Variational Multiscale Nonparametric Regression: Algorithms and Implementation

Miguel del Alamo [1,†], Housen Li [2,†], Axel Munk [2,*] and Frank Werner [3,†]

[1] Department of Applied Mathematics, University of Twente, 7522 NB Enschede, The Netherlands; m.delalamoruiz@utwente.nl

[2] Institute for Mathematical Stochastics, University of Göttingen, 37073 Göttingen, Germany; housen.li@mathematik.uni-goettingen.de

[3] Institute of Mathematics, University of Würzburg, 97070 Würzburg, Germany; frank.werner@mathematik.uni-wuerzburg.de

[*] Correspondence: munk@math.uni-goettingen.de

[†] These authors contributed equally to this work.

Abstract: Many modern statistically efficient methods come with tremendous computational challenges, often leading to large-scale optimisation problems. In this work, we examine such computational issues for recently developed estimation methods in nonparametric regression with a specific view on image denoising. We consider in particular certain variational multiscale estimators which are statistically optimal in minimax sense, yet computationally intensive. Such an estimator is computed as the minimiser of a smoothness functional (e.g., TV norm) over the class of all estimators such that none of its coefficients with respect to a given multiscale dictionary is statistically significant. The so obtained multiscale Nemirowski-Dantzig estimator (MIND) can incorporate any convex smoothness functional and combine it with a proper dictionary including wavelets, curvelets and shearlets. The computation of MIND in general requires to solve a high-dimensional constrained convex optimisation problem with a specific structure of the constraints induced by the statistical multiscale testing criterion. To solve this explicitly, we discuss three different algorithmic approaches: the Chambolle-Pock, ADMM and semismooth Newton algorithms. Algorithmic details and an explicit implementation is presented and the solutions are then compared numerically in a simulation study and on various test images. We thereby recommend the Chambolle-Pock algorithm in most cases for its fast convergence. We stress that our analysis can also be transferred to signal recovery and other denoising problems to recover more general objects whenever it is possible to borrow statistical strength from data patches of similar object structure.

Keywords: non-smooth large-scale optimisation; image denoising; variational estimation; multiscale methods; MIND estimator

MSC: 62G05; 68U10

1. Introduction

Regression analysis has a centuries long history in science and is one of the most powerful and widely used tools in modern statistics. As it aims to discover a functional dependency f between certain variables of interest, it provides important insight into the relationship of such variables. Typically, the data are noisy and specific regression models provide a mathematical framework for the recovery of this unknown relationship f given such noisy observations. Due to the flexibility of such models, they can be accommodated to many different scenarios, e.g., for linearly dependent

data as in linear regression and for data following a general nonlinear structure as in nonlinear regression. The literature on this and the range of applications is vast, we exemplarily refer to the work of Draper and Smith [1] for an overview. Often, these models depend on a few parameters only specifying f already, which then can be estimated from the data in a relatively simple way. In contrast, nonparametric regression avoids such a restrictive modelling and becomes indispensable when prior knowledge is not sufficient for parametric modelling (see, e.g., [2,3]). One of the most fundamental and most studied nonparametric models is the Gaussian nonparametric regression (i.e., denoising) model with independent errors, sometimes denoted as white noise regression model (after a Fourier transformation). In this model, we aim to estimate the unknown regression function $f : [0,1]^d \to \mathbb{R}$ given noisy observations Y_i, which are related to f as

$$Y_i = f(x_i) + \sigma \epsilon_i, \;\; x_i \in \Gamma_n, \;\; i = 1, \ldots, n, \tag{1}$$

where ϵ_i are independent normal random variables with zero mean and unit variance and Γ_n is a discrete grid in $[0,1]^d$ that consists of n points. Note that σ determines the noise level, which we assume for simplicity to be constant (see, however, Section 5). We stress that much of what is addressed in this paper can be generalised to other error models and other domains than the d-dimensional unit cube $[0,1]^d$ (see also Section 5). However, to keep the presentation simple, we restrict ourselves to the Gaussian error and equidistant grids in the d-dimensional unit cube. A mathematical theory of such nonparametric regression problems has a long history in statistics (see [4] for an early reference), as they are among the simplest models where the unknown object is still in a complex function space and not just encoded in a low-dimensional parameter, yet they are general enough to cover many applications (see, e.g., [3]). Consequently, the statistical estimation theory of nonparametric regression is a well investigated area for which plenty of methods have been proposed, such as kernel smoothing [5,6]; global regularisation techniques such as penalised maximum likelihood [7]; ridge regression, which amounts to Tikhonov regularisation [8,9]; or total variation (TV) regularisation [10]. These methods have not been designed a-priori in a spatially adaptive way which could be overcome by a second generation of (sparse) localising regularisation methods originating in the development of wavelets [11]. To fine tune the estimator for first generation methods, usually a simple regularisation parameter has to be chosen (statistically); for wavelet-based estimators, this amounts to properly select (and truncate) the wavelet coefficients methods, e.g., by soft thresholding (see, e.g., [12]). We refer to Tsybakov [13] for an introduction to the modern statistical theory of nonparametric regression, mainly from a minimax perspective. One of the latest developments to the problem of recovering the function f in (1) already dates back to Nemirovski [14] and is implicitly exemplified by the Dantzig selector [15], which can be seen as a hybrid between a sparse approximation with respect to a dictionary and variational (L^1) regularisation. Such hybrid methods were coined by Grasmair et al. [16] as MIND estimators (MultIscale Nemirovski–Dantzig) and are the focus of this paper. We discuss these mainly in the context of statistical image analysis ($d = 2$), but stress that our findings also apply to signal recover $d = 1$ and to other situations where multiscale approaches are advantageous, e.g., for temporal-spatial imaging, where $d = 3, 4$.

1.1. Variational Denoising

One of the most prominent regularisation methods for image analysis is total variation (TV) denoising, which was popularised by Rudin, Osher and Fatemi [10] (for further developments towards a mathematical theory for more general variational methods, see, e.g., [17] and references therein). In the spirit of Tikhonov regularisation, the rationale behind is to enforce certain properties for the function f fitted to the model (1), which are encoded by a convex penalty term R such as the TV seminorm. Consequently, the weighted sum of a least-squares data fidelity term (corresponding to the

maximum likelihood estimation in model (1)), R is minimised over all functions g (e.g., of bounded variation) and the minimiser is taken as the reconstruction or estimator for f, i.e.,

$$\hat{f} \in \arg\min_{g} \left[\frac{1}{2} \sum_{i=1}^{n} (Y_i - g(x_i))^2 + \alpha R(g) \right] \tag{2}$$

with a weighting factor (sometimes called the regularisation parameter) $\alpha > 0$. As soon as R is convex, algorithmically, this leads to a convex optimisation problem, which can be solved efficiently in practice. This also applies to other convex data fidelity, e.g., when the likelihood comes from an exponential family model. To solve (2) numerically, most common are first-order methods (see, e.g., [18]), which boil down to computing the (sub-)gradients of the least-squares term and R.

One of the statistical disadvantages of methods of the form (2) is the usage of the global least-squares term, which to some extend enforces the same smoothness of g everywhere on the domain Γ_n. To overcome this issue, a spatially varying choice of the parameter α has been discussed in case of R being the TV seminorm (see, e.g., [19,20]). Other options are localised least-squares fidelities [21–23] or anisotropic total variation penalties, where the weighting matrix A is also improved iteratively based on the available data (see, e.g., [24]). We discuss another powerful strategy to cope with local inhomogeneity of the signal which has its origin in statistics and has been recently shown to be statistically optimal (in a certain sense to be made precise below) for various choices of penalties R.

1.2. Statistical Multiscale Methods

In the nonparametric statistics literature, multiscale methods have their origin in the discovery of wavelets [11], which Donoho and Johnstone [25] used to construct wavelet thresholding estimators and showed their ability to adapt to certain smoothness classes. Such estimators are computationally simple, as they only involve thresholding with respect to an orthonormal basis (cf. [12]). Following wavelets, myriad multiscale dictionaries tailored to different needs have been developed, such as curvelets [26], shearlets [27], etc. In a nutshell, the superiority of multiscale dictionaries over other bases for denoising and inverse problems lies in their excellent approximation and localising properties, which yields sparse approximations of functions with respect to loss functions that results from norms which average the error, e.g., L^p, $p \geq 1$, Sobolev or Besov norms. However, it is well known that despite this sparseness these estimators tend to present Gibbs-like oscillatory artefacts in statistical settings (which is not well reflected by such norms). This affects the quality of the overall reconstruction critically [28,29].

1.3. Variational Multiscale Methods

Variational multiscale methods offer a solution to the problem of such Gibbs-artefacts, as they combine multiscale methods with classical (non-sparse) regularisation techniques, with the idea of making use of the best of both worlds: the stringent data-fitting properties of (overcomplete) multiscale dictionaries, and the desired (problem dependent) smoothness imposed by a regularisation method, e.g., by TV regularisation.

Definition 1. *Let $\mathcal{F}$ be a class of functions, $R : \mathcal{F} \to \mathbb{R} \cup \{\infty\}$ a convex functional and $\{\phi_\lambda | \lambda \in \Lambda_n\}$ a finite collection of functions. Given observations Y_i as in (1), the MultIscale Nemirovski–Dantzig Estimator (MIND) is defined by the constrained optimisation problem*

$$\hat{f} \in \operatorname*{argmin}_{g \in \mathcal{F}} R(g) \quad \text{such that} \ \max_{\lambda \in \Lambda_n} |\langle \phi_\lambda, g - Y \rangle| \leq q_n \tag{3}$$

for a threshold $q_n > 0$, where we use the inner product $\langle h, g \rangle := n^{-1} \sum_{i=1}^{n} h(x_i) g(x_i)$.

Several particular cases of the MIND have been proposed: the first time by Nemirovski [14] (who gave credit to S. V. Shil'man) in 1985 for a system of indicator functions and a Sobolev seminorm, $R(g) = |g|_{H^s}$, Donoho [12] derived soft-thresholding from this principle where the dictionary $\{\phi_\lambda \mid \lambda \in \Lambda_n\}$ is a wavelet basis (for the latter, see also [30]). For the case of a curvelet frame, we refer to [28] and for a system of indicator functions and the Poisson likelihood to [31,32], where $R(g) = |g|_{BV}$ is the TV seminorm. Notice that, in practice, the choices of R and the dictionary reflect previous knowledge or expectations on the unknown function f: the amount of regularisation that R imposes should be dependent on how smooth we expect f to be; and the dictionary should measure patterns that we expect f to obey, e.g., elongated features in the case of curvelets and piecewise constant areas in the case of indicator functions. Furthermore, we have the following rule of thumb: the more redundant is the dictionary, the better is the statistical properties of the estimator but the more expensive is the computation of a solution to (3).

It is documented that the different proposed instances of the MIND show increased regularity and reduced artefacts as compared with standard multiscale methods (i.e., thresholding type estimators), but still avoid oversmoothing if the threshold q_n is appropriately chosen (see [29,31,32]). For instance, we illustrate in Figure 1 that the MIND with R the TV seminorm outperforms the soft-thresholding estimator under various choices of multiscale dictionaries, including wavelets, curvelets and shearlets. In addition to their good empirical performance, estimators of the form (3) have recently been analysed theoretically and proved to be (nearly minimax) optimal for nonparametric regression [16,29] and certain inverse problems [15,33]. We give a brief review of these theoretical results in Section 2.
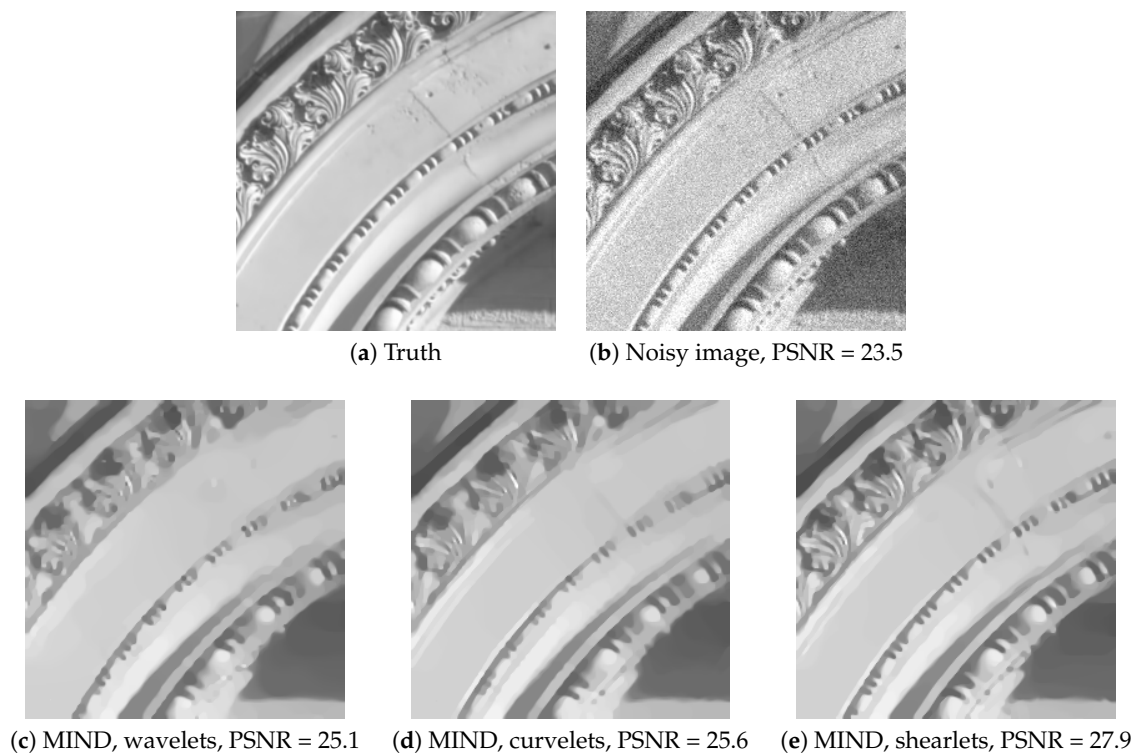

(**a**) Truth   (**b**) Noisy image, PSNR = 23.5

(**c**) MIND, wavelets, PSNR = 25.1   (**d**) MIND, curvelets, PSNR = 25.6   (**e**) MIND, shearlets, PSNR = 27.9

Figure 1. *Cont.*

(**f**) ST, wavelets, PSNR = 24   (**g**) ST, curvelets, PSNR = 23.7   (**h**) ST, shearlets, PSNR = 24.4
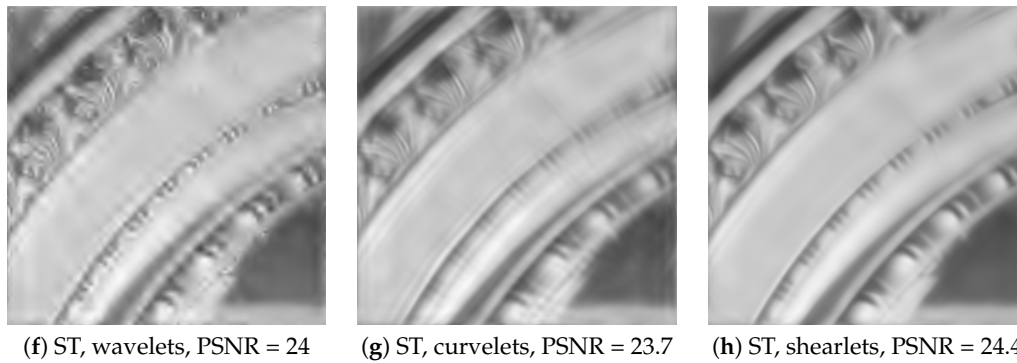
Figure 1. Comparison of MIND with TV regularisation and the soft-thresholding (ST) estimator on the "building" image from Darmstadt Noise Dataset [34] with different dictionaries. Details of dictionaries can be found in Section 4.2. The thresholds for MIND and ST are chosen the same as the 50% quantile of $\|\hat{\sigma}\epsilon\|_{\mathrm{MS}}$ in (5) (see Section 2.2). The noise level σ is estimated by a second order difference-based estimator [35] (see Section 4.3). (**a**) Truth; (**b**) Noisy image; (**c**) MIND, wavelets; (**d**) MIND, curvelets; (**e**) MIND, shearlets; (**f**) ST, wavelets; (**g**) ST, curvelets; (**h**) ST, shearlets.

1.4. Computational Challenges and Scope of the Paper

However, the practical and theoretical superiority of regularised multiscale methods over purely dictionary thresholding methods comes at a cost: instead of simple thresholding, a complex optimisation problem has to be solved, with an increase in computation time and the need of improved optimisation methods. In practice, the problem (3) is non-smooth (e.g., if R is chosen as the TV seminorm) and is furthermore high-dimensional due to a huge number of constraints. For instance, in the case of images of 256×256 pixels, the number of constraints is 1,751,915 for the dictionary of cubes of edge length ≤ 30 and 3,211,264 for that of shearlets.

Besides our theoretical review in Section 2, the goal of this paper is to discuss different particularly successful algorithmic approaches for the solution of (3) in Section 3. Among the first attempts is the usage of the alternating direction method of multipliers (ADMM) in [31,32,36], including a problem-specific convergence analysis. The computational disadvantage of this approach is the necessity to compute projections to the constraint set, which is most generally done by Dykstra's algorithm. Even though this can be efficiently implemented using GPUs (cf. [37,38]), the projection step provides a computational bottleneck of the overall algorithm. Several other convex optimisation algorithms such as Douglas–Rachford–Splitting (see [39]) or the Proximal Alternating Predictor Corrector algorithm (cf. [40]) also require the computation of such projections, and hence their computational performance is similar to the ADMM-based version. In view of this, we furthermore discuss two new approaches based on the Chambolle–Pock algorithm [41] on the one hand and a semismooth Newton method (cf. [42,43]) on the other hand. The former allows avoiding computing projections to the constraint sets, but requires only the computation of the corresponding resolvent operator, which reduces to a high-dimensional soft-thresholding problem. The latter allows solving a regularised version of the optimality conditions by Newton's method and then apply a path-continuation scheme to decrease the amount of regularisation. Both algorithms hence avoid Dykstra's projection step, but still have favourable convergence properties.

In Section 4, we finally discuss the practical advantages and disadvantages of the previously described algorithms along different numerical examples. For all examples checked in this paper, we found that the Chambolle–Pock algorithm is superior compared to the others. Finally, we provide a Matlab implementation of the Chambolle–Pock algorithm for this problem and code to run all examples in https://github.com/housenli/MIND.

2. Theoretical Properties of Variational Multiscale Estimation Methods

In this section, we briefly review some theoretical reconstruction properties of multiscale variational estimators within model (1). Recall that, in the nonparametric regression model (1), we have access to noisy samples of a function f at locations $x_i \in \Gamma_n$. We have some flexibility in choosing the underlying grid Γ_n: it could be for instance an equidistant d-dimensional grid (e.g., as pixels in an image), but other choices are possible (e.g., a polar grid).

2.1. Theoretical Guarantees

Estimators of the form (3) have been analysed from a theoretical viewpoint in a variety of settings. When the dictionary basis functions ϕ_λ are orthogonal, this becomes particularly simple, as then the evaluation functionals (if the truth equals g) $\langle \phi_\lambda, \epsilon \rangle$ become independent in model (3). This is valid for wavelet systems and their statistical analysis is vast (see, e.g., [44–52] for various forms of adaptation and thresholding techniques). If the dictionary is redundant the analysis becomes more difficult (see, however, [53] for the asymptotic validity of hard and soft thresholding in this case) and only in recent years it could be been shown that suitably constructed MINDs with redundant dictionaries perform optimal in a statistical (minimax) sense over certain function spaces. In the following, we summarise some of these results in an informal way. To this end, the notion of minimax optimality is key, which compares a given estimator with the *best* possible estimator in terms of their *worst case error* (see Equation (4) and Definition 3.1 in [13] for a formal definition).

- *Sobolev spaces*: The authors of [14,16] analysed the MIND with $R(g) = |g|_{H^s}$ and $\{\phi_\lambda\}$ being a set of indicator functions of rectangles at different locations and scales. They showed that, for the choice $q_n = C \sigma \sqrt{\log n / n}$ for an explicit constant $C > 0$, the MIND is minimax optimal up to logarithmic factors for estimating functions in the Sobolev space H^s. This means that the MIND's expected reconstruction error is of the same order as the error of the best possible estimator, i.e.,

$$1 \leq \frac{\sup_{|f|_{H^s} \leq L} \mathbb{E}\|f - \hat{f}_{MIND}\|_{L^p}}{\inf_{\text{all estimators } \hat{f}} \sup_{|f|_{H^s} \leq L} \mathbb{E}\|f - \hat{f}\|_{L^p}} \leq C \, Polylog(n). \tag{4}$$

Besides minimax optimality, Grasmair et al. [16] also showed that the MIND with H^s regularisation is also optimal for estimating functions in other Sobolev spaces H^t for certain smoothness indices t, a phenomenon known as adaptation.

- *Bounded variation*: In [29], the MIND with bounded variation regularisation $R(g) = |g|_{BV}$ was considered. It was shown that it is optimal in a minimax sense up to logarithms for estimating functions of bounded variation if $d = 2$. For $d \geq 3$, the discretisation matters further and this could only be shown in a Gaussian white noise model. Such results hold for a variety of dictionaries $\{\phi_\lambda\}$, such as wavelet bases, mixed wavelet and curvelet dictionaries, as well as suitable systems of indicator functions of rectangles.

In addition to theoretical guarantees, these results also provide a way of choosing the threshold parameter q_n in (3). Indeed, both for Sobolev and for bounded variation regularisation, it is shown (see again [16,29]) that the choice $q_n = C \sigma \sqrt{(\log n)/n}$ for an explicit constant $C > 0$ yields asymptotically optimal performance. The constant C depends on the dimension d and smoothness s of the functions, on whether we consider Sobolev or BV regularisation, and on the dictionary $\{\phi_\lambda\}$ we employ.

2.2. Practical Choice of the Threshold

Besides the theoretically (asymptotically) optimal choice of the parameter q_n, a Monte Carlo method for a finite sample choice was proposed by Grasmair et al. [16]. It is based on the observation that the multiscale constraint in (3) can be interpreted as a test statistic for testing whether the data Y is compatible with the function g, in the sense of (1). In fact, the "multiscales" come from not only performing one test, but many tests that focus on different features of g of various sizes, locations and

orientations. From this viewpoint, q_n is interpreted as a critical value for a statistical test, and statistical testing theory suggests that q_n should be chosen as a high quantile of the random variable

$$\|\sigma\epsilon\|_{\mathrm{MS}} := \sigma \max_{\lambda \in \Lambda_n} |n^{-1} \sum_{i=1}^{n} \epsilon_i \phi_\lambda(x_i)|. \tag{5}$$

This interpretation yields a practical way of choosing q_n: we simply pre-estimate σ, simulate independent realisations of the noise ϵ, compute their values in (5) and finally set q_n to be a quantile of that sample. This choice of q_n yields good practical performance (see Section 4) and is compatible with the theoretically optimal choice for n large enough. Methods for pre-estimating σ from the data can be found, e.g., in [35] (see Section 4.3 for simulations of the MIND estimator with estimated noise level).

3. Computational Methods

In this section, we discuss different algorithmic approaches to compute the MIND $\hat{f}$ in (3). First, we stress that the optimisation problem (3) is a non-smooth, convex, high-dimensional optimisation problem, which allows in principle to exploit any optimisation method designed for such situations. In the following, we restrict to three different approaches that we found particularly suited for our scenario. To set the notation in this section, we rewrite (3) as

$$\min_{v \in \mathbb{R}^n} J(v), \qquad J(v) := F(Kv) + G(v), \tag{6}$$

where F and G are lower semi-continuous, proper convex functionals given by

$$F(w) := 1_{\leq 0}(w - KY - q_n) + 1_{\leq 0}(-w + KY - q_n) \ \text{ for } w \in \mathbb{R}^{\#\Lambda_n},$$
$$G(v) := R(v) \ \text{ for } v \in \mathbb{R}^n.$$

and K is the linear (bounded) operator that maps from $\mathbb{R}^n$ to $\mathbb{R}^{\#\Lambda_n}$ and is defined by

$$[Kg_n]_\lambda := \langle \phi_\lambda, g_n \rangle \ \text{ for any } \lambda \in \Lambda_n.$$

Here, and in what follows, $1_{\leq 0}$ denotes the indicator function of the negative half-space, which is

$$1_{\leq 0}(v) = \begin{cases} 0 & v_i \leq 0 \text{ for all possible } i, \\ \infty & \text{else.} \end{cases}$$

Due to the convexity of F and G, the problem (6) can equivalently be solved by finding a root of the subdifferential $\partial J(v)$ (see, e.g., [54]), which is a generalisation of the classical derivative. Such roots correspond to stationary points of the evolution equation

$$\partial v(t) \in -\partial J(v(t)), \qquad t > 0.$$

Two fundamental algorithms for the solution of this equation arise from applying the explicit or implicit Euler method, which leads to the steepest decent procedure

$$v_{k+1} \in (I - \lambda_k \partial J)(v_k), \qquad k \in \mathbb{N} \tag{7a}$$

and the so-called proximal point method

$$v_{k+1} \in (I + \lambda_k \partial J)^{-1}(v_k), \qquad k \in \mathbb{N} \tag{7b}$$

respectively, where I denotes the identity operator and $\lambda_k > 0$ is a step size parameter. For (7a), λ_k should be chosen such that $(I - \lambda_k \partial J)$ is a contraction to ensure convergence. For continuously

differentiable J, it therefore suffices to choose $\lambda_k < L^{-1}$ with the Lipschitz constant L of ∇J; in practice, one often uses $\lambda_k = (2\hat{L})^{-1}$ with an estimator $\hat{L}$ of the Lipschitz constant (generated, e.g., by a power method). In the case of (7b), convergence can be obtained for any choice of $\lambda_k > 0$, but the computation of the so-called resolvent operator $(I + \lambda_k \partial J)^{-1}$ is clearly as difficult as the original problem. Motivated by these two methods, different algorithms have been suggested. Therefore, the subdifferential $\partial J(v)$ of J is split into the subdifferentials of F and G, which allow for a much simpler computation of the corresponding resolvent operators. This makes use of the formula

$$\partial J(v) = K^* \partial F\,(Kv) + \partial G(v), \tag{8}$$

which holds true whenever there exists some vector $v \in \mathbb{R}^n$ such that both F and G are finite and continuous at Kv and v respectively (see e.g., Prop. 5.6 in [55]). In our situation, this is the case whenever R is continuous at a point in the interior of the feasible set. With the help of (8), we can also derive the necessary and (due to convexity) sufficient first-order optimality conditions

$$-K^* w \in \partial G\,(v)\,, \tag{9a}$$
$$Kv \in \partial F^\star\,(w) \tag{9b}$$

with the conjugate functional

$$F^\star\,(w) := \sup_{v \in \mathbb{R}^n}\, \left[w^\top v - F\,(v) \right].$$

In the following, we present three different algorithms to solve (6) either via a specific splitting of ∂J or via the first-order optimality conditions: the Chambolle–Pock primal dual algorithm, an ADMM method and a semismooth Newton method.

3.1. The Chambolle–Pock Method

The primal dual algorithm by Chambolle and Pock [41] is based on a reformulation of the optimality conditions (9) as fixed point equations. If we multiply the second condition by a parameter $\tau > 0$ and add w on both sides yields

$$w + \tau Kv \in w + \tau \partial F^\star(w).$$

Similarly, the first condition yields

$$v - \delta K^* w \in v + \delta \partial G(v).$$

Hence, the solutions v and w of the optimality conditions can be found by repetitively applying the resolvent operators of G and the dual of F, which are given by

$$(I + \tau\, \partial G)^{-1}(v) := \text{argmin}_{x \in \mathbb{R}^n} \frac{\|x - v\|^2}{2\tau} + R(v) \tag{10a}$$

$$(I + \delta\, \partial F^\star)^{-1}(w) := \text{argmin}_{z \in \mathbb{R}^{\#\Lambda_n}} \frac{\|z - w\|^2}{2\delta} + q_n \sum_{\lambda \in \Lambda_n} \left| z_\lambda - \delta Y_\lambda \right|. \tag{10b}$$

This combined with an extrapolation step yields the Chambolle–Pock algorithm. It can also be interpreted as a splitting of the subdifferential ∂J into those of F and G, and then applying proximal point steps (i.e., backwards steps) to ∂G and the dual of ∂F.

The first proximal operator in (10) depends on the regularising functional $R(\cdot)$, which is typically convex, so the computation can be done efficiently. For instance, it can be solved by quadratic programming [56] if $R(\cdot)$ is a Sobolev norm and with Chambolle's algorithm [57] if $R(\cdot)$ is the TV penalty.

The second proximal operator in (10) involves the high-dimensional constraint. However, the solution to (10b) is simply the soft-thresholding operator applied to $\delta^{-1}w - KY$ with threshold $q_n \delta$. Altogether, the Chambolle–Pock algorithm applied to (3) is given in Algorithm 1.

Algorithm 1: Chambolle–Pock algorithm

Require: $\delta, \tau > 0, \theta \in [0,1], k = 0, (v_0, w_0) \in X \times Y$, stopping criterion

1: **while** stopping criterion not satisfied **do**

2: $\quad w_{k+1} \leftarrow (I + \delta\,\partial F^\star)^{-1}(w_k + \delta\,K\,\widetilde{v}_k)$

3: $\quad v_{k+1} \leftarrow (I + \tau\,\partial G)^{-1}(v_k - \tau\,K^*\,w_{k+1})$

4: $\quad \widetilde{v}_{k+1} \leftarrow v_N + \theta(v_{k+1} - v_k)$

5: $\quad k \leftarrow k + 1$

6: **end while**

7: Return (v_k, w_k)

To run the Chambolle–Pock algorithm, we need to choose the step sizes δ and τ, and convergence of the algorithm is ensured whenever $\tau\delta \le \|K\|_{op}^{-2}$. Due to the different difficulty of the two subproblems, it is reasonable to choose $\delta \ne \tau$ and, in particular, to choose $\delta > \tau$. More precisely, the subproblem with respect to F is more involved because of the large number of constraints, so this requires a much smaller step size, which by the Moreau's identity is equivalent to choosing a much larger δ. We observe in practice that the choices $\delta = \|K\|_{op}^{-1}\sqrt{n}$ and $\tau = \|K\|_{op}^{-1}/\sqrt{n}$ yield good results (see Section 4).

We remark that, in the Chambolle–Pock method applied to this problem, the high-dimensionality appears only in the very simple problem of soft-thresholding. This is a very convenient way of dealing with it and, as shown below, makes the Chambolle–Pock-algorithm superior over the ADMM method.

3.2. ADMM Method

The alternating direction methods of multipliers (ADMM) can be seen as a variant of the augmented Lagrangian method (see [58,59] for classical references). To derive it, we introduce a slack variable $h \in \mathbb{R}^{\#\Lambda_n}$ and rewrite (3) it into the equivalent problem

$$\operatorname*{argmin}_{v \in \mathbb{R}^n, w \in \mathbb{R}^{\#\Lambda_n}} F(w) + G(v) \ \text{ subject to } Kv = w.$$

By the convex duality theory, it is equivalent to find the saddle point of the augmented Lagrangian $L_\lambda(v, w; h)$, that is,

$$\operatorname*{argmin}_{v,w} \max_{h \in \mathbb{R}^{\#\Lambda_n}} F(w) + G(v) + \langle h, Kv - w \rangle + \frac{\lambda}{2}\|Kv - w\|^2$$

where $h \in \mathbb{R}^{\#\Lambda_n}$ is the Lagrangian multiplier and $\lambda > 0$. As its name suggests, the ADMM algorithm solves this saddle point problem alternately over v, w and h in a Gauß–Seidel fashion (i.e., successive displacement). The details are given in Algorithm 2 below.

The usage of the ADMM for the problem (3) was first proposed by Frick et al. [31]. One central difference compared to the Chambolle–Pock algorithm is that the ADMM splitting avoids the usage of $F^\star$. Instead, it deals with the high-dimensional constraint in an explicit way, which ultimately results in a slower performance.

From the steps performed by the ADMM in Algorithm 2, the first one (Line 2) involves the proximal operator of R and can typically be dealt with with a standard algorithm (see the discussion

in the Chambolle–Pock algorithm, Section 3.1). The second step (Line 3) is more challenging, as it involves solving the optimisation problem

$$w_k = \operatorname*{argmin}_w \frac{\lambda}{2} \| w - (Kv_k + \lambda^{-1} h_{k-1}) \|^2 \quad \text{subject to} \quad \max_{\lambda \in \Lambda_n} |w_\lambda - \langle \phi_\lambda, Y \rangle| \le q_n. \tag{11}$$

In other words, we have to find the orthogonal projection of the point $Kv_N + \lambda^{-1} h_{N-1}$ to the feasible set

$$\left\{ w \in \mathbb{R}^{\#\Lambda_n} \mid \max_{\lambda \in \Lambda_n} |w_\lambda - \langle \phi_\lambda, Y \rangle| \le q_n \right\}.$$

This set is the intersection of $2\#\Lambda_n$ half-spaces, and it is known to be non-empty (as it always contains $\{ \langle Y, \phi_\lambda \rangle \mid \lambda \in \Lambda_n \}$). The projection problem (11) can be solved by Dykstra's algorithm [60,61], which converges linearly [62] (see [63] for an efficient stopping rule).

Finally, it follows from Corollary 3.1 in [64] that the ADMM has a linear convergence guarantee for the problem (6).

Algorithm 2: Alternating direction method of multipliers (ADMM)

Require: data $Y \in \mathbb{R}^n$, step size $\lambda > 0$, tolerance $\epsilon > 0$, initial values v_0, w_0, h_0

1: **while** $\max\{ \| Kv_k - w_k \|, \| K(v_k - v_{k-1}) \} > \epsilon$ **do**

2:    $v_k = \operatorname{argmin}_v \dfrac{\lambda}{2} \| Kv - (w_{k-1} - \lambda^{-1} h_{k-1}) \|^2 + G(v)$

3:    $w_k = \operatorname{argmin}_w \dfrac{\lambda}{2} \| w - (Kv_k + \lambda^{-1} h_{k-1}) \|^2 + F(w)$

4:    $h_k = h_{k-1} + \lambda (Kv_k - w_k)$

5:    $k \leftarrow k + 1$

6: **end while**

7: Return (v_k, w_k)

3.3. Semismooth Newton Method

Besides the optimality conditions (9), the problem (3) can also be solved by the so-called Karush–Kuhn–Tucker conditions, which are necessary and sufficient as well. Using the notation of this section, the original problem (3) is equivalent to

$$\min_{v \in \mathbb{R}^n} R(v) \quad \text{such that} \quad q_n - Kv + KY \ge 0, q_n - KY + Kv \ge 0. \tag{12}$$

This is a convex optimisation problem with linear inequality constraints, and if we introduce $B : \mathbb{R}^n \to \mathbb{R}^{2\#\Lambda_n}$ as $v \mapsto Bv$ with $B = (K^\top, -K^\top)^\top$, then a vector v is a solution to (12) if and only if there exists a vector of Lagrange multipliers $\lambda \in \mathbb{R}^{2\#\Lambda_n}_{\ge 0}$ such that

$$\partial R(v) \ni -B^\top \lambda, \tag{13a}$$

$$q_n - Bv + BY \ge 0, \tag{13b}$$

$$\lambda_i (q_n - Bv + BY)_i = 0 \quad \text{for all} \quad 1 \le i \le 2\#\Lambda_n. \tag{13c}$$

The latter two conditions can be reformulated as

$$\lambda_i = \max\{ 0, \lambda_i + c(Bv - BY - q_n)_i \} \quad \text{for all} \quad 1 \le i \le 2\#\Lambda_n. \tag{14}$$

The immediately visible advantage of this formulation over the original problem is that the inequality constraints have been transformed into equations. Now, suppose for a moment that the

functional R is twice differentiable. Then, $\partial R(v)$ is single valued and (13a) is a differentiable equation for v and λ. It seems a natural approach to solve the corresponding system of Equations (13a) and (14) via an analog of Newton's method. On the other hand, even if R was twice differentiable, the max function in (14) is not differentiable in the classical sense. The application of Newton's method is however still possible, as the max function turns out to be semismooth (see Definition 2.5 in [42]). For a general operator B, it is however not clear if the overall system (13a) and (14) can also be described as finding the root of a semismooth function. Therefore, one introduces a regularisation parameter $\beta \in (0,1)$ and replaces (14) by the regularised equation $\lambda_i = \beta \max\{0, \lambda_i + c\,(Bv - BY - q_n)_i\}$ for all $1 \leq i \leq 2\#\Lambda_n$, which is in turn equivalent to

$$\lambda_i = \max\left\{0, \frac{\beta c}{1 - \beta}\,(Bv - BY - q_n)_i\right\} \quad \text{for all} \quad 1 \leq i \leq 2\#\Lambda_n. \tag{15}$$

This system is now explicit in λ and yields in combination with (13a) the overall system

$$\partial R\,(v) \ni -\frac{1}{\delta}B^\top \max\{0, (Bv - BY - q_n)\} \tag{16}$$

with the new regularisation parameter $\delta := (1 - \beta)/(\beta c) \in (0, \infty)$. This system can—for twice differentiable R and under appropriate assumptions on B satisfied in our example—now be shown to be semismooth (cf. [65]). Furthermore, for $\delta \searrow 0$, the solution of the regularised system (16) converges towards a solution of the original system (13). This follows from the fact that (15) is in fact the Moreau–Yosida regularisation of the second optimality condition (9b). In practice, this limiting process is realised by a path-continuation scheme, sustaining the superlinear convergence behaviour.

For differentiable R, such as the Sobolev seminorm $R(v) = |v|_{H^s}$, the implementation of the semismooth Newton method with path-continuation is now straightforward: the overall system to be solved can now be written as

$$T_\delta(v) = 0, \quad \text{where } T_\delta = \partial|v|_{H^s} + \delta^{-1}B^T \max\{0, Bv - BY - q_n\}. \tag{17}$$

Denote by $\mathcal{D}_k[T_\delta]$ the generalised derivative of the functional at the position u_k. Then, we initialise the iteration at $\delta_0 > 0$ and $u_{0,\delta}$ and solve the linear equations

$$\mathcal{D}_k[T_\delta]u_{k+1,\delta} = \mathcal{D}_k[T_\delta]u_{k,\delta} + T_\delta(u_{k,\delta}) \quad \text{for } k \geq 0$$

iteratively until a stopping criterion is satisfied. Then, the parameter δ is decreased by a fixed factor (say $1/2$) and the iteration is started again with $u_{k,\delta}$ as the initial guess. This continuation process is stopped until a global error criterion is reached, which is formulated in terms of the number and magnitude of the violated constraints (see Algorithm 3). Note that, for $\delta > 0$, the underlying minimisation problem is strictly convex (this is an immediate property of the Moreau–Yosida regularisation, cf. [43]) and the subdifferential is Lipschitz continuous with Lipschitz constant $\sim 1/\delta$, and hence the Newton iteration for (17) will converge for all initial guesses in a ball with radius $\sim\delta$ around 0. Hence, whenever the inner iteration in Algorithm 3 does not converge, it should be re-started with a larger value of δ. However, a larger value of δ clearly prolongs the run-time of Algorithm 3, and hence the initial value δ should not be too large.

In the case of a non-differentiable R such as the TV-seminorm, we introduce another regularisation for R, resulting in its Huber regularisation

$$TV_\beta\,(v) = \int \min\{\beta^{-1}\,|\nabla v|^2, |\nabla v|\}\,dx.$$

This functional is differentiable for any $\beta > 0$, and to compute the limit $\beta \searrow 0$ we again apply a path-continuation strategy (cf. [43]). In this case, the superlinear convergence behaviour is sustained.

However, we remark that the path-following routine for two parameters (δ and β) turns out to be unstable in practice. Instead, it might be desirable to look for the Moreau–Yosida regularisation of *both* the constraint and R, that is, to regularise the optimality conditions (13) in one step with just one parameter. We leave this idea to be pursued in future work.

Finally, we remark that this superlinear convergence of the semismooth Newton method is a huge theoretical advantage over the Chambolle–Pock and the ADMM algorithms. In practice, however, the situation is more complex, as the convergence speed of Newton's method depends strongly on the initialisation. In Table 1, we provide a summary of the comparison between the three methods.

Algorithm 3: Semismooth Newton method for $R(v) = |v|_{H^s}$

Require: data $Y \in \mathbb{R}^n$, step size $\Delta\delta > 0$, tolerance ϵ, $\delta_{\min}, \rho_{min}, r_{min} > 0$, initial value δ

1: $v_{old} \leftarrow 0 \in \mathbb{R}^n$

2: *ratio*, *res* $\leftarrow 1$

3: **while** $\delta > \delta_{min}$ **do**

4:      **while** *ratio* $> \rho_{min}$ or *res* $> r_{min}$ **do**

5:          $v_{new} \leftarrow$ solution to equation: $\nabla T_\delta(v_{old})v_{new} = \nabla T_\delta(v_{old})v_{old} - T_\delta(v_{old})$ with tolerance ϵ

6:          *ratio* $\leftarrow \frac{\#\{q_n - Bv_{new} + BY < 0\}}{2\#\Lambda_n}$

7:          *res* $\leftarrow \frac{\sqrt{\|\max\{Kv_{new} - KY - q_n, 0\}\|^2 + \|\max\{KY - q_n - Kv_{new}, 0\}\|^2}}{\|v_{new}\|}$

8:      **end while**

9:      $v_{old} \leftarrow v_{new}$

10:     $\delta \leftarrow \delta \cdot \Delta\delta$

11: **end while**

12: Return v_{old}

Table 1. Summary of the comparison of the Chambolle–Pock, ADMM and semismooth Newton algorithms.

	Dependence on Initialisation	Theor. Convergence Speed	Practical Performance
Chambolle–Pock	no	linear	Good for smooth and nonsmooth R
ADMM	no	linear	Too slow
Semismooth Newton	yes	superlinear	Good for smooth R Unstable otherwise

4. Numerical Study

In this section, we compare the practical performance of the Chambolle–Pock, ADMM and semismooth Newton algorithms in computing MIND and demonstrate the empirical performance of MIND with respect to various choices of dictionaries. We refer to, e.g., [16,28,29,31,32,36,38] for further numerical examinations of MIND. We select in particular Sobolev H^1 and TV seminorms as the regularisation functional, with the former differentiable but the latter non-differentiable and the 50%-quantile of $\|\sigma\epsilon\|_{MS}$ in (5) as q_n for MIND. Concerning measure of image quality, we consider the peak signal-to-noise ratio (PSNR [66]), the structural similarity index measure (SSIM [67]) and the visual information fidelity (VIF [68]) criteria. The implementation in MATLAB for the Chambolle–Pock algorithm, together with code that reproduces all the following numerical examples, is available at https://github.com/housenli/MIND.

4.1. Comparison of Three Algorithms

Here, we compare the convergence speed of the Chambolle–Pock, ADMM and semismooth Newton algorithms in solving (3) or equivalently (6). We stress that the relative performance of the three algorithms remains similar over different settings. Thus, as a particular example, we choose the "cameraman" image (256×256 pixels) as the function f in model (1) and the indicator functions of dyadic (partition) system of cubes (cf. Definition 2.2 in [16]), which consists of cubes

$$\left[i2^{-\ell}, (i+1)2^{-\ell} \right] \times \left[j2^{-\ell}, (j+1)2^{-\ell} \right] \quad \subseteq \quad [0,1]^2 \qquad \text{for all possible } i, j, \ell \in \mathbb{N} \qquad (18)$$

as the dictionary $\{\phi_\lambda | \lambda \in \Lambda_n\}$. The noise level σ is assumed to be known and is chosen such that the signal-to-noise ratio (SNR) $\max_{i=1,\ldots,n} |f(x_i)|/\sigma = 30$, see Figure 2. Besides the aforementioned image quality measures, we employ objective values $R(g_k)$, (relative) constraint gaps $\left(\max_{\lambda \in \Lambda_n} |\langle \phi_\lambda, g_k - Y \rangle| - q_n \right) / q_n$, and the distance $\|g_k - g_\infty\|$ to the limit solution g_∞ to examine the evolution of iterations g_k. The limit solution g_∞ is obtained for each algorithm after a large number of iterations, that is 3×10^4 outer iterations for the Chambolle–Pock, 3×10^3 outer iterations for the ADMM, and the largest possible number of iterations until which the path-continuation scheme remains stable for the semismooth Newton method (in the considered setup, 29 outer iterations in case of H^1 regularisation). As mentioned in Section 3.3, the semismooth Newton method is unstable for the case of TV regularisation, which is thus not reported here. All limit solutions g_∞ are shown in Figures 3 and 4 and are visually quite the same in each figure (while MIND with TV regularisation leads to a slightly better result than that with H^1 regularisation). This indicates that all three algorithms are able to find the global solution of (3) to a desirable accuracy.
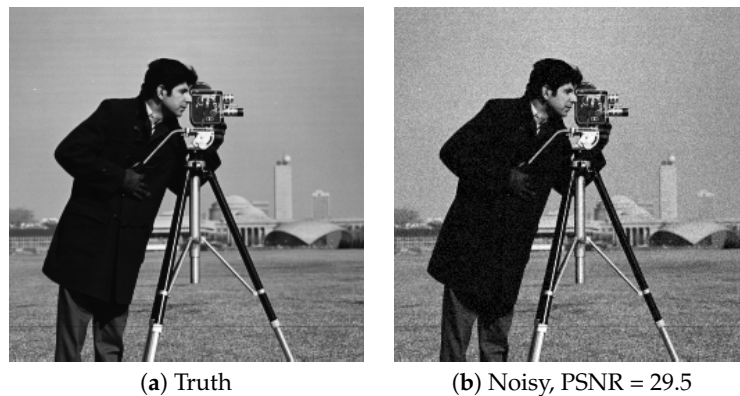

(**a**) Truth   (**b**) Noisy, PSNR = 29.5

Figure 2. Image "cameraman" and its noisy version with SNR = 30. (**a**) Truth; (**b**) Noisy.

(**a**) Chambolle–Pock   (**b**) ADMM   (**c**) Semismooth Newton
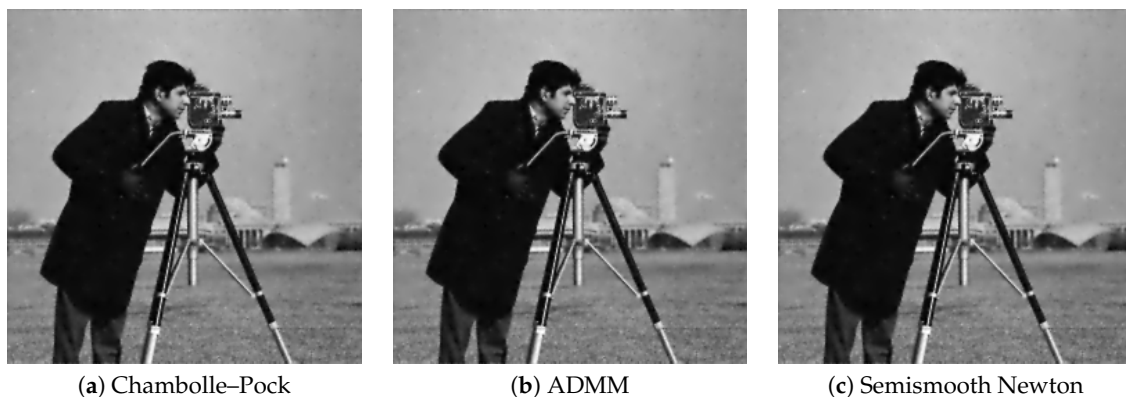
Figure 3. Limit solutions of all three algorithms for MIND with H^1 (PSNR = 30 for all) regularisation after long iterations. (**a**) Chambolle–Pock; (**b**) ADMM; (**c**) Semismooth Newton.

To reduce the burden of practitioners on parameter tuning, we set the parameters in the Chambolle–Pock (Algorithm 1) by default as $\theta = 1$, $\delta = \|K\|_{op}^{-1}\sqrt{n}$ and $\tau = \|K\|_{op}^{-1}/\sqrt{n}$ in all settings. We note that the performance of the Chambolle–Pock, which is reported below, can be further improved by fine tuning θ and δ and possibly also by preconditioning [69], for every choice of regularisation functional and dictionary. In contrast, the parameters of the ADMM and the semismooth Newton algorithms are fine tuned towards the best performance for each case. More precisely, for ADMM, we choose $\lambda = 50$ in the case of H^1 and $\lambda = 0.5$ in the case of TV, and, for the semismooth Newton, we choose $\delta = 1/3$ and $\Delta\delta = 1/3$ in the case of H^1. The performance of three algorithms over time is shown in Figures 5 and 6 for H^1 and TV regularisation, respectively. In both cases, the Chambolle–Pock clearly outperformed the ADMM with respect to all considered criteria. The semismooth Newton algorithm, as is ensured by the theory, exhibited a faster convergence rate than the other two, but this happened only at a very late stage. Moreover, the path-continuation scheme is sensitive to the choice of parameters and makes it difficult to decide the right stopping stage in general. In summary, from a practical perspective, the overall performance of the Chambolle–Pock algorithm was found to be superior compared to the other algorithms. We speculate, however, that a hybrid combination of the Chambolle–Pock and the semismooth Newton algorithm (switching to the semismooth Newton algorithm after a burn in period using the Chambolle Pock algorithm) might lead to further improvement.
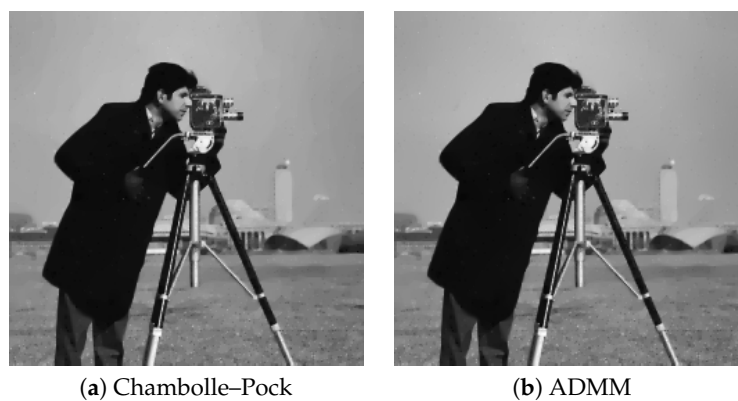

(**a**) Chambolle–Pock　　　　　　　　　　(**b**) ADMM

Figure 4. Limit solutions of Chambolle–Pock and ADMM algorithms for MIND with TV (PSNR = 30.6 for both) regularisation after long iterations. As described in the main text, the semismooth Newton method showed an unstable behaviour in combination with TV regularisation, and hence the result is not documented here. (**a**) Chambolle–Pock; (**b**) ADMM.
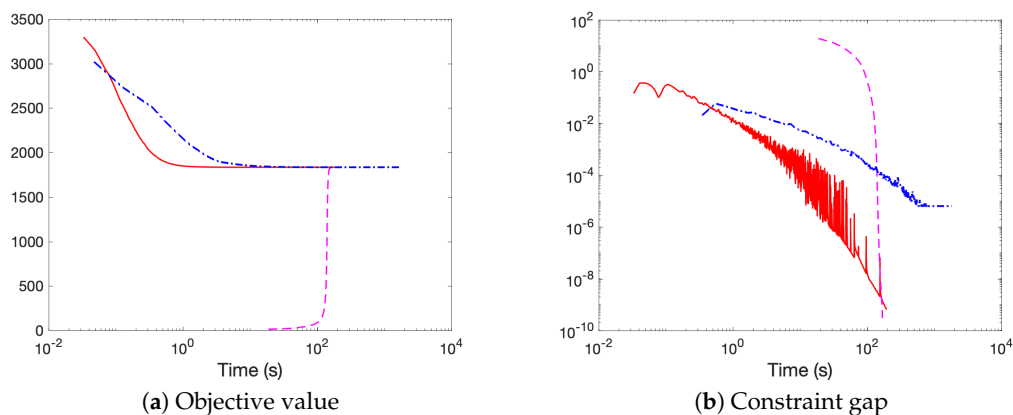

(**a**) Objective value　　　　　　　　　　(**b**) Constraint gap

Figure 5. *Cont.*

(**c**) Distance to limit

(**d**) PSNR
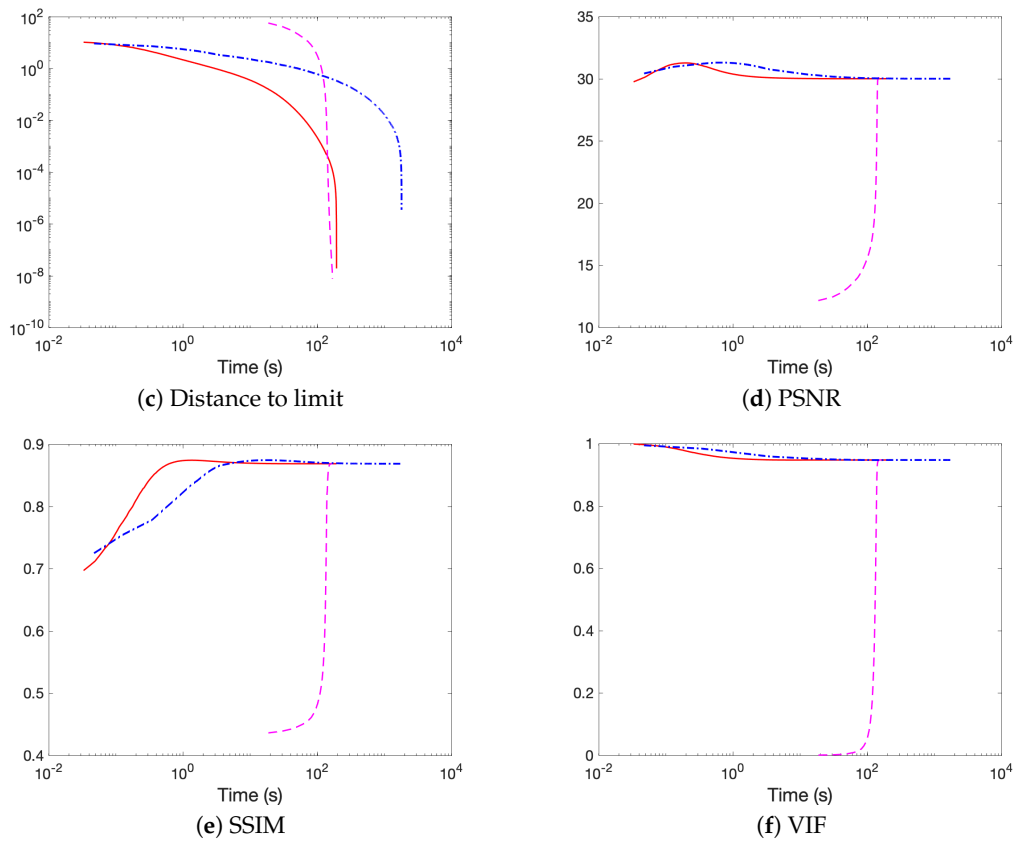
(**e**) SSIM

(**f**) VIF

Figure 5. Performance of Chambolle-Pock (line), ADMM (dash-dot) and semismooth Newton (dash) algorithms for MIND with H^1 regularisation over time. (**a**) Objective value; (**b**) Constraint gap; (**c**) Distance to limit; (**d**) PSNR; (**e**) SSIM; (**f**) VIF.

4.2. Comparison of Different Dictionaries

We next investigate the choice of dictionary $\{\phi_\lambda | \lambda \in \Lambda_n\}$ and its impact on the practical performance. Five different dictionaries are considered. The first consists of the indicator functions of dyadic cubes defined in (18), and the second is composed of the indicator functions of small cubes of edge length ≤ 30, i.e.,

$$\frac{1}{n} [i, i + \ell] \times [j, j + \ell] \quad \subseteq \quad [0, 1]^2 \qquad \text{for } \ell = 1, \ldots, 30, \text{ and all possible } i, j \in \mathbb{N}.$$

The third is the system of tensor wavelets, in particular, the Daubechies' symlets with six vanishing moments. The fourth is the frame of curvelets and the last is that of shearlets, both of which are constructed using default parameters in packages CurveLab (http://www.curvelet.org) and ShearLab3D (http://shearlab.math.lmu.de), respectively. As shown in Figures 3 and 4, different algorithms lead to almost the same final result, we only report the Chambolle–Pock algorithm, due to its empirically fast convergence.
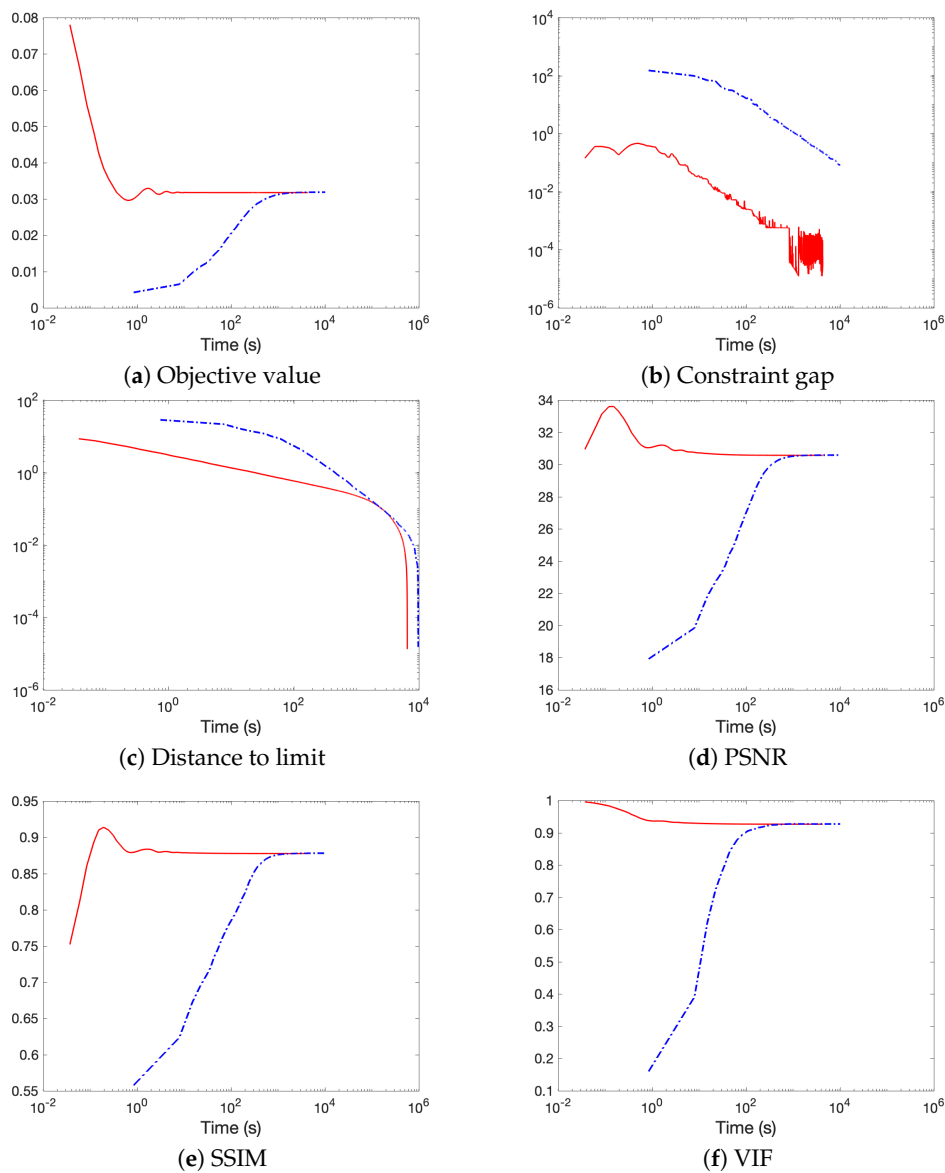
Figure 6. Performance of Chambolle–Pock (line) and ADMM (dash-dot) algorithms for MIND with TV regularisation over time. (**a**) Objective value; (**b**) Constraint gap; (**c**) Distance to limit; (**d**) PSNR; (**e**) SSIM; (**f**) VIF.

As a reflection of versatility of images, we consider three test images of different types. They are a magnetic resonance tomography image of mouse "brain' (cf. Figure 7a) from Radiopaedia.org (rID: 67777), a "cell" image (cf. Figure 8a) taken from [70] and a mouse "BIRN" image (cf. Figure 9a) from cellimagelibrary.org (doi:10.7295/W9CCDB17) (see also Figure 1 in the Introduction as well as Figure 10 in Section 4.3 for the case of natural images). All images are rescaled to 256 × 256 pixels via bicubic interpolation. The SNR is set to 20 over all cases (see Figure 11 for noisy image)s). The results on all test images are shown in Figures 7–9. In general, the dictionary of shearlets performed the best, which is followed by that of curvelets, while the rest (i.e., dyadic cubes, small cubes and wavelets) had similar performance. One exception is the "cell" image, where the dictionary of indicator functions of cubes, in particular of small cubes, was better at detecting tiny white balls (see Figure 8). This is because of the similarity of such features and the cubes at small scales in the dictionary. The average performance over 10 random repetitions measured by aforementioned image quality measures as well as mean integrated squared error

$$\text{MISE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - f(x_i))^2$$

is reported in Table 2, which is consistent with the virtual inspections in Figures 7–9. We note that 10 repetitions are sufficient here, since the variation in each repetition is comparably small (cf. standard deviations reported in parenthesis in Table 2). As a remark, we emphasise that the difference in performance due to the choice of dictionaries is generally negligible, but the dictionary of shearlets slightly outperforms the other choices.

(**a**) Truth   (**b**) Dyadic cubes, PSNR = 27.5   (**c**) Small cubes, PSNR=28.2

(**d**) Wavelets, PSNR = 28.1   (**e**) Curvelets, PSNR = 29.1   (**f**) Shearlets, PSNR = 30.6
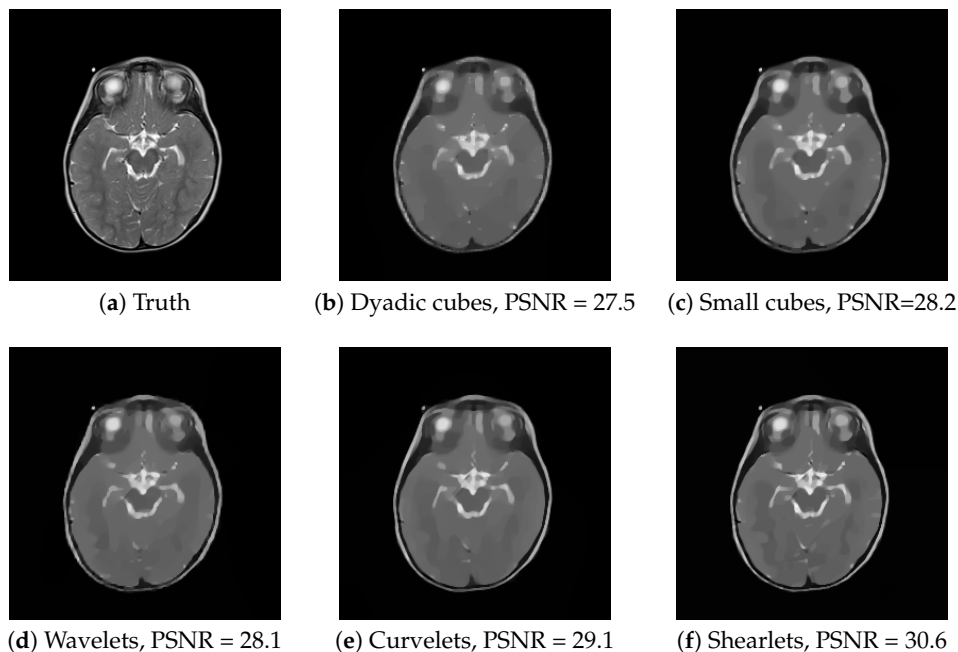
Figure 7. Results on "brain" by MIND with TV regularisation and different dictionaries. (**a**) Truth; (**b**) Dyadic cubes, PSNR = 27.5; (**c**) Small cubes, PSNR = 28.2; (**d**) Wavelets, PSNR = 28.1; (**e**) Curvelets, PSNR = 29.1; (**f**) Shearlets, PSNR = 30.6.

(**a**) Truth   (**b**) Dyadic cubes, PSNR = 31.8   (**c**) Small cubes, PSNR = 33.6

(**d**) Wavelets, PSNR = 32.9   (**e**) Curvelets, PSNR = 32.9   (**f**) Shearlets, PSNR = 33.9
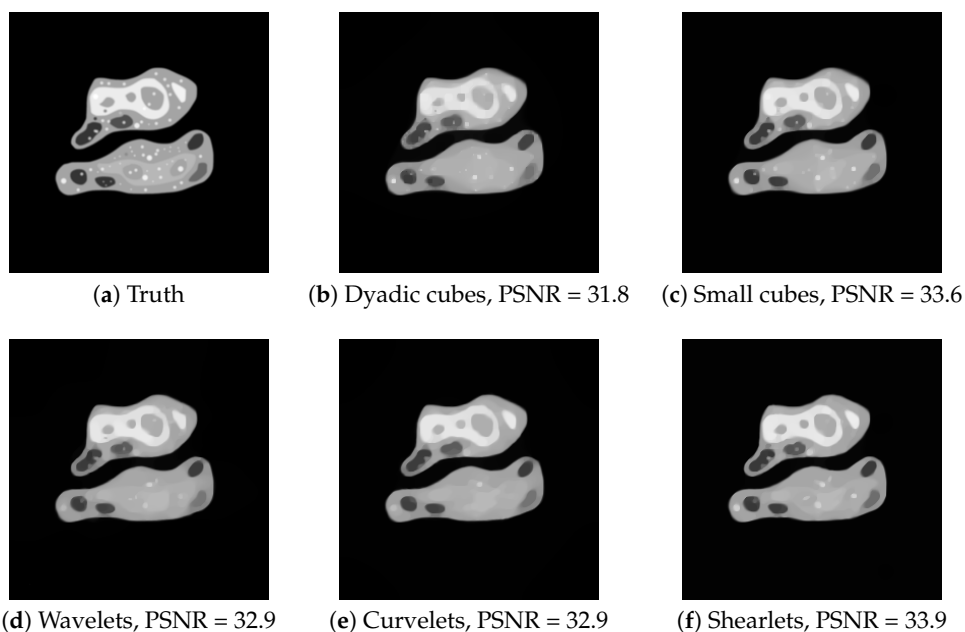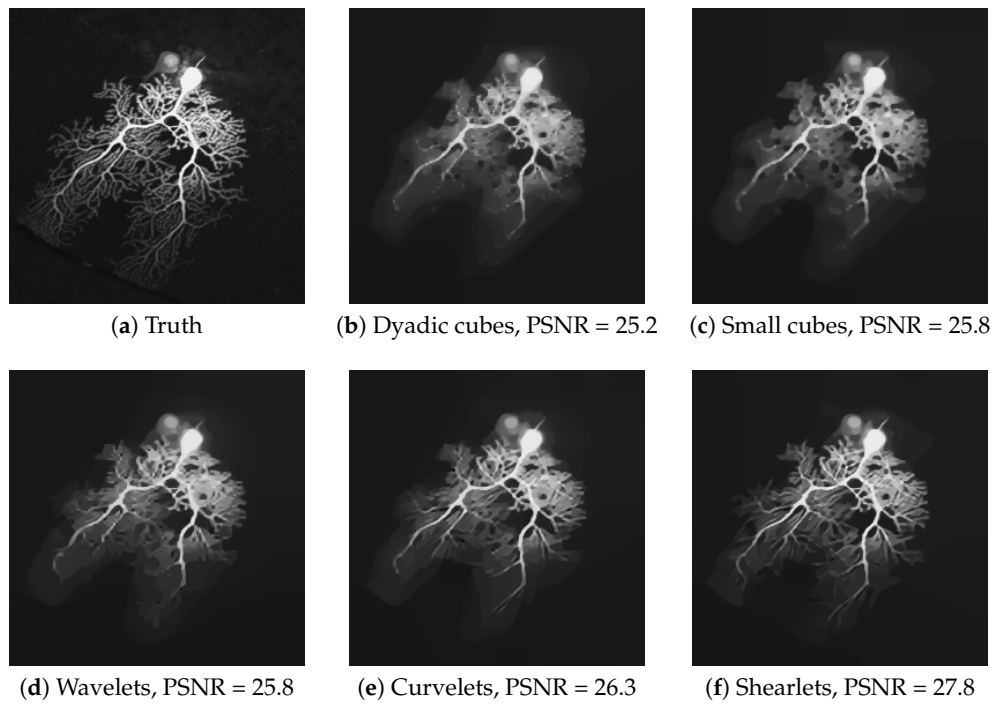
Figure 8. Results on "cell" by MIND with TV regularisation and different dictionaries. (**a**) Truth; (**b**) Dyadic cubes, PSNR = 31.8; (**c**) Small cubes, PSNR = 33.6; (**d**) Wavelets, PSNR = 32.9; (**e**) Curvelets, PSNR = 32.9; (**f**) Shearlets, PSNR = 33.9.

(**a**) Truth　　　　　(**b**) Dyadic cubes, PSNR = 25.2　　　　　(**c**) Small cubes, PSNR = 25.8

(**d**) Wavelets, PSNR = 25.8　　　　　(**e**) Curvelets, PSNR = 26.3　　　　　(**f**) Shearlets, PSNR = 27.8
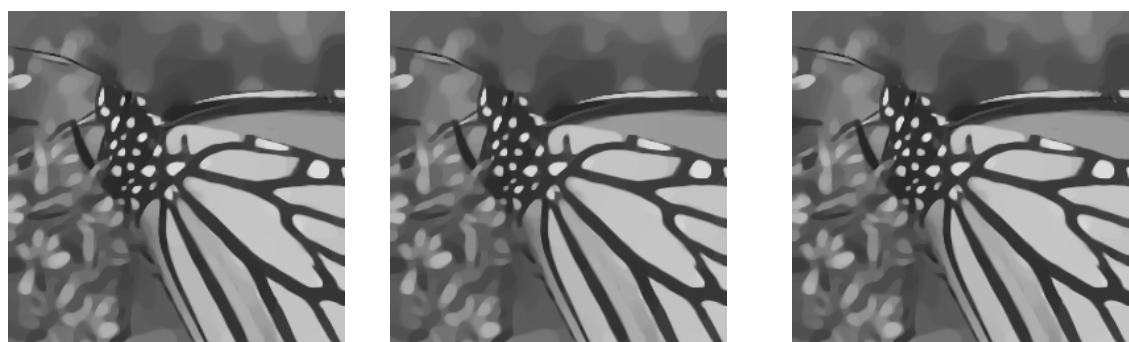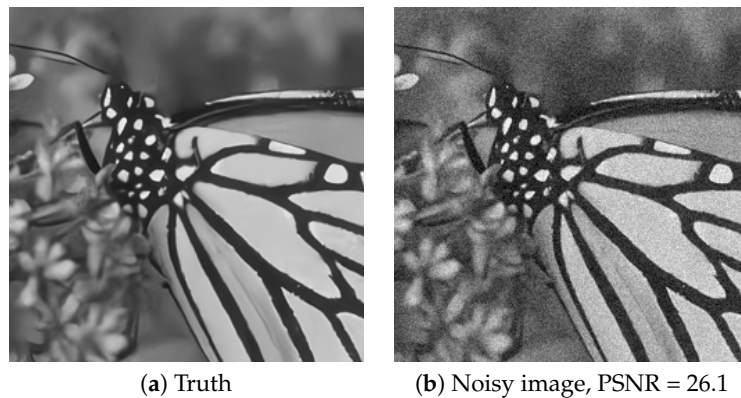
Figure 9. Results on "BIRN" by MIND with TV regularisation and different dictionaries. (**a**) Truth; (**b**) Dyadic cubes, PSNR = 25.2; (**c**) Small cubes, PSNR = 25.8; (**d**) Wavelets, PSNR = 25.8; (**e**) Curvelets, PSNR = 26.3; (**f**) Shearlets, PSNR = 27.8.

(**a**) Truth　　　　　(**b**) Noisy image, PSNR = 26.1

(**c**) MIND with true σ, PSNR=28.9　　　(**d**) MIND with $\hat{\sigma}$ by [35], PSNR = 28　　　(**e**) MIND with $\hat{\sigma}$ by [71], PSNR = 29.1

Figure 10. Results on "butterfly" with unknown noise level: (**c**–**e**) MIND with total variation and shearlets dictionary is shown, with threshold q_n the 50%-quantile of $\|\tilde{\sigma}\epsilon\|_{\mathrm{MS}}$, where $\tilde{\sigma}$ is true σ, estimated $\hat{\sigma}$ by Munk et al. [35] and Liu et al. [71], respectively. (**a**) Truth; (**b**) Noisy image, PSNR = 26.1; (**c**) MIND with true σ; (**d**) MIND with $\hat{\sigma}$ by [35]; (**e**) MIND with $\hat{\sigma}$ by [71].

(**a**) Brain   (**b**) Cell   (**c**) BIRN
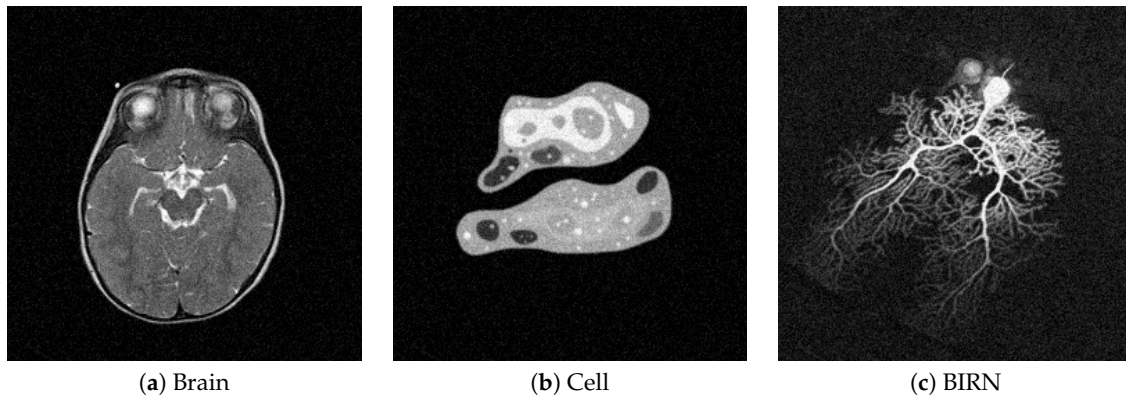
Figure 11. Noisy images of "brain", "cell" and "BIRN" with SNR = 20 and PSNR = 26. (**a**) Brain; (**b**) Cell; (**c**) BIRN.

Table 2. Average performance of MIND with *TV* regularisation and various dictionaries, over 10 repetitions. Values in parenthesis are standard deviations. The best value in each row is in bold.

		Dyadic Cubes	Small Cubes	Wavelets	Curvelets	Shearlets
"brain"	MISE	0.00176 (5.4×10^{-6})	0.00149 (1.9×10^{-6})	0.00147 (2.9×10^{-5})	0.0013 (2.5×10^{-5})	**0.000871** (3×10^{-6})
	PSNR	27.5 (0.013)	28.3 (0.0055)	28.3 (0.085)	28.9 (0.086)	**30.6** (0.015)
	SSIM	0.871 (3.1×10^{-6})	0.806 (0.002)	**0.872** (0.0023)	0.852 (0.0091)	0.715 (0.0043)
	VIF	0.726 (0.00058)	0.823 (0.00036)	0.767 (0.0021)	0.809 (0.0022)	**0.852** (0.00086)
"cell"	MISE	0.000671 (1.3×10^{-6})	0.000438 (8.5×10^{-7})	0.000509 (5×10^{-7})	0.000554 (1.3×10^{-5})	**0.000412** (2.6×10^{-6})
	PSNR	31.7 (0.0082)	33.6 (0.0084)	32.9 (0.0043)	32.6 (0.11)	**33.9** (0.027)
	SSIM	0.912 (0.001)	0.859 (0.008)	**0.924** (8.9×10^{-5})	0.841 (0.018)	0.636 (0.0091)
	VIF	0.86 (8.2×10^{-5})	0.913 (0.0004)	0.884 (0.0004)	0.888 (0.00057)	**0.917** (0.00037)
"BIRN"	MISE	0.00301 (1.6×10^{-6})	0.00266 (1.4×10^{-5})	0.00269 (1×10^{-5})	0.00237 (7.1×10^{-6})	**0.00167** (8.6×10^{-6})
	PSNR	25.2 (0.0023)	25.7 (0.023)	25.7 (0.016)	26.2 (0.013)	**27.8** (0.022)
	SSIM	0.791 (0.00012)	0.802 (0.00053)	0.81 (6.5×10^{-5})	0.82 (0.00075)	**0.858** (0.00076)
	VIF	0.718 (0.0018)	0.811 (0.0018)	0.739 (0.0024)	0.762 (0.00096)	**0.838** (0.00058)

4.3. Unknown Noise Level

We now investigate the issue of unknown noise level σ. This is illustrated by implementing two different estimators for the noise level into the MIND estimator with total variation regularisation and shearlets dictionary. We stress, however, that the results are to be expected to be similar for other choices of the dictionary and regularisation functional. The MIND optimisation is solved by the Chambolle–Pock algorithm. However, in contrast to previous simulations the 50%-quantile of $\|\sigma\epsilon\|_{\mathrm{MS}}$ in (5) with known σ is now replaced by an estimated noise level $\hat{\sigma}$ which then gives a new threshold q_n for MIND. We display all results for the well known test image "butterfly" (256×256 pixels, see, e.g., [72]). The SNR is set to 20, or equivalently $\sigma = 12.6$. The choice of test image and SNR is purely arbitrary, and the results would remain quite the same for other choices. We investigate the influence on MIND for two different noise level estimators. One is a second-order difference-based estimator by Munk et al. [35], which tends to overestimate the noise level, and the other one is a patch based estimator by Liu et al. [71], which often underestimates the noise level (see Figure 12). The comparison results are summarised in Figure 10. Visually, there is little difference between the use of MIND with known noise level and that of the estimated ones. A slight improvement in terms of PSNR is even shown for the use of the estimator that tends to underestimate the noise level. However, a caution should be taken that an underestimated noise level leads to a sacrifice in statistical confidence. For instance, it is no longer possible to guarantee that the true image f lies in the constraint of (3) with given probability 50%. In contrast, when using the true noise level or an overestimated noise level, such a statistical confidence statement is valid. In short, there is nearly no loss of performance of MIND due to the unknown noise level in practice. A heuristic explanation of this is that the mean squared error of the estimated noise level σ scaled as $O(n^{-1})$, whereas for estimating f this rate is slower. Hence, its effect on the overall performance is of minor order.
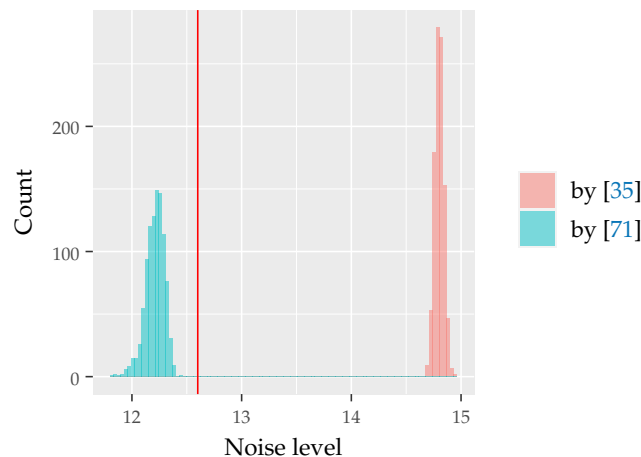
Figure 12. Performance of two noise level estimators by Munk et al. [35] and Liu et al. [71] over 1000 repetitions, with the true σ indicated by a vertical red line, in case of "butterfly" image with SNR = 20.

5. Conclusions and Discussion

5.1. Conclusions

In this paper, we discuss three different algorithms for the computation of variational multiscale estimators (MIND). Among them, we find the Chambolle–Pock algorithm to perform best in practice and to be equally suited for smooth and non-smooth regularisation functionals $R(\cdot)$, the latter case including the prominent TV regularisation functional. Further, there is no need to tune parameters that are related to the Chambolle–Pock algorithm, as the default choice already delivers desirable performance.

Concerning the MIND estimator itself, we find that, for two-dimensional imaging, in nearly all cases it performs best when employing shearlets as the dictionary. However, if more specific a-priori knowledge about the shape of features is available, improvements can be achieved with a dictionary consisting of basis functions that resemble such features. The MIND estimator can be easily modified accordingly. For example, as we demonstrate, in the case of images with bubble-like structures, the choice of cubes as dictionary leads to very promising results.

In contrast to many other statistical recovery methods relying on subtle regularisation parameter choices which are often difficult to choose and interpret, the only relevant parameter for MIND (besides parameters for the numerical optimisation step) is a significance level α between 0 and 1. This has a clear statistical meaning: it guarantees that the estimator is no rougher than the truth in terms of the regularisation functional $R(\cdot)$ with probability at least $1 - \alpha$.

Finally, the unknown noise level can be easily estimated by a difference-based or patch-based estimator, both being easy to compute in $O(n)$ steps.

5.2. Extensions

5.2.1. Bump Signals and Inverse Problems

Even though our findings give a quite satisfactory answer to the question of computability of variational multiscale estimators, some additional questions arise. The first one concerns model (1). We focus on the nonparametric regression model, and in particular on image denoising. Concerning this, we stress that variational multiscale estimators can be applied to other settings, as well. The most obvious one is signal recovery when $d = 1$, and specific structure of signals can be incorporated as well, such as locally constant signals [73], where the number of constant segments is incorporated into the regularisation functional. Although the resulting functional is no longer convex, the MIND estimator can still be computed efficiently exploiting a dynamic program.

A further extensions concerns general linear inverse problems in any dimension d (see, e.g., [33]). In such noisy inverse problems, the only difference concerns the dictionary $\{\phi_\lambda\}$, which has to be chosen in a way akin to the wavelet-vaguelette transform [74]. Otherwise, the structure of the estimator (3) remains unchanged, and in particular the optimisation problem to be solved remains the same. We hence expect the findings of this paper to apply to variational multiscale estimators for inverse problems as well.

5.2.2. Different Noise Models

The second possible extension of the present paper concerns the noise model. We consider Gaussian noise with homogeneous (i.e., not depending on the spatial location x) yet unknown variance. We remark that the extension to heterogeneous variance is of interest as in many applications the variance varies with the signal. To this end, the residuals $Y - f$ in (3) have to be standardised by a local estimator of the variance (see, e.g., [75]).

We finally mention that variational multiscale estimators have been proposed for non-Gaussian noise models as well (see, e.g., [32] for Poisson noise). For non-Gaussian data, the constraint $|\langle \phi_\lambda, g - Y \rangle| \leq q_n$ in (3) is replaced by a constraint on the likelihood-ratio statistic, which offers a route to generalise this further, e.g., to exponential family models (see [76]). In some cases (e.g., Poisson), after a variance stabilising transformation, that constraint can be turned into a linear constraint as in the Gaussian case, so our findings are expected to apply there as well. For general noise, however, deeper study will be required, as nonlinearity in the constraint may lead to nonconvexity of the optimisation problem (3).

References

1. Draper, N.R.; Smith, H. *Applied Regression Analysis*, 3rd ed.; Wiley Series in Probability and Statistics: Texts and References Section; John Wiley & Sons, Inc.: New York, NY, USA, 1998. [CrossRef]
2. Bowman, A.W.; Azzalini, A. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*; OUP Oxford: Oxford, UK, 1997; Volume 18.
3. Fan, J.; Gijbels, I. *Local Polynomial Modelling and Its Applications*; Monographs on Statistics and Applied Probability; CRC Press: Boca Raton, FL, USA, 1996; Volume 66.
4. Stone, C.J. Optimal global rates of convergence for nonparametric regression. *Ann. Stat.* **1982**, *10*, 1040–1053. [CrossRef]
5. Nadaraya, E.A. On estimating regression. *Theory Probab. Appl.* **1964**, *9*, 141–142. [CrossRef]
6. Watson, G.S. Smooth regression analysis. *Sankhya Indian J. Stat. Ser. A* **1964**, *26*, 359–372.
7. Eggermont, P.; LaRiccia, V. Maximum likelihood estimation of smooth monotone and unimodal densities. *Ann. Stat.* **2000**, *28*, 922–947.
8. Phillips, D.L. A technique for the numerical solution of certain integral equations of the first kind. *J. ACM* **1962**, *9*, 84–97. [CrossRef]
9. Morozov, V.A. Regularization of incorrectly posed problems and the choice of regularization parameter. *Zhurnal Vychislitel'noi Mat. I Mat. Fiz.* **1966**, *6*, 170–175. [CrossRef]
10. Rudin, L.I.; Osher, S.; Fatemi, E. Nonlinear total variation based noise removal algorithms. *Phys. D Nonlinear Phenom.* **1992**, *60*, 259–268. [CrossRef]

11. Daubechies, I. *Ten Lectures on Wavelets*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1992; Volume 61.

12. Donoho, D.L. De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* **1995**, *41*, 613–627. [CrossRef]

13. Tsybakov, A.B. *Introduction to Nonparametric Estimation*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.

14. Nemirovski, A. Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk. SSR Teckhn. Kibernet* **1985**, *3*, 50–60.

15. Candès, E.J.; Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **2007**, *35*, 2313–2351. [CrossRef]

16. Grasmair, M.; Li, H.; Munk, A. Variational multiscale nonparametric regression: Smooth functions. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*; Institut Henri Poincaré: Paris, France, 2018; Volume 54, pp. 1058–1097.

17. Scherzer, O.; Grasmair, M.; Grossauer, H.; Haltmeier, M.; Lenzen, F. *Variational Methods in Imaging*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.

18. Burger, M.; Sawatzky, A.; Steidl, G. First Order Algorithms in Variational Image Processing. In *Splitting Methods in Communication, Imaging, Science, and Engineering*; Glowinski, R., Osher, S.J., Yin, W., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 345–407. [CrossRef]

19. Hintermüller, M.; Rincon-Camacho, M. An adaptive finite element method in L^2-TV-based image denoising. *Inverse Probl. Imaging* **2014**, *8*, 685–711. [CrossRef]

20. Hintermüller, M.; Papafitsoros, K.; Rautenberg, C.N. Analytical aspects of spatially adapted total variation regularisation. *J. Math. Anal. Appl.* **2017**, *454*, 891–935. [CrossRef]

21. Hintermüller, M.; Langer, A.; Rautenberg, C.N.; Wu, T. Adaptive Regularization for Image Reconstruction from Subsampled Data. In *Imaging, Vision and Learning Based on Optimization and PDEs*; Tai, X.C., Bae, E., Lysaker, M., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–26.

22. Dong, Y.; Hintermüller, M.; Rincon-Camacho, M.M. A multi-scale vectorial L^τ-TV framework for color image restoration. *Int. J. Comput. Vis.* **2011**, *92*, 296–307. [CrossRef]

23. Dong, Y.; Hintermüller, M.; Rincon-Camacho, M.M. Automated regularization parameter selection in multi-scale total variation models for image restoration. *J. Math. Imaging Vis.* **2011**, *40*, 82–104. [CrossRef]

24. Lenzen, F.; Berger, J. Solution-Driven Adaptive Total Variation Regularization. In *Scale Space and Variational Methods in Computer Vision*; Aujol, J.F., Nikolova, M., Papadakis, N., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 203–215.

25. Donoho, D.L.; Johnstone, J.M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **1994**, *81*, 425–455. [CrossRef]

26. Candès, E.J.; Donoho, D.L. *Curvelets: A Surprisingly Effective Nonadaptive Representation for Objects with Edges*; Technical Report; Department of Statistics Stanford University: Stanford, CA, USA, 2000.

27. Labate, D.; Lim, W.Q.; Kutyniok, G.; Weiss, G. Sparse multidimensional representation using shearlets. In *Wavelets XI*; International Society for Optics and Photonics: Bellingham, WA, USA, 2005; Volume 5914, p. 59140U.

28. Candès, E.J.; Guo, F. New multiscale transforms, minimum total variation synthesis: Applications to edge-preserving image reconstruction. *Signal Process.* **2002**, *82*, 1519–1543. [CrossRef]

29. Del Alamo, M.; Li, H.; Munk, A. Frame-constrained total variation regularization for white noise regression. *arXiv* **2020**, arXiv:1807.02038.

30. Malgouyres, F. Mathematical analysis of a model which combines total variation and wavelet for image restoration. *J. Inf. Process.* **2002**, *2*, 1–10.

31. Frick, K.; Marnitz, P.; Munk, A. Statistical multiresolution Dantzig estimation in imaging: Fundamental concepts and algorithmic framework. *Electron. J. Stat.* **2012**, *6*, 231–268. [CrossRef]

32. Frick, K.; Marnitz, P.; Munk, A. Statistical multiresolution estimation for variational imaging: With an application in Poisson-biophotonics. *J. Math. Imaging Vis.* **2013**, *46*, 370–387. [CrossRef]

33. Del Álamo, M.; Munk, A. Total variation multiscale estimators for linear inverse problems. *Inf. Inference J. IMA* **2019**. [CrossRef]

34. Plotz, T.; Roth, S. Benchmarking denoising algorithms with real photographs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1586–1595.

35. Munk, A.; Bissantz, N.; Wagner, T.; Freitag, G. On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 19–41. [CrossRef]

36. Frick, K.; Marnitz, P.; Munk, A. Shape-constrained regularization by statistical multiresolution for inverse problems: Asymptotic analysis. *Inverse Probl.* **2012**, *28*, 065006. [CrossRef]

37. Lebert, J.; Künneke, L.; Hagemann, J.; Kramer, S.C. Parallel Statistical Multi-resolution Estimation. *arXiv* **2015**, arXiv:1503.03492.

38. Kramer, S.C.; Hagemann, J.; Künneke, L.; Lebert, J. Parallel statistical multiresolution estimation for image reconstruction. *SIAM J. Sci. Comput.* **2016**, *38*, C533–C559. [CrossRef]

39. Morken, A.F. An algorithmic Framework for Multiresolution Based Non-Parametric Regression. Master's Thesis, NTNU, Trondheim, Norway, 2017.

40. Luke, D.R.; Shefi, R. A globally linearly convergent method for pointwise quadratically supportable convex-concave saddle point problems. *J. Math. Anal. Appl.* **2018**, *457*, 1568–1590. [CrossRef]

41. Chambolle, A.; Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.* **2011**, *40*, 120–145. [CrossRef]

42. Hintermüller, M. Semismooth Newton Methods and Applications. In *Lecture Notes for the Oberwolfach-Seminar on "Mathematics of PDE-Constrained Optimization"*; Department of Mathematics, Humboldt-University of Berlin: Berlin, Germany, 2010.

43. Clason, C.; Kruse, F.; Kunisch, K. Total variation regularization of multi-material topology optimization. *ESAIM Math. Model. Numer. Anal.* **2018**, *52*, 275–303. [CrossRef]

44. Lepskii, O.V. On a Problem of Adaptive Estimation in Gaussian White Noise. *Theory Probab. Appl.* **1991**, *35*, 454–466. [CrossRef]

45. Donoho, D.L.; Johnstone, I.M.; Kerkyacharian, G.; Picard, D. Wavelet shrinkage: Asymptopia? *J. R. Stat. Soc. Ser. B* **1995**, *57*, 301–369. With discussion and a reply by the authors. [CrossRef]

46. Donoho, D.L.; Johnstone, I.M. Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* **1995**, *90*, 1200–1224. [CrossRef]

47. Weyrich, N.; Warhola, G.T. Wavelet shrinkage and generalized cross validation for image denoising. *IEEE Trans. Image Process.* **1998**, *7*, 82–90. [CrossRef] [PubMed]

48. Härdle, W.; Kerkyacharian, G.; Picard, D.; Tsybakov, A. *Wavelets, Approximation, and Statistical Applications*; Lecture Notes in Statistics; Springer: New York, NY, USA, 1998; Volume 129, p. xviii+265. [CrossRef]

49. Cai, T.T. On block thresholding in wavelet regression: Adaptivity, block size, and threshold level. *Stat. Sin.* **2002**, *12*, 1241–1273.

50. Zhang, C.H. General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Stat.* **2005**, *33*, 54–100. [CrossRef]

51. Abramovich, F.; Benjamini, Y.; Donoho, D.L.; Johnstone, I.M. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Stat.* **2006**, *34*, 584–653. [CrossRef]

52. Cai, T.T.; Zhou, H.H. A data-driven block thresholding approach to wavelet estimation. *Ann. Stat.* **2009**, *37*, 569–595. [CrossRef]

53. Haltmeier, M.; Munk, A. Extreme value analysis of empirical frame coefficients and implications for denoising by soft-thresholding. *Appl. Comput. Harmon. Anal.* **2014**, *36*, 434–460. [CrossRef]

54. Rockafellar, R.T. *Convex Analysis*; Princeton University Press: Princeton, NJ, USA, 2015.

55. Ekeland, I.; Témam, R. *Convex Analysis and Variational Problems*, english ed.; Volume 28, Classics in Applied Mathematics; Society for Industrial and Applied Mathematics (SIAM): Philadelphia, PA, USA, 1999; p. xiv+402. [CrossRef]

56. Nesterov, Y.; Nemirovsky, A. *Interior-Point Polynomial Algorithms in Convex Programming*; Society for Industrial and Applied Mathematics (SIAM): Philadelphia, PA, USA, 1994; p. ix+396. [CrossRef]

57. Chambolle, A. An Algorithm for Total Variation Minimization and Applications. *J. Math. Imaging Vis.* **2004**, *20*, 89–97.

58. Powell, M.J.D. A method for nonlinear constraints in minimization problems. In *Optimization (Sympos., Univ. Keele, Keele, 1968)*; Academic Press: London, UK, 1969; pp. 283–298.

59. Hestenes, M.R. Multiplier and gradient methods. *J. Optim. Theory Appl.* **1969**, *4*, 303–320. [CrossRef]

60. Dykstra, R.L. An algorithm for restricted least squares regression. *J. Am. Stat. Assoc.* **1983**, *78*, 837–842. [CrossRef]

61. Boyle, J.P.; Dykstra, R.L. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in Order Restricted Statistical Inference*; Springer: Berlin/Heidelberg, Germany, 1986; pp. 28–47.

62. Deutsch, F.; Hundal, H. The rate of convergence of Dykstra's cyclic projections algorithm: The polyhedral case. *Numer. Funct. Anal. Optim.* **1994**, *15*, 537–565. [CrossRef]

63. Birgin, E.G.; Raydan, M. Robust stopping criteria for Dykstra's algorithm. *SIAM J. Sci. Comput.* **2005**, *26*, 1405–1414. [CrossRef]

64. Deng, W.; Yin, W. On the global and linear convergence of the generalized alternating direction method of multipliers. *J. Sci. Comput.* **2016**, *66*, 889–916. [CrossRef]

65. Hintermüller, M.; Kunisch, K. Path-following methods for a class of constrained minimization problems in function space. *SIAM J. Optim.* **2006**, *17*, 159–187. [CrossRef]

66. Hore, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In Proceedings of the IEEE 2010 20th International Conference on Pattern Recognition, New York, NY, USA, 23–26 July 2010; pp. 2366–2369.

67. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]

68. Sheikh, H.R.; Bovik, A.C. Image information and visual quality. *IEEE Trans. Image Process.* **2006**, *15*, 430–444. [CrossRef]

69. Pock, T.; Chambolle, A. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In Proceedings of the IEEE 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1762–1769.

70. Giewekemeyer, K.; Krueger, S.P.; Kalbfleisch, S.; Bartels, M.; Salditt, T.; Beta, C. X-ray propagation microscopy of biological cells using waveguides as a quasipoint source. *Phys. Rev. A* **2011**, *83*. [CrossRef]

71. Liu, X.; Tanaka, M.; Okutomi, M. Single-image noise level estimation for blind denoising. *IEEE Trans. Image Process.* **2013**, *22*, 5226–5237. [CrossRef]

72. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef]

73. Frick, K.; Munk, A.; Sieling, H. Multiscale change point inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2014**, *76*, 495–580. With 32 discussions by 47 authors and a rejoinder by the authors. [CrossRef]

74. Donoho, D.L. Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **1995**, *2*, 101–126. [CrossRef]

75. Brown, L.D.; Levine, M. Variance estimation in nonparametric regression via the difference sequence method. *Ann. Stat.* **2007**, *35*, 2219–2232. [CrossRef]

76. König, C.; Munk, A.; Werner, F. Multidimensional multiscale scanning in exponential families: Limit theory and statistical consequences. *Ann. Stat.* **2020**, *48*, 655–678. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.