^{k+1} - \tilde{x}^k)\ , \end{cases}$$

$$\text{where}\quad \Pi^{\mathrm{iso}}_z(\tilde{z})_i = \begin{cases} \tilde{z}_i & \text{if}\quad \|z_i\|_2 \leqslant \lambda + \beta\ , \\ \frac{\lambda}{\|z_i\|_2}\left(\tilde{z}_i - \left\langle \tilde{z}_i, \frac{z_i}{\|z_i\|_2}\right\rangle \frac{z_i}{\|z_i\|_2}\right) & \text{otherwise}\ . \end{cases}$$

Unlike the ℓ_1 case, the re-fitted solution is not $\tilde{x}^k$ itself, but following Subsection 6.4, can be obtained at the last iteration k as

$$(61)\qquad \mathcal{R}_{x^k}(y) = (1-\rho)x^k + \rho\tilde{x}^k \quad \text{with} \quad \rho = \frac{\langle \Phi(\tilde{x}^k - x^k),\, y - \Phi x^k\rangle}{\|\Phi(\tilde{x}^k - x^k)\|_2^2}\ .$$

6.7. Example for the non-local means. In this section, we specify the update rule (51) to an acceleration of the block-wise non-local means inspired from [12, 13]. We use the procedure of [38] to correctly handle the central pixel. Again, one can check that this implementation can be written in the general form considered in the update rule (50), where the fixed point solution is obtained directly at the first iteration.

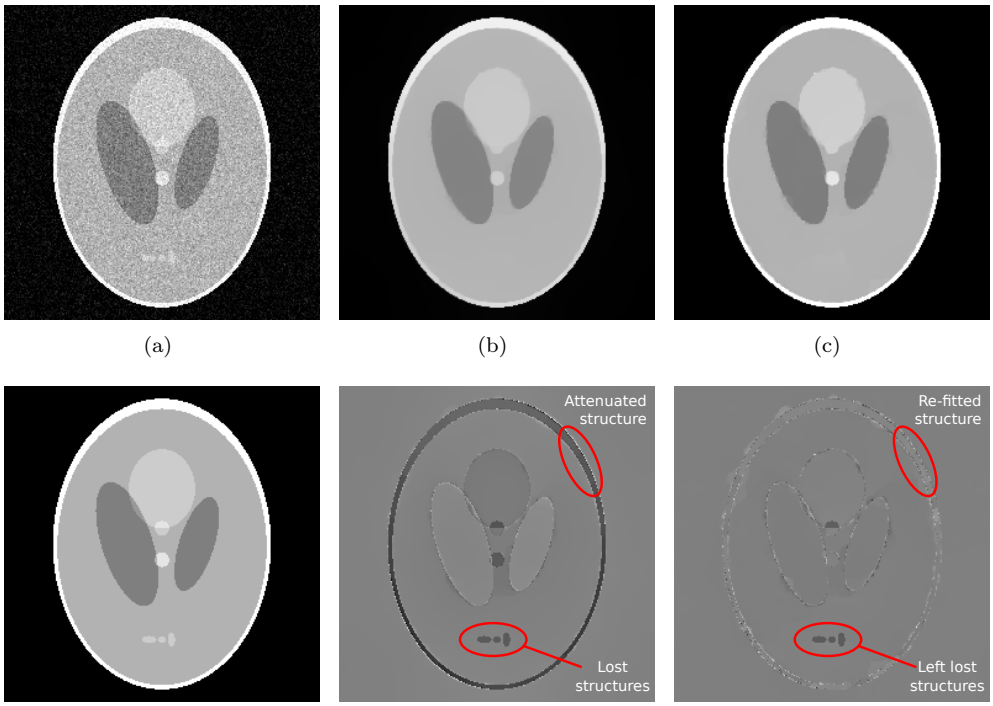


FIG. 7. *(a) Noisy observation $y = x_0 + w$ and (d) noise-free signal x_0. (b) Iso-TV denoising $\hat{x}(y)$ and (e) the residual $\hat{x}(y) - x_0$. (c) Our re-fitting $\mathcal{R}_{\hat{x}}(y)$ and (f) the residual $\mathcal{R}_{\hat{x}}(y) - x_0$.*

The pseudo-code obtained with the algorithmic differentiation scheme, as described in (51) is given in Figure 6. All variables with suffix $'$ correspond to the directional derivative obtained by using the chain rule on the original variables. This fast computation relies on the fact that all convolutions can be computed with integral tables, or in the Fourier domain, leading to a global complexity in $O(s^2 n)$, resp. $O(s^2 n \log n)$, for both the computation of the estimator $\hat{x}(y)$ and its directional derivative Jd. Recall that the covariant re-fitting is obtained from its general form with two steps, by computing first $\hat{x}(y)$, and applying next the proposed pseudo-code in the direction $d = y - \hat{x}(y)$.

7. Numerical experiments and comparisons with related approaches. In this section, we first give illustrations of our CLEAR method on toy image restoration problems. Then, we evaluate the performance of the re-fitting and discuss about its benefit in several scenarios, and we compare our method with related popular approaches from the literature.

7.1. Denoising with isotropic total-variation (iso-TV). Figure 7 gives an illustration of our covariant re-fitting of the 2D iso-TV, where the regularization parameter λ has been chosen large enough in order to highlight the behavior of the re-fitting. We apply it for the denoising (*i.e.,* $\Phi = \mathrm{Id}$) of an approximate $8bits$ piecewise constant image damaged by AWGN with standard deviation $\sigma = 20$, known as

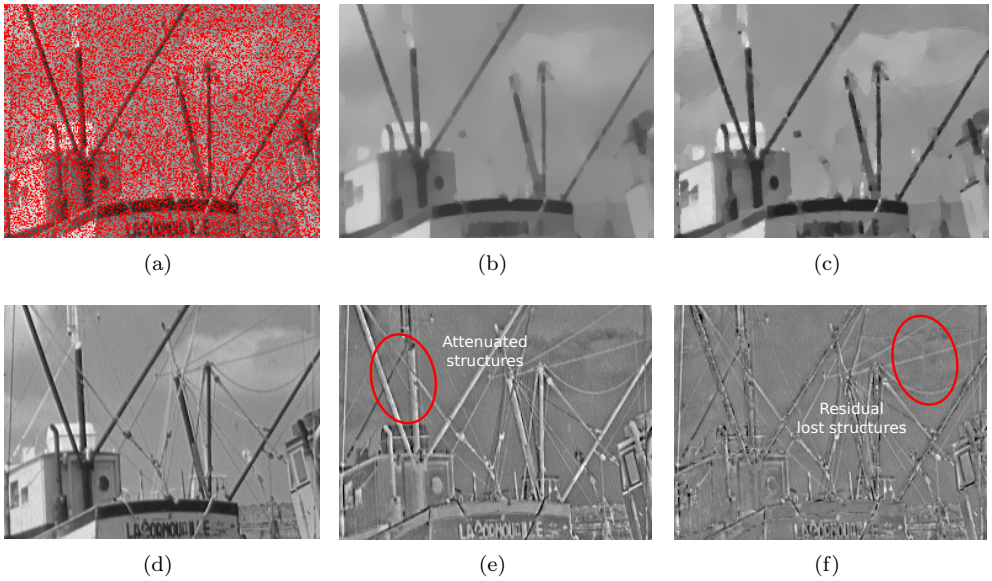


FIG. 8. *(a) Partial and noisy observation* $y = \Phi x_0 + w$ *(red indicates missing pixels) and (d) noise-free signal* x_0*. (b) Iso-TV restoration* $\hat{x}(y)$ *and (e) the residual* $\hat{x}(y) - x_0$*. (c) Our re-fitting* $\mathcal{R}_{\hat{x}}(y)$ *and (f) the residual* $\mathcal{R}_{\hat{x}}(y) - x_0$*.*

the *Shepp-Logan phantom*. As discussed in Subsection 3.6, iso-TV introduces a significant loss of contrast [41], typically for thin detailed structures, which are re-enhanced in our result.

The residuals $\hat{x}(y) - x_0$ and $\mathcal{R}_{\hat{x}}(y) - x_0$ highlight that our re-fitting technique re-enhances efficiently the attenuated structure while it leaves the lost structures unchanged. Nevertheless, after re-fitting, some small residuals around the edges appear. In fact, in the vicinity of edges, iso-TV finds (barely visible) discontinuities that are not in accordance with the underlying image. This creates an overload of small constant regions. When re-fitting is performed, all such regions are re-fitted to the noisy data, and such regions becomes barely visible artifacts. In other words, the re-fitting has re-enforced the presence of a modeling problem, resulting to an increase of residual variance, that iso-TV had originally compensated by attenuating the amplitudes.

7.2. De-masking with isotropic total-variation (iso-TV). Figure 8 gives another illustration of our covariant re-fitting of the 2D iso-TV used for the restoration of an approximate 8*bits* piece-wise constant image damaged by AWGN with standard deviation $\sigma = 20$, known as *Boat*. The observation operator Φ is chosen as a random mask removing 40% of pixels. Again, iso-TV introduces a significant loss of contrast, typically for thin details such as the contours of the objects, which are re-enhanced in our re-fitting result.

Inspecting the map of residuals in Figure 8.(e)-(f) illustrates again that our re-fitting technique eliminates most of the bias to the price of a small variance increase. This becomes clear by looking at the mast and the ropes of the Boat. While the mast was preserved by iso-TV and re-enhanced by our re-fitting, the ropes remains lost for both the iso-TV and our re-fitting.

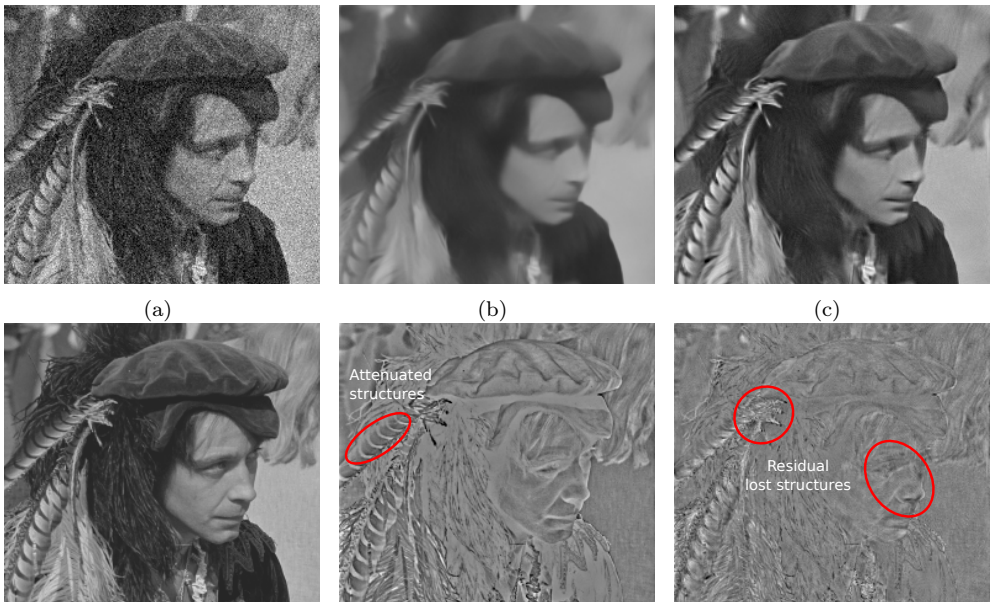


FIG. 9. *(a) Noisy observation $y = x_0 + w$ and (d) noise-free signal x_0. (b) Non-local means $\hat{x}(y)$ and (e) the residual $\hat{x}(y) - x_0$. (c) Our re-fitting $\mathcal{R}_{\hat{x}}(y)$ and (f) the residual $\mathcal{R}_{\hat{x}}(y) - x_0$.*

7.3. Denoising with non-local means. Figure 9 gives another illustration of our re-fitting procedure for the block-wise non-local means algorithm used in a denoising problem of the 8*bits* image *Pirate*, enjoying many repetitive patterns and damaged by AWGN with standard deviation $\sigma = 20$. Again, we choose a regularizing kernel $\varphi(\cdot) = \exp(\cdot/h), h > 0$ that leads to strong smoothing in order to highlight the behavior of the re-fitting. Our re-fitting technique provides again favorable results: many details are enhanced compared to the dull standard version. This reveals that the standard non-local means is actually able to well capture the repetitions of many patterns but this information is not used properly to create a satisfying result. The re-fitting produces a sharper result, by enforcing the correct use of all the structures identified by patch comparisons.

The maps of residuals Figure 9.(e)-(f) highlight that our re-fitting technique suppresses efficiently this dull effect while it preserves the model originally captured by patch redundancy. Again the suppression of this phenomenon is counter balanced by an increase of the residual variance prominent where the local patch redundancy assumption is violated.

In these examples, the overall residual norm is clearly reduced by the re-fitting because the amount of reduced bias surpasses the increased of residual variance. This favorable behavior depends on the internal parameters of the original estimator acting on the bias-variance trade-off. This is the subject of the next section.

7.4. A bias-variance analysis for the covariant re-fitting. Previous experiments have revealed that while CLEAR tends to reduce the bias, it increases (as expected) the residual variance. It is therefore important to understand under which

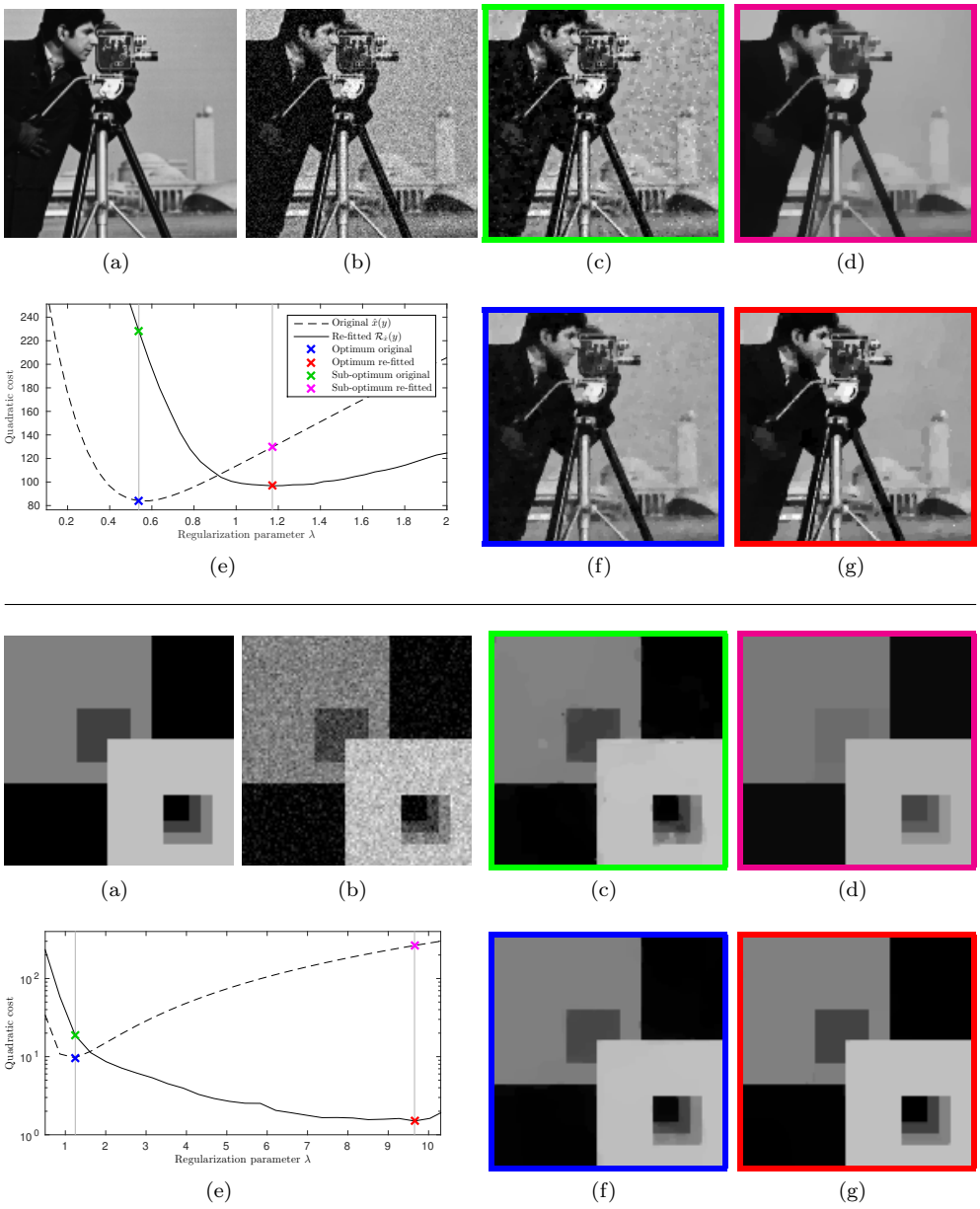


FIG. 10. *Experiment with aniso-TV: (top) poorly approximate piece-wise constant case. (bottom) pure piece-wise constant case. (a) Noise-free image x_0. (b) Noisy image $y = x_0 + w$. (e) Mean square error of the aniso-TV estimator $\hat{x}(y)$ and its re-fitted version $\mathcal{R}_{\hat{x}}(y)$ with respect to the parameter λ. Two values of λ are selected corresponding to (c) re-fitted version for a sub-optimal λ value, (d) original version for a sub-optimal λ value, (f) original version for its optimal λ value, (g) re-fitted version for its optimal λ value.*

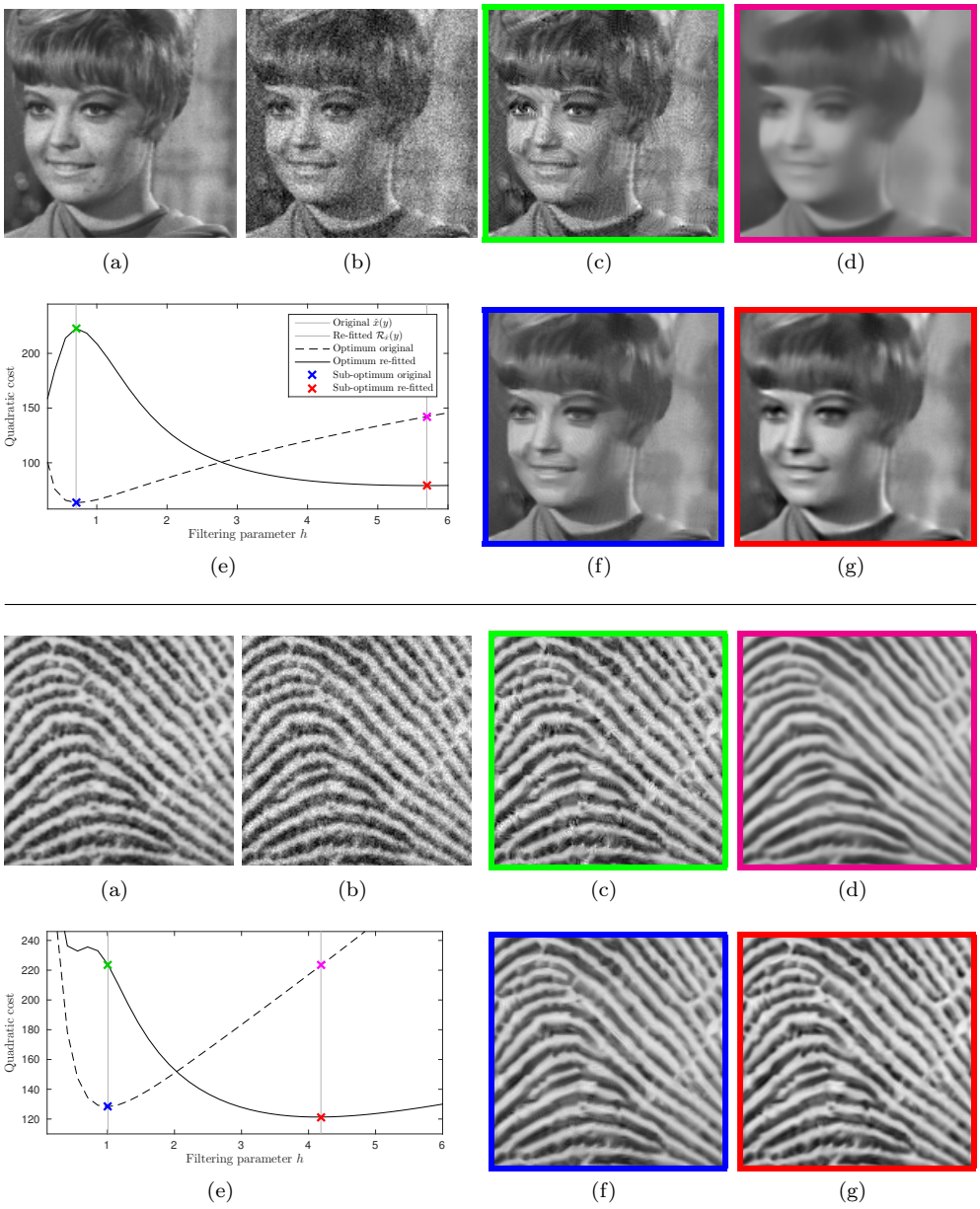


FIG. 11. *Experiment with block-wise non-local means: (top) moderate patch redundancy case. (bottom) high patch redundancy case. (a) Noise-free image x_0. (b) Noisy image $y = x_0 + w$. (e) Mean square error of the estimator $\hat{x}(y)$ and its re-fitted version $\mathcal{R}_{\hat{x}}(y)$ with respect to the parameter h. Two values of h are selected corresponding to (c) re-fitted version for a sub-optimal h value, (d) original version for a sub-optimal h value, (f) original version for its optimal h value, (g) re-fitted version for its optimal h value.*

conditions the bias-variance trade-off is in favor of our re-fitting technique.

Figure 10 illustrates the evolution of performance, measured in terms of mean squared error (MSE), of both the aniso-TV and its re-fitting version as a function of the regularization parameter λ. Two images are considered: a crude approximation of a piece-wise constant image (the *Cameraman* in the top), and an actual piece-wise constant image (in the bottom).

This experiment highlights that optimal results for both approaches are not reached at the same λ value. Visual inspection of the optima seems to demonstrate that due to the bias, the optimal solution of aniso-TV is reached for a λ value that promotes a model subspace that is not in accordance with the underlying signal: typically the presence of an overload of (barely visible) transitions in homogeneous areas. These transitions become clear when looking at the re-fitted version where each small region is re-fitted on the noisy data, revealing an excessive residual variance. Conversely, the optimal λ value for the re-fitting seems to retrieve the correct model, *i.e.,* with transitions that are closely in accordance with the underlying signal. Moreover, comparing their relative performance, when both are used at their own optimal λ value, reveals that our re-fitting brings a significant improvement if the underlying image is in fact piece-wise constant.

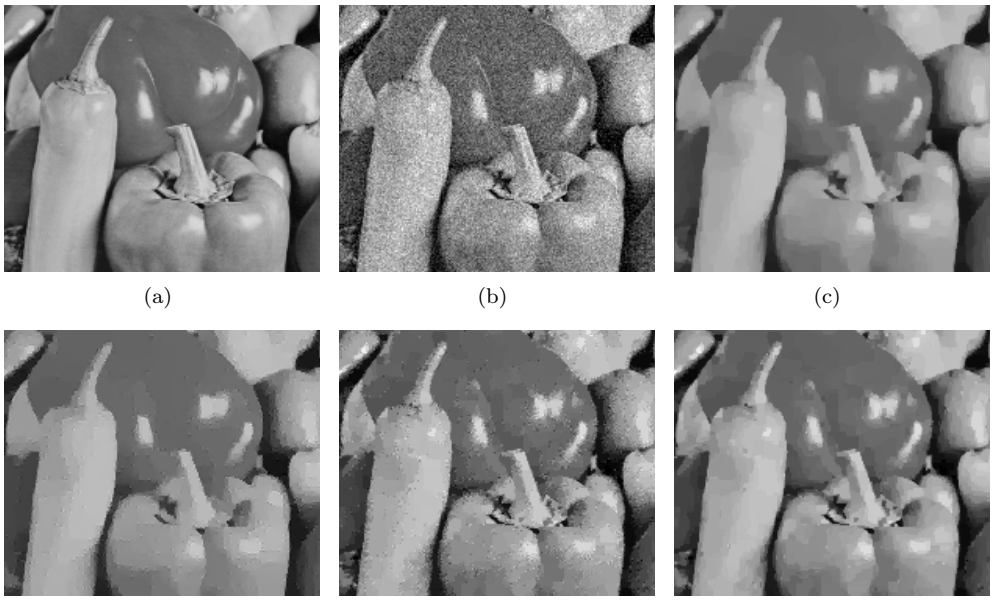
Figure 11 provides a similar illustration of the evolution of performance for the block-wise non-local means and its re-fitted version as a function of the smoothing parameter h of the kernel function $\varphi(\cdot) = \exp(\cdot/h)$. Two images are considered: a crude approximation of an image with redundant patterns (the *Lady* in the top part of the figure), and an image with many redundant patterns (the *Fingerprint* in the bottom part of the figure). Similar conclusions can be made from this experiment. In particular, comparing their relative performance, when both are used at their own optimal h value, seems to demonstrate that the re-fitting brings an improvement when most of patches of the underlying image are redundant.

While it is difficult to make a general statement, we can reasonably claim from these experiments that the re-fitting is all the more relevant in terms of MSE than the underlying image x_0 is in accordance to the retrieved subspace model $\mathcal{M}_{\hat{x}}(y)$. In other words, re-fitting is relatively safe when the original restoration technique was chosen wisely with respect to the underlying image of interest. Beyond the MSE performance, visual inspection of the re-fitted results at their optimal parameters choice might nevertheless be preferable. Even tough the MSE is not necessarily improved, intensities and contrasts are recovered better.

7.5. Comparisons with other techniques devoted to the ℓ_1 case. We detail hereafter two different alternative strategies devoted to re-enhance the solution of the ℓ_1 analysis regularization.

Iterative hard-thresholding. As shown earlier, the hard-thresholding is the re-fitted version of the soft-thresholding. Given an iterative solver $(k, y) \mapsto x^k$ composed of linear and soft-thresholding (such as the primal-dual algorithm), one could consider replacing all soft-thresholding by hard-thresholding while keeping linearities unchanged: a technique often referred to as "iterative hard-thresholding" [3]. Unfortunately, the convergence of such techniques is not ensured in general (typically when $\Phi \neq \mathrm{Id}$), and even if so, they usually do not converge to the sought re-fitting $\mathcal{R}_{\hat{x}}(y)$.

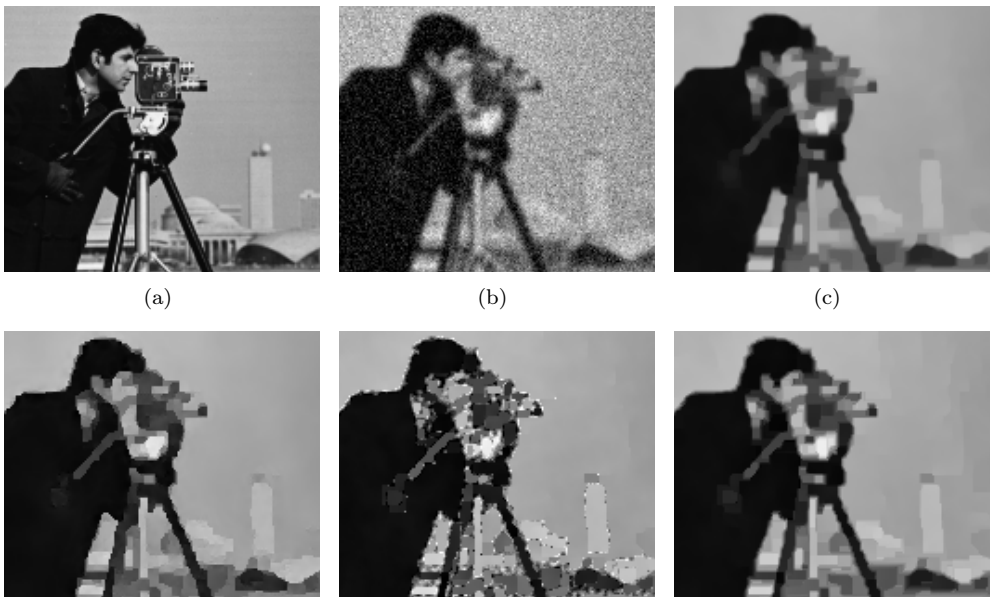
Co-support identification based post re-fitting. Another solution referred to as post re-fitting, and studied in, *e.g.,* [20, 35, 2, 27, 2], consists in identifying the (co-)support $\mathcal{I} = \{i \ : \ (\Gamma\hat{x}(y))_i \neq 0\}$ and solving a least-square problem constrained



(a) (b) (c)

(d) (e) (f)

FIG. 12. *(a) Noise-free image x_0. (b) Noisy image $y = x_0 + w$. (c) Original aniso-TV estimate $\hat{x}(y)$. Enhanced results by (d) iterative hard-thresholding, (e) post re-fitting with support identification, and (f) our proposed covariant re-fitting.*



(a) (b) (c)

(d) (e) (f)

FIG. 13. *(a) Original image x_0. (b) Blurred and noisy image $y = \Phi x_0 + w$. (c) Original aniso-TV estimate $\hat{x}(y)$. Enhanced results by (d) iterative hard-thresholding, (e) post re-fitting with support identification, and (f) our proposed covariant re-fitting.*

to $\{x : (\Gamma x)_{\mathcal{I}^c} = 0\}$ typically with a conjugate gradient. However, $\hat{x}(y)$ is usually obtained thanks to a converging sequence x^k, and unfortunately, $\mathrm{supp}(\Gamma x^k)$ can be far from $\mathrm{supp}(\Gamma \hat{x}(y))$ even though x^k can be made arbitrarily close to $\hat{x}(y)$. Such erroneous support identifications can lead to results that strongly deviates from the solution $\mathcal{R}_{\hat{x}}(y)$.

Figure 12 provides a comparison of our re-fitting method with the two other approaches mentioned earlier for the aniso-TV used on a 8*bits* image damaged by AWGN with standard deviation $\sigma = 20$, known as *peppers*. The iterative hard-thresholding approach does not preserve the model space: transitions are not localized at the same positions as in the original solution and suspicious oscillations are created. The post re-fitting and our approach have both improved $\hat{x}(y)$ by enhancing each piece and preserving the location of transitions. Our method is nevertheless more stable than support identification which produces many errors due to wrong co-support identification.

Figure 13 provides another illustration highlighting the problem of support identification in an ill-posed problem. It consists of an 8bits image damaged by a Gaussian blur of 2px and AWGN with standard deviation $\sigma = 20$, known as *Cameraman*. Again, while aniso-TV reduces the contrast, the re-fitting recovers the original amplitudes and keep unchanged the discontinuities. Post re-fitting offers comparable results to ours except for suspicious oscillations due to wrong co-support identification.

In contrast with the support identification, CLEAR does not require the identification of the co-support, nor the identification of the model subspace $\mathcal{M}_{\hat{x}}(y)$. This is an appealing property since the co-support of the optimal solution $\hat{x}(y)$ is difficult to identify, particularly in the analysis context. Being computed during the iterations of the original algorithm, jointly with $\hat{x}(y)$, our re-fitting strategy also provides more stable solutions.

7.6. Comparisons with boosting strategies. We detail hereafter other popular alternatives designed to re-enhance results of an arbitrary estimator.

Twicing and boosting. The boosting iterations introduced in [5] is a simple approach that consists in re-injecting to the current solution $\tilde{x}^k$ a filtered version of its residual $y - \Phi\tilde{x}^k$. The idea is that if parts of the signal where lost at iteration k, they might be retrieved in the residual. Given $\tilde{x}^0 = 0$, the iterations reads

$$\tilde{x}^{k+1} = \tilde{x}^k + \hat{x}(y - \Phi\tilde{x}^k) \ . \tag{62}$$

The first iterate is $\tilde{x}^1 = \hat{x}(y)$, and $\tilde{x}^2$ is known as the twicing estimate [46]. When k increases, the bias of this estimator tends to decreases while its variance increases, see for instance [42]. In denoising (*i.e.,* when $\Phi = \mathrm{Id}$) with a linear estimator $\hat{x}(y) = Wy$ (*e.g.,* the Tikhonov regularization), we get $\tilde{x}^k = (\mathrm{Id} - (\mathrm{Id} - W)^k)y$, for $k > 0$. In particular, the twicing reads as $\tilde{x}^2 = (2W - W^2)y$. Recall that the covariant re-fitting reads in this case as $\mathcal{R}_{\hat{x}}(y) = (W + \rho W - \rho W^2)y$ and it coincides when $\rho = 1$. Such an approach is also popular for instance in kernel smoothing in non-parametric statistics [11].

Iterative Bregman refinement. In [32], the authors proposed an iterative procedure, originally designed to improve iso-TV results, given by

$$\tilde{x}^{k+1} = \hat{x}\left(y + \sum_{i=1}^{k}(y - \Phi\tilde{x}^k)\right) \ . \tag{63}$$

Unlike boosting that iteratively filters the residual, the idea is to filter a modified version of the input y amplified by adding the sum of the residuals. When $\Phi = \mathrm{Id}$ and $\hat{x}(y) = Wy$, the iterative Bregman refinement reads as $\tilde{x}^k = (\mathrm{Id} - (\mathrm{Id} - W)^k)y$, and it coincides with the boosting approach. This has led to several related approaches, and we refer the interested reader to [6, 50, 24, 33] for more details.

SOS-boosting. In [36], the authors follows a similar idea by iteratively filtering a strengthened version of the input y. Their method, named Strengthen Operate Substract boosting (SOS-boosting) performs iteratively the following update

$$\tilde{x}^{k+1} = \tau\hat{x}(y + \alpha\Phi\tilde{x}^k) - (\tau\alpha + \tau - 1)\tilde{x}^k \ . \tag{64}$$

where α and τ are two real parameters. The first one controls the emphasis of the solution (and the convergence of the sequence), while the second one controls the rate of convergence. When $\Phi = \mathrm{Id}$ and $\hat{x}(y) = Wy$, the SOS refinement with $\tau = 1$ reads as $\tilde{x}^k = Wy + \alpha(W - \mathrm{Id})\tilde{x}^{k-1}$, and in particular, for $k = 2$, we get $\tilde{x}^2 = (W - \alpha W + \alpha W^2)y$ which coincides with our covariant re-fitting for the choice $\alpha = -\rho$. Note that for all the estimators we have considered, we have always observed that $\rho > 0$, contrarily to [36], where $\alpha > 0$ is implicitly assumed. Hence, we cannot concludes that the two models match in a specific setting. Another difference, is that while we provide an automatic way to compute ρ, see for instance Equation (54), the α parameter of the SOS-boosting must be tuned by the practitioner, which my be cumbersome for the practitioner (*e.g.,* when using cross-validation on a specific dataset of images and/or for specified noise levels).

SAIF-boosting. As described in [29], the diffusion of a filter consists in iteratively re-applying the filter to the current estimate $\tilde{x}^{k+1} = \hat{x}\left(\tilde{x}^k\right)$. The authors of [42] noticed that, unlike the boosting method of [5], the bias of this estimator increases and its variance decreases with k. As a consequence, the authors suggest mixing the two approaches by deciding at each iteration between performing a boosting or a diffusion step. To that end, they proposed a plug-in risk estimator that crudely estimates the MSE from a pre-filtered version of y. This approach is in fact applied locally on image patches, and is referred to as the Spatially Adapted Iterative Filter (SAIF)-boosting. Unlike the other techniques, the SAIF-boosting cannot be used as a black-box. Indeed, it requires to perform an eigen decomposition of $\hat{x}$ locally for each patch of y. This can be efficiently done for some kernel-based averaging filters, but can be very challenging for arbitrary estimators, such as for instance iso-TV.

It is worth mentioning that boosting approaches are scarcely used for linear estimators. The successive re-application of a non-linear estimator $\hat{x}$ allows to recover parts of the signal that were lost at former iterations. Nevertheless, to boost the solution, the internal parameters of $\hat{x}$ often need to be re-adapted at each iteration, leading to cumbersome tuning of parameters in practice, even if this can be done using cross-validation on a specific set of images for each targeted noise level investigated. Unlike boosting methods, a re-fitting approach should not modify the regularity and the structure of the first estimate. This is the reason why CLEAR only considers the linearization of $\hat{x}$ at y through the Jacobian.

It is important to have in mind that theoretical results of the aforementioned methods are well grounded in the case where, even though $\hat{x}$ is non-linear, it acts locally as an averaging filter. In other words, locally, there exists a row stochastic linear operator W, *i.e.,* $W1_n = 1_n$, (or even bi-stochastic, symmetric or independent of y) such that $\hat{x}(y) = Wy$. Such theoretical results could not be applied to the

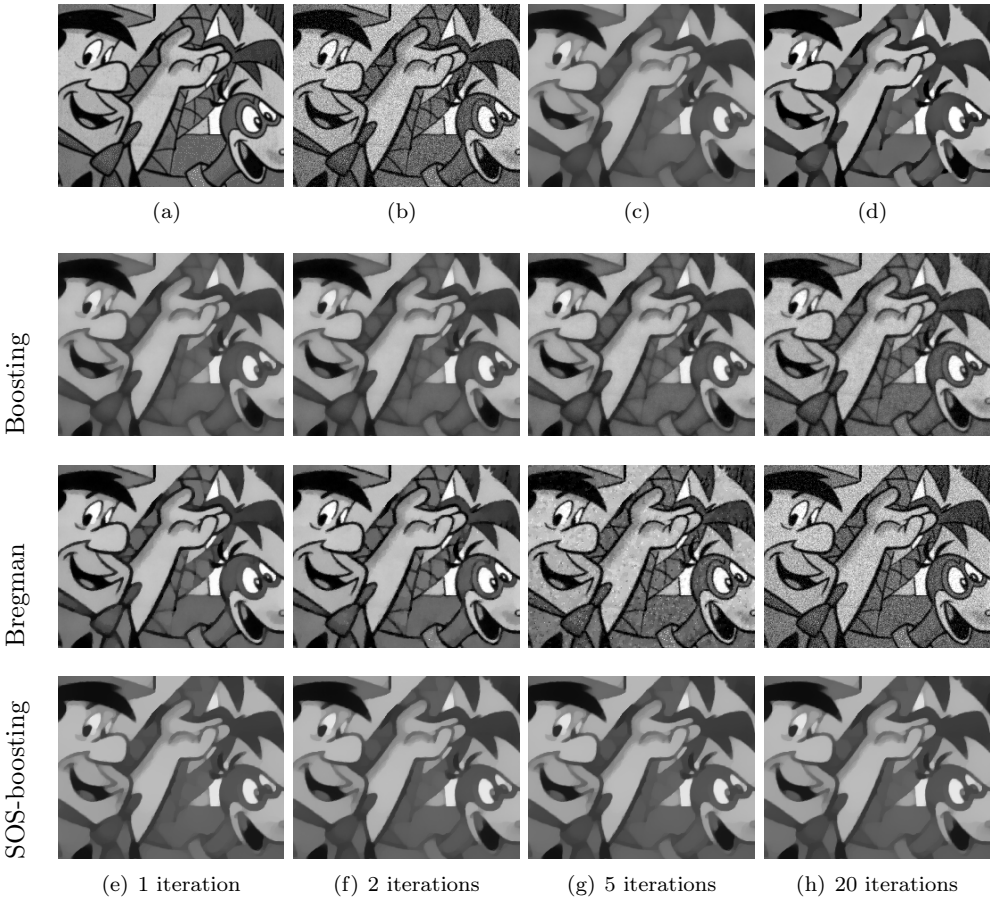


FIG. 14. *(a) A cartoon image x_0 with large constant pieces. (b) Its noisy version $y = x_0 + w$. (c) Result $\hat{x}(y)$ of iso-TV. (d) Our covariant re-fitting $\mathcal{R}_{\hat{x}}(y)$. (e,f,g,h) From top to bottom, results of boosting [46, 5], Bregman iterations [6] and SOS-boosting [36] at respectively* 1, 2, 5 *and* 20 *iterations.*

soft-thresholding (nor to more advanced methods we have considered). Indeed, from Subsection 3.3, for the soft-thresholding, a natural candidate for W is the diagonal matrix defined as

$$W_{ii} = \begin{cases} 1 - \frac{\lambda}{|y_i|} & \text{if} \quad |y_i| > \lambda \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{65}$$

which is not row stochastic. In this context, the matrix W is not the Jacobian, which is given in this case by the diagonal matrix

$$J_{ii} = \begin{cases} 1 & \text{if} \quad |y_i| > \lambda \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{66}$$

which, unlike W, is row stochastic and locally a projector. A second limitation is that even though W is row stochastic, it might still encode a bias part. Typically, for the ℓ_1 analysis described in Subsection 3.5, a natural candidate for W is

$$W = U(\Phi U)^+ - U(U^\top \Phi^\top \Phi U)^{-1} U^\top (\Gamma^\top)_{\mathcal{I}} R \ , \tag{67}$$

FIG. 15. *(a) A cartoon image x_0 with large constant pieces. (b) Its noisy version $y = x_0 + w$. (c) Result $\hat{x}(y)$ of non-local means. (d) Our covariant re-fitting $\mathcal{R}_{\hat{x}}(y)$. (e,f,g,h) From top to bottom, results of boosting [46, 5], Bregman iterations [6] and SOS-boosting [36] at respectively* 1, 2, 5 *and* 20 *iterations.*

where R is a diagonal matrix with diagonal elements $R_{ii} = \lambda/|y_i|$ for $i \in \mathcal{I}$ and 0 otherwise. One can check that if $1_n \in \mathrm{Ker}[\Gamma]$ (which holds true for aniso-TV), then W is row stochastic. However, as seen in Subsection 3.5, the quantity $U(U^\top \Phi^\top \Phi U)^{-1} U^\top (\Gamma^\top)_{\mathcal{I}} R$ is the term responsible for the systematic contraction of the ℓ_1 analysis regularization (this simplifies to $\lambda/|y_i|$ for the soft-thresholding). As a consequence, the bias cannot be corrected by a single application of W. The Jacobian $J = U(\Phi U)^+$, which is again row stochastic, is free of this contraction term. Therefore our covariant re-fitting gets rid of this bias term after one single application of the Jacobian J.

Figure 14 and Figure 15 provide a comparison of our covariant re-fitting with the three first aforementioned boosting approaches, on two 8-bits images (*Flinstones* and *Barbara*) damaged by AWGN with standard deviation $\sigma = 20$, respectively for iso-TV denoising and for non-local means. Note that the α and τ parameters of the SOS-boosting approach have been tuned to offer the most satisfying results, even though, we did not observe a significant impact in the iso-TV case. As expected,

our covariant re-fitting provides results re-enhanced towards the amplitudes of the noisy inputs. While our re-fitting preserves the structural content and smoothness of the original dull solution, the boosting approaches re-inject structural contents that were not originally preserved. Sometimes they do not even re-fit towards the original amplitudes, or they present a large amount of residual noise. Note that in these experiments, the original estimator was chosen to be significantly biased. We believe that unlike re-fitting methods, boosting techniques might be more relevant to improve the quality of near unbiased estimates.

8. Conclusion. We have introduced a generalization of the popular least-square re-fitting technique, originally introduced to reduce the systematic contraction of the Lasso estimator.

Together with this generalization, a generic implementation has been given for a wide class of ill-posed linear problem solvers. This implementation requires a computational overload of at most a factor of about three compared to the original solver; this factor can even be reduced to about two for most popular estimators in image processing.

While the classical re-fitting is inspired by the standard Lasso debiasing step (*i.e.,* least-square re-fitting on the estimated support), our generalization leverages the Jacobian of the estimator and does not rely on the notion of support. In particular, the proposed implementation only requires chain rules and differentiating the considered solver. It is free from the unstable step of identifying the underlying support. In practice, this has the benefit of increasing the stability compared to classical implementations.

For estimators such as Tikhonov regularization, total-variation or non-local means, numerical experiments have demonstrated the efficiency of the CLEAR technique in retrieving correct intensities while respecting the structure of the original biased estimator. Moreover, it has been shown in practice that re-fitting is beneficial when the underlying signal structure is well captured by the original estimator. Otherwise, re-fitting leads to too simplistic approximations, typically reflecting an inaccurate prior model. In other words, if the considered estimator is adequate with respect to the application context, then re-fitting is recommended.

We have highlighted the importance in distinguishing boosting approaches from the re-fitting one. In particular, re-fitting should be preferred in applications where the content of the original solution must be preserved. While boosting approaches are mostly used to enhance near unbiased estimators (typically coming from combinatorial or non-convex problems), the re-fitting is all the more relevant for estimators that present biases. For instance, re-fitting is essential for estimators solution of a convex problem that require a large bias correction to accurately retrieve the content of the signal of interest, a canonical example being the isotropic total-variation (iso-TV). Nevertheless, we believe that the notion of Jacobian based re-fitting could be of interest for boosting applications and we leave this to future work.

8.1. Acknowledgment. The authors would like to acknowledge the financial support of the GDR 720 ISIS (Information, Signal, Image et viSion). This study has also been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the "Investments for the future" Programme IdEx Bordeaux - CPU (ANR-10- IDEX-03-02).

Appendix A. Sketch of proofs. This section details how to retrieve closed-form expressions of some of the estimators studied in the paper, that are either well

known results or easy to derive.

A.1. Retrieving the Least-square solution. We aim at retrieving here the solution of the constrained least-square problem

$$\operatorname*{argmin}_{x\in\mathbb{R}^N} \|\Phi x - y\|^2 + \iota_C(x) \tag{68}$$

where $C = b + \mathrm{Im}[A]$, $b \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times n}$. Note that the initial problem can be recast as

$$\operatorname*{argmin}_{x\in C} \|\Phi x - y\|^2 = b + A \cdot \operatorname*{argmin}_{t\in\mathbb{R}^n} \underbrace{\|\Phi A t - (y - \Phi b)\|^2}_{F(t,y)} \ . \tag{69}$$

The problem being differentiable and convex, a minimum can be obtained by studying its first optimality conditions given by

$$\frac{\partial F(t,y)}{\partial t} = 0 \Leftrightarrow A^\top \Phi^\top \Phi A t = A^\top \Phi^\top (y - \Phi b) \ . \tag{70}$$

In particular $t = (\Phi A)^+(y - \Phi b)$ is a solution, and hence $x = b + At$, i.e., $x = b + A(\Phi A)^+(f - \Phi b)$ is a solution.

A.2. Retrieving the Tikhonov solution. We consider the optimization problem defined, for $\Gamma \in \mathbb{R}^{m\times p}$ and $\lambda > 0$, as

$$\operatorname*{argmin}_{x\in\mathbb{R}^N} \frac{1}{2}\|\Phi x - y\|^2 + \frac{\lambda}{2}\|\Gamma x\|^2 \ . \tag{71}$$

The objective function being again differentiable and convex, a minimum can be obtained by studying the first order optimality conditions given by

$$\Phi^\top(\Phi x - y) + \lambda\Gamma^\top\Gamma x = 0 \Leftrightarrow (\Phi^\top\Phi + \lambda\Gamma^\top\Gamma)x = \Phi^\top f \ . \tag{72}$$

Provided $\operatorname{Ker}\Phi \cap \operatorname{Ker}\Gamma = \{0\}$ and $\lambda > 0$, the quantity $\Phi^\top\Phi + \lambda\Gamma^\top\Gamma$ is invertible and $x = (\Phi^\top\Phi + \lambda\Gamma^\top\Gamma)^{-1}\Phi^\top y$ is the unique solution of (71).

A.3. Retrieving the hard-thresholding solution. We consider the minimization problem defined, for $\lambda > 0$, as

$$\operatorname*{argmin}_{x\in\mathbb{R}^N} \left\{ E(x,y) = \frac{1}{2}\|x - y\|^2 + \frac{\lambda^2}{2}\|x\|_0 \right\}. \tag{73}$$

The problem is separable meaning that $\left[\operatorname*{argmin}_{x\in\mathbb{R}^N} E(x,y)\right]_i = \operatorname*{argmin}_{x_i\in\mathbb{R}} E_i(x_i,y_i)$ with

$$E_i(x_i,y_i) = \frac{1}{2}\begin{cases} y_i^2 & \text{if} \quad x_i = 0 \ , \\ (x_i - y_i)^2 + \lambda^2 & \text{otherwise} \ . \end{cases} \tag{74}$$

Since $y_i^2 \leqslant \min_{x_i}[(x_i - y_i)^2 + \lambda^2] \Leftrightarrow |y_i| \leqslant \lambda$, we get

$$\min_{x_i} E_i(x_i,y_i) = \frac{1}{2}\begin{cases} y_i^2 & \text{if} \quad |y_i| \leqslant \lambda \ , \\ \lambda^2 & \text{otherwise} \ , \end{cases} \tag{75}$$

which is reached by setting $x_i = 0$ when $|y_i| \leqslant \lambda$ and $x_i = y_i$ otherwise.

A.4. Retrieving the soft-thresholding solution. We consider the minimization problem defined, for $\lambda > 0$, as

$$\operatorname*{argmin}_{x \in \mathbb{R}^N} \left\{ E(x,y) = \frac{1}{2}\|x-y\|^2 + \lambda \|x\|_1 \right\} , \tag{76}$$

which is, as for the hard-thresholding, separable with

$$E_i(x_i,y_i) = \frac{1}{2}(x_i-y_i)^2 + \lambda |x_i| \ . \tag{77}$$

The problem being convex, a minimum is reached when zero belongs to its subdifferential given by

$$\partial E_i(x_i,y_i) = x_i - y_i + \lambda \begin{cases} \operatorname{sign}(x_i) & \text{if} \quad |x_i| > 0 \ , \\ [-1,1] & \text{otherwise} \ . \end{cases} \tag{78}$$

Hence, x_i is solution if

$$0 \in \partial E_i(x_i,y_i) \Leftrightarrow x_i \in y_i - \lambda \begin{cases} \operatorname{sign}(x_i) & \text{if} \quad |x_i| > 0 \ , \\ [-1,1] & \text{otherwise} \ , \end{cases} \tag{79}$$

which holds true by setting $x_i = 0$ when $|y_i| \leqslant \lambda$ and $x_i = y_i - \lambda \operatorname{sign}(y_i)$ otherwise.

A.5. Retrieving the non-local means. The periodical boundary conditions lead to the following relationship, where $l \in [-b,b] \times [-b,b]$, given by

$$\begin{aligned} F(x,y) &= \frac{1}{2}\sum_{i,j} w_{i,j} \|\mathcal{P}_i x - \mathcal{P}_j y\|^2 = \frac{1}{2}\sum_{i,j} w_{i,j} \sum_l (x_{i+l} - y_{j+l})^2 \\ &= \frac{1}{2}\sum_{i,j} \left[\sum_l w_{i,j} (x_{i+l} - y_{j+l})^2 \right] = \frac{1}{2}\sum_{i,j} \left[\sum_l w_{i-l,j-l} (x_i - y_j)^2 \right] \\ &= \frac{1}{2}\sum_{i,j} \left[\sum_l w_{i-l,j-l} \right] (x_i - y_j)^2 = \frac{1}{2}\sum_{i,j} \bar{w}_{i,j} (x_i - y_j)^2 \ . \end{aligned} \tag{80}$$

For all $i \in [p_1] \times [p_2]$, studying the first optimality conditions gives

$$\frac{\partial F(x,y)}{\partial x_i} = 0 \Leftrightarrow \sum_j \bar{w}_{i,j}(x_i - y_j) = 0 \Leftrightarrow x_i = \frac{\sum_j \bar{w}_{i,j} y_j}{\sum_j \bar{w}_{i,j}} \ . \tag{81}$$

Appendix B. Proof of Theorem 28.

Before turning to the proof of this theorem, let us introduce a first lemma.

LEMMA 37. *For all y, let $\hat{x}(y)$ be a solution of*

$$\hat{x}(y) \in \operatorname*{argmin}_x F(y - \Phi x) + G(x) \ , \tag{82}$$

with F, G two convex functions and G being 1-homogeneous. Then for all $\varepsilon \in [0,1]$, the following holds

$$(1-\varepsilon)\hat{x}(y) \in \operatorname*{argmin}_x F(y - \varepsilon \Phi \hat{x}(y) - \Phi x) + G(x) \ . \tag{83}$$

Proof. Note that if G is a convex and $1-$homogeneous function, then G is sub-additive, *i.e.,* $G(a)+G(b) \geqslant G(a+b)$. Next, assume that, for some $\varepsilon \in [0,1]$, Eq. (83) does not hold, so that there exists v such that

$$F(y-\varepsilon\Phi\hat{x}(y)-\Phi v)+G(v) < F(y-\varepsilon\Phi\hat{x}(y)-(1-\varepsilon)\Phi\hat{x}(y))+G((1-\varepsilon)\hat{x}(y)) \;, \\ < F(y-\Phi\hat{x}(y))+(1-\varepsilon)G(\hat{x}(y)) \;. \tag{84}$$

It follows that

$$F(y-\varepsilon\Phi\hat{x}(y)-\Phi v)+G(v)+\varepsilon G(\hat{x}(y)) < F(y-\hat{x}(y))+G(\hat{x}(y)) \;. \tag{85}$$

We also have $G(v)+\varepsilon G(\hat{x}(y)) = G(v)+G(\varepsilon\hat{x}(y)) \geqslant G(v+\varepsilon\hat{x}(y))$, since G is 1-homogeneous and sub-additive. Hence, for $w = v+\varepsilon\hat{x}(y)$, we get

$$F(y-\Phi w)+G(w) < F(y-\Phi\hat{x}(y))+G(\hat{x}(y)) \;, \tag{86}$$

which is in contradiction with $\hat{x}(y) \in \operatorname{argmin} F(y-\Phi x)+G(x)$, and then concludes the proof of this lemma. ∎

Proof of Theorem 28. By virtue of Lemma 37 and definition of $\hat{x}(y)$, we have $(1-\varepsilon)\hat{x}(y) = \hat{x}(y-\varepsilon\Phi\hat{x}(y))$ since $\hat{x}(y)$ is supposed to be the unique solution for all y. Now, recall that the linear Jacobian operator applied to $\Phi\hat{x}(y)$ is the directional derivative of $\hat{x}(y)$ in the direction $\Phi\hat{x}(y)$, then, for almost all y, we get

$$J\Phi\hat{x}(y) \triangleq \lim_{\varepsilon\to 0} \frac{\hat{x}(y)-\hat{x}(y-\varepsilon\Phi\hat{x}(y))}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{\hat{x}(y)-(1-\varepsilon)\hat{x}(y)}{\varepsilon} = \hat{x}(y) \;, \tag{87}$$

which concludes the proof. ∎

Appendix C. Proof of Theorem 36.

Before turning to the proof of this theorem, let us introduce a first lemma.

LEMMA 38. *The re-fitting $\mathcal{R}_{\hat{x}}(y)$ of the ℓ_1 analysis regularization $x^\star = \hat{x}(y)$ is the solution of the saddle-point problem*

$$\min_{\tilde{x}\in\mathbb{R}^p} \max_{\tilde{z}\in\mathbb{R}^m} \|\Phi\tilde{x}-y\|^2 + \langle \Gamma\tilde{x}, \tilde{z}\rangle - \iota_{S_{\mathcal{I}}}(\tilde{z}) \;, \tag{88}$$

where $\iota_{S_{\mathcal{I}}}$ is the indicator function of the convex set $S_{\mathcal{I}} = \{p \in \mathbb{R}^m \;:\; p_{\mathcal{I}} = 0\}$.

Proof. As ΦU has full column rank, the re-fitting of the solution (15) is the unique solution of the constrained least-square estimation problem (see Example 9)

$$\mathcal{R}_{\hat{x}}(y) = U(\Phi U)^+ y = \operatorname*{argmin}_{\tilde{x}\in\mathcal{M}_{\hat{x}}(y)} \|\Phi\tilde{x}-y\|^2 \;. \tag{89}$$

Remark that $\tilde{x} \in \mathcal{M}_{\hat{x}}(y) = \operatorname{Ker}[\mathrm{Id}^t_{\mathcal{I}^c}\Gamma] \Leftrightarrow (\Gamma\tilde{x})_{\mathcal{I}^c} = 0 \Leftrightarrow \iota_{S_{\mathcal{I}^c}}(\Gamma\tilde{x}) = 0$, where $S_{\mathcal{I}^c} = \{p \in \mathbb{R}^m \;:\; p_{\mathcal{I}^c} = 0\}$.

Using Fenchel transform, $\iota_{S_{\mathcal{I}^c}}(\Gamma\tilde{x}) = \max_{\tilde{z}} \langle \Gamma\tilde{x}, \tilde{z}\rangle - \iota^*_{S_{\mathcal{I}^c}}(\tilde{z})$, where $\iota^*_{S_{\mathcal{I}^c}}$ is the convex conjugate of $\iota_{S_{\mathcal{I}^c}}$. Observing that $\iota_{S_{\mathcal{I}}} = \iota^*_{S_{\mathcal{I}^c}}$ concludes the proof. ∎

Given Lemma 38, replacing $\Pi_{z^k+\sigma\Gamma v^k}$ in (58) by the projection onto $S_{\mathcal{I}}$, *i.e.,*

$$\Pi_{S_{\mathcal{I}}}(\tilde{z})_{\mathcal{I}^c} = \tilde{z}_{\mathcal{I}^c} \quad \text{and} \quad \Pi_{S_{\mathcal{I}}}(\tilde{z})_{\mathcal{I}} = 0 \;, \tag{90}$$

leads to the primal-dual algorithm of [8] applied to problem (88) which converges to the re-fitted estimator $\mathcal{R}_{\hat{x}}(y)$. It remains to prove that the projection $\Pi_{z^k+\sigma\Gamma v^k}$ defined in (58) converges to $\Pi_{S_{\mathcal{I}}}$ in finite time.

Proof of Theorem 36. First consider $i \in \mathcal{I}$, *i.e.*, $|\Gamma x^\star|_i > 0$. By assumption on α, $|\Gamma x^\star|_i \geq \alpha > 0$. Necessary $z_i^\star = \lambda \operatorname{sign}(\Gamma x^\star)_i$ in order to maximize (55). Hence, $|z^\star + \sigma\Gamma x^\star|_i \geq \lambda + \sigma\alpha$. Using the triangle inequality shows that

$$\lambda + \sigma\alpha \leqslant |z^\star + \sigma\Gamma x^\star|_i \leqslant |z^\star - z^k|_i + \sigma|\Gamma x^\star - \Gamma v^k|_i + |z^k + \sigma\Gamma v^k|_i \ . \tag{91}$$

Choose $\varepsilon > 0$ sufficiently small such that $\sigma\alpha - \varepsilon(1+\sigma) > \beta$. From the convergence of the primal-dual algorithm of [8], the sequence (z^k, x^k, v^k) converges to $(z^\star, x^\star, x^\star)$. Therefore, for k large enough, $|z^\star - z^k|_i < \varepsilon$, $|\Gamma x^\star - \Gamma v^k|_i < \varepsilon$, and

$$|z^k + \sigma\Gamma v^k|_i \geqslant \lambda + \sigma\alpha - \varepsilon(1+\sigma) > \lambda + \beta \ . \tag{92}$$

Next consider $i \in \mathcal{I}^c$, *i.e.*, $|\Gamma x^\star|_i = 0$, where by definition $|z^\star|_i \leqslant \lambda$. Using again the triangle inequality shows that

$$|z^k + \sigma\Gamma v^k|_i \leqslant |z^k - z^\star|_i + \sigma|\Gamma v^k - \Gamma x^\star|_i + |z^\star|_i \ . \tag{93}$$

Choose $\varepsilon > 0$ sufficiently small such that $\varepsilon(1+\sigma) < \beta$. As $(z^k, x^k, v^k) \to (z^\star, x^\star, x^\star)$, for k large enough, $|z^k - z^\star|_i < \varepsilon$, $|\Gamma v^k - \Gamma x^\star|_i < \varepsilon$, and

$$|z^k + \sigma\Gamma v^k|_i < \lambda + \varepsilon(1+\sigma) \leqslant \lambda + \beta \ . \tag{94}$$

It follows that for k sufficiently large $|z^k + \sigma\Gamma v^k|_i \leqslant \lambda + \beta$ if and only if $i \in \mathcal{I}^c$, and hence $\Pi_{z^k + \sigma K v^k}(\tilde{z}) = \Pi_{S_{\mathcal{I}}}(\tilde{z})$. As a result, all subsequent iterations of (58) will solve (88), and hence from Lemma 38 this concludes the proof of the theorem. ∎

REFERENCES

[1] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer, 2011.

[2] A. Belloni and V. Chernozhukov, *Least squares after model selection in high-dimensional sparse models*, Bernoulli, 19 (2013), pp. 521–547.

[3] T. Blumensath and M. E. Davies, *Iterative thresholding for sparse approximations*, J. Fourier Anal. Appl., 14 (2008), pp. 629–654.

[4] A. Buades, B. Coll, and J.-M. Morel, *A review of image denoising algorithms, with a new one*, SIAM J. Multiscale Model. Simul., 4 (2005), pp. 490–530.

[5] P. Bühlmann and B. Yu, *Boosting with the L2 loss: regression and classification*, J. Am. Statist. Assoc., 98 (2003), pp. 324–339.

[6] M. Burger, G. Gilboa, S. Osher, J. Xu, et al., *Nonlinear inverse scale space methods*, Communications in Mathematical Sciences, 4 (2006), pp. 179–212.

[7] A. Chambolle, V. Duval, G. Peyré, and C. Poon, *Geometric properties of solutions to the total variation denoising problem*, Preprint 1602.00087, Arxiv, 2016.

[8] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vis., 40 (2011), pp. 120–145.

[9] P. L. Combettes and J.-C. Pesquet, *A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery*, Selected Topics in Signal Processing, IEEE Journal of, 1 (2007), pp. 564–574.

[10] P. L. Combettes and J.-C. Pesquet, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer-Verlag, 2011, ch. Proximal Splitting Methods in Signal Processing, pp. 185–212.

[11] P.-A. Cornillon, N. W. Hengartner, and E. Matzner-Løber, *Recursive bias estimation for multivariate regression smoothers*, ESAIM: Probability and Statistics, 18 (2014), pp. 483–502.

[12] J. Darbon, A. Cunha, T. F. Chan, S. Osher, and G. J. Jensen, *Fast nonlocal filtering applied to electron cryomicroscopy*, in ISBI, 2008, pp. 1331–1334.

[13] C.-A. Deledalle, V. Duval, and J. Salmon, *Non-local methods with shape-adaptive patches (NLM-SAP)*, J. Math. Imaging Vis., 43 (2012), pp. 103–120.

[14] C.-A. Deledalle, N. Papadakis, and J. Salmon, *Contrast re-enhancement of total-variation regularization jointly with the Douglas-Rachford iterations*, in Signal Processing with Adaptive Sparse Structured Representations, 2015.
[15] C.-A. Deledalle, N. Papadakis, and J. Salmon, *On debiasing restoration algorithms: applications to total-variation and nonlocal-means*, in SSVM, 2015, pp. 129–141.
[16] C.-A. Deledalle, S. Vaiter, G. Peyré, and J. M. Fadili, *Stein unbiased gradient estimator of the risk (SUGAR) for multiple parameter selection*, SIAM J. Imaging Sci., 7 (2014), pp. 2448–2487.
[17] D. L. Donoho and J. M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, 81 (1994), pp. 425–455.
[18] J. Douglas and H. Rachford, *On the numerical solution of heat conduction problems in two and three space variables*, Transactions of the American mathematical Society, (1956), pp. 421–439.
[19] V. Duval, J.-F. Aujol, and Y. Gousseau, *A bias-variance approach for the nonlocal means*, SIAM J. Imaging Sci., 4 (2011), pp. 760–788.
[20] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist., 32 (2004), pp. 407–499.
[21] M. Elad, P. Milanfar, and R. Rubinstein, *Analysis versus synthesis in signal priors*, Inverse problems, 23 (2007), p. 947.
[22] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, CRC Press, 1992.
[23] E. I. George, *Minimax multiple shrinkage estimation*, Ann. Statist., 14 (1986), pp. 188–205.
[24] G. Gilboa, *A total variation spectral framework for scale and texture analysis*, SIAM J. Imaging Sci., 7 (2014), pp. 1937–1961.
[25] A. Griewank and A. Walther, *Evaluating derivatives: principles and techniques of algorithmic differentiation*, Society for Industrial and Applied Mathematics (SIAM), 2008.
[26] A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.
[27] J. Lederer, *Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions*, arXiv preprint arXiv:1306.0113, (2013).
[28] Y. Lin and H. H. Zhang, *Component selection and smoothing in multivariate nonparametric regression*, Ann. Statist., 34 (2006), pp. 2272–2297.
[29] P. Milanfar, *A tour of modern image filtering: New insights and methods, both practical and theoretical*, IEEE Signal Processing Magazine, 30 (2013), pp. 106–128.
[30] U. Naumann, *Optimal jacobian accumulation is np-complete*, Mathematical Programming, 112 (2008), pp. 427–441.
[31] G. R. Obozinski, M. J. Wainwright, and M. I. Jordan, *High-dimensional support union recovery in multivariate regression*, in Advances in Neural Information Processing Systems, 2008, pp. 1217–1224.
[32] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, *An iterative regularization method for total variation-based image restoration*, SIAM J. Multiscale Model. Simul., 4 (2005), pp. 460–489.
[33] S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin, *Sparse recovery via differential inclusions*, Appl. Comput. Harmon. Anal., (2016).
[34] S. Ramani, T. Blu, and M. Unser, *Monte-Carlo SURE: a black-box optimization of regularization parameters for general denoising algorithms*, IEEE Trans. Image Process., 17 (2008), pp. 1540–1554.
[35] P. Rigollet and A. B. Tsybakov, *Exponential screening and optimal rates of sparse estimation*, Ann. Statist., 39 (2011), pp. 731–471.
[36] Y. Romano and M. Elad, *Boosting of image denoising algorithms*, SIAM J. Imaging Sci., 8 (2015), pp. 1187–1219.
[37] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena, 60 (1992), pp. 259–268.
[38] J. Salmon, *On two parameters for denoising with non-local means*, Signal Processing Letters, IEEE, 17 (2010), pp. 269–272.
[39] X. Shen and J. Ye, *Adaptive model selection*, J. Am. Statist. Assoc., 97 (2002), pp. 210–221.
[40] C. Stein, *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, in Proc. 3th Berkeley Sympos. Math. Statist. and Prob., vol. 1, 1956, pp. 197–206.
[41] D. Strong and T. Chan, *Edge-preserving and scale-dependent properties of total variation regularization*, Inverse problems, 19 (2003), p. S165.
[42] H. Talebi, X. Zhu, and P. Milanfar, *How to saif-ly boost denoising performance*, IEEE Trans. Image Process., 22 (2013), pp. 1470–1485.

[43] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B, (1996), pp. 267–288.
[44] R. J. Tibshirani, *The lasso problem and uniqueness*, Electron. J. Statist., 7 (2013), pp. 1456–1490.
[45] A. N. Tikhonov, *On the stability of inverse problems*, Dokl. Akad. Nauk SSSR, 39 (1943), pp. 176–179.
[46] J. W. Tukey, *Exploratory data analysis*, Reading, Mass., 1977.
[47] S. Vaiter, C.-A. Deledalle, G. Peyré, C. Dossal, and J. Fadili, *Local behavior of sparse analysis regularization: Applications to risk estimation*, Appl. Comput. Harmon. Anal., 35 (2013), pp. 433–451.
[48] S. Vaiter, C.-A. Deledalle, G. Peyré, J. M. Fadili, and C. Dossal, *The degrees of freedom of partly smooth regularizers*, to appear in Annals of the Institute of Statistical Mathematics, (2016).
[49] D. Van De Ville and M. Kocher, *Nonlocal means with dimensionality reduction and sure-based parameter selection*, IEEE Trans. Image Process., 20 (2011), pp. 2683–2690.
[50] J. Xu and S. Osher, *Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising*, IEEE Trans. Image Process., 16 (2007), pp. 534–544.
[51] J. Ye, *On measuring and correcting the effects of data mining and model selection*, J. Am. Statist. Assoc., (1998), pp. 120–131.
[52] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. Roy. Statist. Soc. Ser. B, 68 (2006), pp. 49–67.