Bilevel Image Denoising using Gaussianity tests

Jérôme Fehrenbach, Mila Nikolova, Gabriele Steidl, and Pierre Weiss

Université de Toulouse, CNRS, IMT (UMR5219) and ITAV (USR 3505),
jerome.fehrenbach@math.univ-toulouse.fr
ENS Cachan, CNRS, CMLA, nikolova@cmla.ens-cachan.fr
University of Kaiserslautern, steidl@mathematik.uni-kl.de
Université de Toulouse, CNRS, IMT (UMR5219) and ITAV (USR 3505),
pierre.armand.weiss@gmail.com

Abstract. We propose a new methodology based on bilevel programming to remove additive white Gaussian noise from images. The lower-level problem consists of a parameterized variational model to denoise images. The parameters are optimized in order to minimize a specific cost function that measures the residual Gaussianity. This model is justified using a statistical analysis. We propose an original numerical method based on the Gauss-Newton algorithm to minimize the outer cost function. We finally perform a few experiments that show the well-foundedness of the approach. We observe a significant improvement compared to standard TV-ℓ^2 algorithms and show that the method automatically adapts to the signal regularity.

Keywords: Bilevel programming, image denoising, Gaussianity tests, convex optimization.

1 Introduction

In this paper, we consider the following simple image formation model:

$$u_b = u_c + b \tag{1}$$

where $u_c \in \mathbb{R}^n$ denotes a clean image, $b \in \mathbb{R}^n$ is a white Gaussian noise of variance σ^2 and $u_b \in \mathbb{R}^n$ is the noisy image. Our aim is to denoise u_b, i.e. to retrieve an approximation of u_c knowing u_b.

1.1 Variational denoising

The standard way to achieve image restoration using variational methods consists in solving an optimization problem of the form

$$\text{Find } u^*(\alpha) = \operatorname*{arg\,min}_{u \in \mathbb{R}^n} \alpha R(u) + \frac{1}{2\sigma^2}\|u - u_b\|_2^2, \tag{2}$$

where α is a regularization parameter and $R : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a regularizing term such as total variation (TV) [11] or alternative priors. This approach can be justified using a Bayesian point of view, assuming that images are random vectors with a density $\mathbb{P}(u) \propto \exp\left(-\alpha R(u)\right)$ and using the fact that b is a white Gaussian noise. This reasoning is widespread in the imaging community since the seminal paper [4]. Despite its success in applications, it suffers from serious drawbacks. First, it is now well known that Bayesian estimators strongly deviate from the data and noise models [1, 10]. Second, one needs to design a probability density function that describes the set of natural images. This task is extremely hard and simple models (e.g. based on total variation) are very unlikely to correctly describe the density of natural images. This problem is studied and discussed thoroughly in [9]. As a consequence, denoising models such as (2) are only partially satisfactory and the residuals $b^*(\alpha) = u_b - u^*(\alpha)$ obtained by solving (2) are usually non-white. This is illustrated in Figure 1.

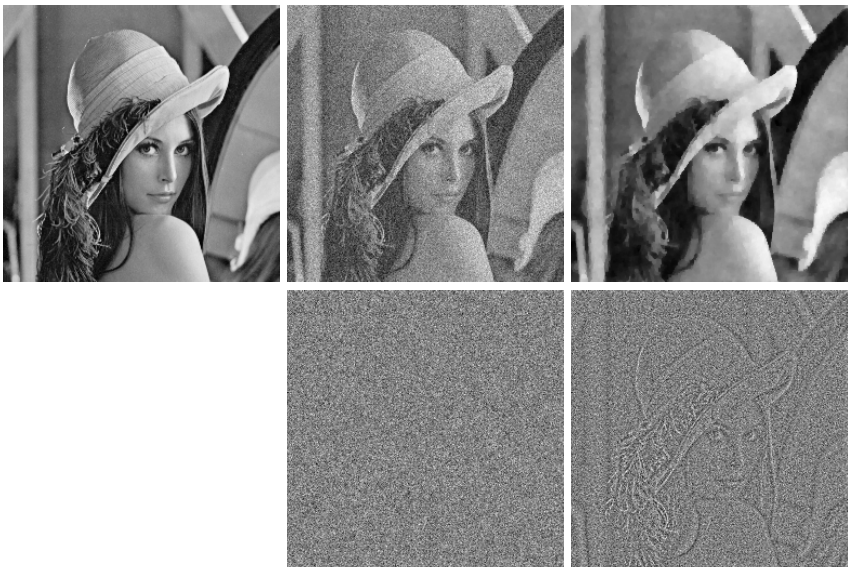

Fig. 1. An example of TV-ℓ^2 denoising. Top-left: original image. Top-mid: noisy image. Top-right: denoising result. Bottom-mid: noise. Bottom-right: retrieved residual $u^*(\alpha) - u_b$. The residual contains a lot of structure, showing the limits of this approach.

In this work, we depart from the standard setting (2). Our starting observation is that in many applications, one has a quite good knowledge of the noise properties and only a very rough idea of the image contents. The data and regularization terms should thus play modified roles: the regularization should be adaptive to the image contents while the data term should measure Gaussianity in a more efficient way than the standard ℓ^2-norm. This idea is not new and led to state-of-the-art results in wavelet thresholding based methods [7].

1.2 The proposed framework

Instead of fixing the regularization term, we propose to use a parametric restoration model and to optimize the parameters with a bilevel programming approach, in order to make the residual "look" Gaussian. This idea is close in spirit to the recent work [5] and very different from the simple model (2), where the regularization term R is fixed and the Gaussianity measure is just the ℓ^2-norm.

We propose using a parameterized model of type:

$$u^*(\alpha) = \operatorname*{arg\,min}_{u \in \mathbb{R}^n} \sum_{i=1}^{p} \alpha_i \phi_i(R_i u) + \frac{1}{2\sigma^2}\|u - u_b\|_2^2, \tag{3}$$

where

- $\alpha = (\alpha_i)_{i=1}^p$ is a non negative vector of regularization parameters,
- $\phi_i : \mathbb{R}^{m_i} \to \mathbb{R}$, $i \in \{1, \dots, p\}$ are C^2 symmetric functions (typically smoothed l^1-norms),
- $R_i \in \mathbb{R}^{m_i \times n}$ are known analysis-based operators.

Model (3) thus encompasses total variation like regularization. It is however more flexible since the vector of parameters α can be chosen differently depending on the image contents. Since the residual $b^*(\alpha) = u_b - u^*(\alpha)$ plays an important role in this paper, we use the change of variable $b = u_b - u$ and denote

$$J_\alpha(b) := \sum_{i=1}^{p} \alpha_i \phi_i(R_i(u_b - b)) + \frac{1}{2}\|b\|_2^2.$$

Let $G : \mathbb{R}^n \to \mathbb{R}$ denote a $\mathcal{C}^1$ function that measures noise Gaussianity. The proposed denoising model consists in finding $\alpha^* \in \mathbb{R}_+^p$ and $b^*(\alpha^*) \in \mathbb{R}^n$ solutions of the following bi-level programming problem:

$$\begin{cases} \min\limits_{\alpha \geq 0} g(\alpha) := G(b^*(\alpha)) \\ \text{with } b^*(\alpha) = \operatorname*{arg\,min}\limits_{b \in \mathbb{R}^n} J_\alpha(b). \end{cases} \tag{4}$$

The lower-level problem $\min\limits_{b \in \mathbb{R}^n} J_\alpha(b)$ corresponds to a denoising step with a fixed regularization vector, while the upper-level problem corresponds to a parameter optimization.

1.3 Contributions of the paper

The first contribution of this paper is the variational formulation (4) with a new cost function $g(\alpha)$ (derived in Section 2). This function is motivated by a statistical analysis of white Gaussian noise properties. The bilevel problem (4) shares a connection with [5] and was actually motivated by this paper. In [5], the authors propose to *learn* the parameters using an image database, while our method simply uses the noisy image, making the parameter estimation self-contained.

Moreover, the proposed methodology makes our algorithm *auto-adaptive* to the image contents, meaning that the denoising model adapts to the type of image to denoise.

The second contribution of this paper is the design of an optimization method based on Gauss-Newton's algorithm in Section 3. The preliminary numerical experiments we performed suggest that it is very efficient, while being simpler to implement than the semi-smooth Newton based method proposed in [5].

Finally, we present preliminary denoising experiments in Section 4, showing the well-foundedness of the proposed approach.

2 Measuring Residual Gaussianity

In this section, we propose a function g that measures the residuals Gaussianity and whiteness and allows identifying the vector $\alpha \in \mathbb{R}^p$.

2.1 The case $p = 1$

In this paper, we assume that the discrete image domain Ω satisfies $|\Omega| = n$. To expose our ideas, let us begin with the simple case where only one regularizer is used, i.e. $p = 1$. A basic idea to select the regularization parameter is to find α such that $\|b^*(\alpha)\|_2^2 \simeq \sigma^2 n$, since $\mathbb{E}(\|b\|_2^2) = \sigma^2 n$. One could thus set $g(\alpha) = \frac{1}{2}\left(\|b^*(\alpha)\|_2^2 - \sigma^2 n\right)^2$. This idea is similar to Morozov's discrepancy principle [8].

This simple method is however unlikely to provide satisfactory results with more than 1 regularizer (i.e. $p > 1$), since many vectors $\alpha \in \mathbb{R}_+^p$ may lead to $\|b^*(\alpha)\|_2^2 = \sigma^2 n$. Said differently, the function g here does not allow identifying a unique α since there are two many degrees of freedom in the model. Moreover, the accurate knowledge of the noise distribution b is boiled down to a simple scalar corresponding to the mean of the ℓ^2-norm. Our aim below is therefore to construct measures of Gaussianity allowing to identify the parameters and to better characterize the noise distribution.

2.2 The case $p > 1$

The idea proposed in the case of a single parameter can be generalized by defining a set of q Euclidean semi-norms $(\|\cdot\|_{M_i}^2)_{i=1}^q$. These semi-norms are defined by

$$\|x\|_{M_i}^2 := \|M_i x\|_2^2,$$

where $M_i \in \mathbb{R}^{m_i \times n}$. Let $b \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$ be white Gaussian noise with $\mu_i = \mathbb{E}(\|b\|_{M_i}^2)$ and $v_i = \mathrm{Var}(\|b\|_{M_i}^2)$. A natural idea to extend the principle presented in Subsection 2.1 consists in setting

$$g(\alpha) := \frac{1}{2}\sum_{i=1}^{q} \frac{(\|b^*(\alpha)\|_{M_i}^2 - \mu_i)^2}{v_i}. \tag{5}$$

This choice can be justified using a maximum likelihood approach. In cases where n is large enough and where the singular values of M_i are sufficiently spread out, the distribution of $\|b\|_{M_i}^2$ is well approximated by a normal distribution $\mathcal{N}(\mu_i, v_i)$. The probability density function of $\|b\|_{M_i}^2$ approximately satisfies

$$f_{M_i}(b) \propto \exp\left(-\frac{\left(\|b\|_{M_i}^2 - \mu_i\right)^2}{2v_i}\right).$$

The random variables $(\|b\|_{M_i}^2)_{i=1}^q$ are not independent in general. However, if the matrices M_i are chosen in such a way that they measure different noise properties (e.g. different frequencies components), the likelihood of the random vector $(\|b\|_{M_i}^2)_{i=1}^q$ is approximately equal to

$$f(b) \propto \mathop{\mathit{\Pi}}_{i=1}^{q} f_{M_i}(b). \tag{6}$$

Using a maximum likelihood approach to set the parameter α leads to minimizing $-\log(f(b^*(\alpha)))$, i.e. to set g as in equation (5).

2.3 The choice of M_i

In this paper, we propose to analyse residuals using Fourier decompositions: we construct a partition $\Omega = \cup_{i=1}^q \Omega_i$ of the discrete Fourier domain and set $M_i = F\mathrm{diag}(\mathbf{1}_{\Omega_i})F^*$, where F denotes the discrete Fourier transform and $\mathbf{1}_{\Omega_i}$ denotes a vector equal to 1 on Ω_i and 0 elsewhere. In other words the matrices M_i correspond to discrete convolutions with filters $\varphi_i = F\mathbf{1}_{\Omega_i}$. For this specific choice, it is quite easy to show that

$$\mu_i = n\sigma^2\|\varphi_i\|^2$$

and that

$$v_i = n\sigma^4\|\varphi_i\|^4.$$

Moreover, the random variables $\|b\|_{M_i}^2$ are independent. Therefore, the likelihood (6) is a good approximation of the random vector $(\|b\|_{M_i}^2)_{i=1}^q$ as soon as the cardinals $|\Omega_i|$ are sufficiently large, due to the central limit theorem.

The rationale behind a partition of the Fourier domain is that residuals containing image structures usually exhibit anormal structured spectra. This phenomenon is illustrated in Figure 2. The Fourier transform of white Gaussian noise is still white Gaussian noise. Therefore, if the residual was "correct", its Fourier transforms should "look" white. The spectrum of a residual obtained using a TV-ℓ^2 minimization (Figure 2, middle) is clearly not white. In particular, the modulus of its Fourier transform is too low in the center of the frequency domain. On the contrary, it is too large on directions orthogonal to the main components of the image: the vertical stripes and the diagonal elements of Lena's hat.

In this paper, we propose to define the sets Ω_i similarly to frequency tilings of curvelet or shearlet transforms [2, 6]. This is illustrated in Figure 2, right. Each set Ω_i corresponds to the union of a trapezoid and its symmetric with respect to the origin. Its size and angular resolution increases in a dyadic way with the frequencies.

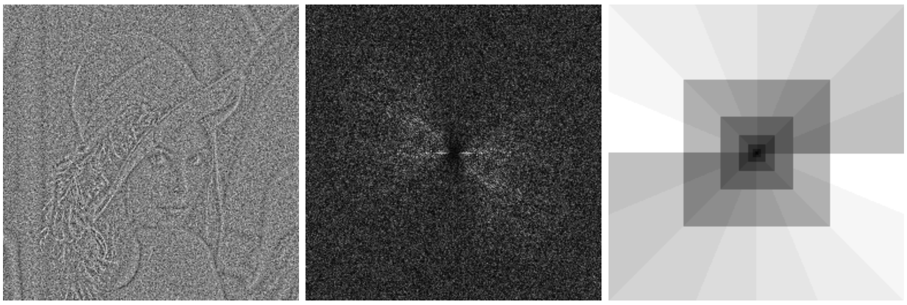

Fig. 2. Analysis of residuals in the Fourier domain. Left: residual of a TV-ℓ^2 minimization. Middle: discrete Fourier transform modulus of the residual. This modulus should be an i.i.d. sequence with constant mean. It exhibits a lot of structure, especially in the low frequencies. Right: frequency tiling proposed to analyse the spectrum.

In order to assess whether a residual is likely to correspond to white Gaussian noise, we will make use of the standard score (or z-score) defined by

$$z_i = \frac{\|b\|_{M_i}^2 - \mu_i}{\sqrt{v_i}}.$$

This score measures the (signed) number of standard deviations $\|b\|_{M_i}^2$ is above the mean. For sufficiently large n (which is typical for contemporary pictures), $\|b\|_{M_i}^2$ can be assimilated to a Gaussian random variable and therefore $\mathbb{P}(|z_i| \geq k) \simeq 1 - \operatorname{erf}\left(\frac{k}{\sqrt{2}}\right)$. The values are displayed in Table 1. As can be seen in this table, it is extremely unlikely that $|z_i|$ be larger than 3. Using the frequency tiling proposed in Figure 2, composed of 45 tiles, we get a maximum z-score $\max\limits_{i\in\{1,\cdots,q\}} |z_i| = 30.3$ and a mean z-score of 6.0. By looking at Table 1, it is clear that such a residual is *extremely* unlikely to correspond to white Gaussian noise.

k	0	1	2	3	4	5	6	7
$P(\lvert z_i\rvert \geq k)$	1	$3.2 \cdot 10^{-1}$	$4.6 \cdot 10^{-2}$	$2.7 \cdot 10^{-3}$	$6.3 \cdot 10^{-5}$	$5.7 \cdot 10^{-7}$	$2.0 \cdot 10^{-9}$	$2.6 \cdot 10^{-12}$

Table 1. Probability that a standard normally distributed random variable deviates from its mean more than k times its standard deviation.

3 A Bilevel Programming Approach Based on a Gauss-Newton Algorithm

In this section, we describe a numerical method based on the Gauss-Newton algorithm to solve problem (4) with

$$g(\alpha) = \sum_{i=1}^{q} \frac{\left(\|\varphi_i \star b^*(\alpha)\|^2 - \mu_i\right)^2}{2v_i}.$$

The solution of bilevel programs of type (4) is a well studied problem with many different solution algorithms, see, e.g., the monograph [3]. Bilevel problems are usually NP-hard so that only local minima can be expected. Similarly to standard optimization, there exists multiple algorithms which should be chosen depending on the context (problem dimension, lower and upper-level problem regularity, convexity,...). In this paper, we suggest using the following combination:

- Handle the positivity constraint $\alpha_i \geq 0$ by writing $\alpha = \exp(\beta)$, allowing to have an unconstrained minimization problem with parameter β.
- Use the implicit function theorem to estimate the Jacobian $\mathrm{Jac}_{b^*}(\alpha)$ (i.e. the first order variations of b^* w.r.t. α).
- Use this information to design a Gauss-Newton algorithm.

The advantage of the Gauss-Newton algorithm is that it usually converges much faster than gradient descent methods since the metric adapts to the local function curvatures. It is also much simpler to use than the semi-smooth approach suggested in [5] while still showing a very good performance.

The change of variable $\alpha = \exp(\beta)$ ensures that $\alpha > 0$ without bringing any extra difficulty in the design of the numerical algorithm since the chain rule allows a straightforward modification. More precisely we aim at minimizing

$$h(\beta) = g(\exp(\beta)),$$

and we use the following identity:

$$Dh(\beta) = Dg(\exp(\beta))\Sigma,$$

where Σ is the diagonal matrix with entries $\exp(\beta_i)$. Next, remark that function h can be rewritten as

$$h(\beta) = \frac{1}{2}\|f(\beta)\|_2^2 = \frac{1}{2}\|F(b^*(\exp(\beta)))\|_2^2 \tag{7}$$

with

$$F(b) := \begin{pmatrix} F_1(b) \\ \vdots \\ F_q(b) \end{pmatrix}, \qquad f(\beta) := \begin{pmatrix} f_1(\exp(\beta)) \\ \vdots \\ f_q(\exp(\beta)) \end{pmatrix},$$

$$f_i(\alpha) := F_i(b^*(\alpha)) \qquad \text{and} \qquad F_i(b) := \frac{\|\varphi_i \star b\|_2^2 - \mu_i}{\sqrt{2}v_i}. \tag{8}$$

Then the k-th iteration of the Gauss-Newton algorithm adapted to functions of type (7) reads as follows:

1. Set
$$d^k = \arg\min_{d \in \mathbb{R}^p} \|f(\beta^k) + \mathrm{Jac}_f(\beta^k) d\|_2^2 \tag{9}$$
2. Set
$$\beta^{k+1} = \beta^k + d^k \tag{10}$$

The descent direction d^k computed in (9) satisfies

$$\mathrm{Jac}_f(\beta^k)^T \mathrm{Jac}_f(\beta^k) d^k = -\mathrm{Jac}_f(\beta^k)^T f(\beta^k).$$

The lower-level problem in (4) is solved using an accelerated proximal gradient descent algorithm on the dual of (3), see, e.g., [12]. We do not detail further this algorithm for lack of space.

4 Numerical results

4.1 A test example

To begin with, we perform a simple denoising experiment to validate the overall principle and the numerical algorithm. We consider the following simple denoising model:

$$\min_{u \in \mathbb{R}^n} \alpha_1 \phi(\partial_x u) + \alpha_2 \phi(\partial_y u) + \frac{1}{2}\|u - u_b\|_2^2, \tag{11}$$

where $\phi(x) := \sqrt{x^2 + \epsilon^2}$ is an approximation of the ℓ^1-norm, ∂_x and ∂_y are first order difference operators in the x and y directions, respectively. We use the smooth 64×64 images which are constant along the x or y axes in Figure 3. The algorithm is initialized with $\alpha = (1, 1)$. After 20 iterations of our Gauss-Newton algorithm, the regularization parameters become $\alpha = (186.3, 0.03)$ for the image constant in the x-direction and $\alpha = (0.03, 155.11)$ for the image constant in the y-direction. This choice basically corresponds to a very strong regularization in the direction of the level lines of the image: the method is capable of automatically detecting the smoothness directions.

We compare the output of our algorithm with a TV-ℓ^2 model, where the regularization coefficient is chosen in order to maximize the mean square error (hence the choice of this optimal coefficient requires the knowledge of the ground truth image).

Compared to the TV-ℓ^2 model, the denoising results are significantly better. In particular, no structure can be found in the residual of the proposed method, while a lot of structure is apparent in the residual of the TV-ℓ^2 model. The maximum z-score is 88.2 for the TV-ℓ^2 algorithm and 1.8 for the bilevel approach.

Regarding the numerical behavior, even though we performed 20 iterations, a satisfactory and stable solution is found after just 6 iterations of our Gauss-Newton algorithm. The cost function with respect to the iteration number is displayed in Figure 4.

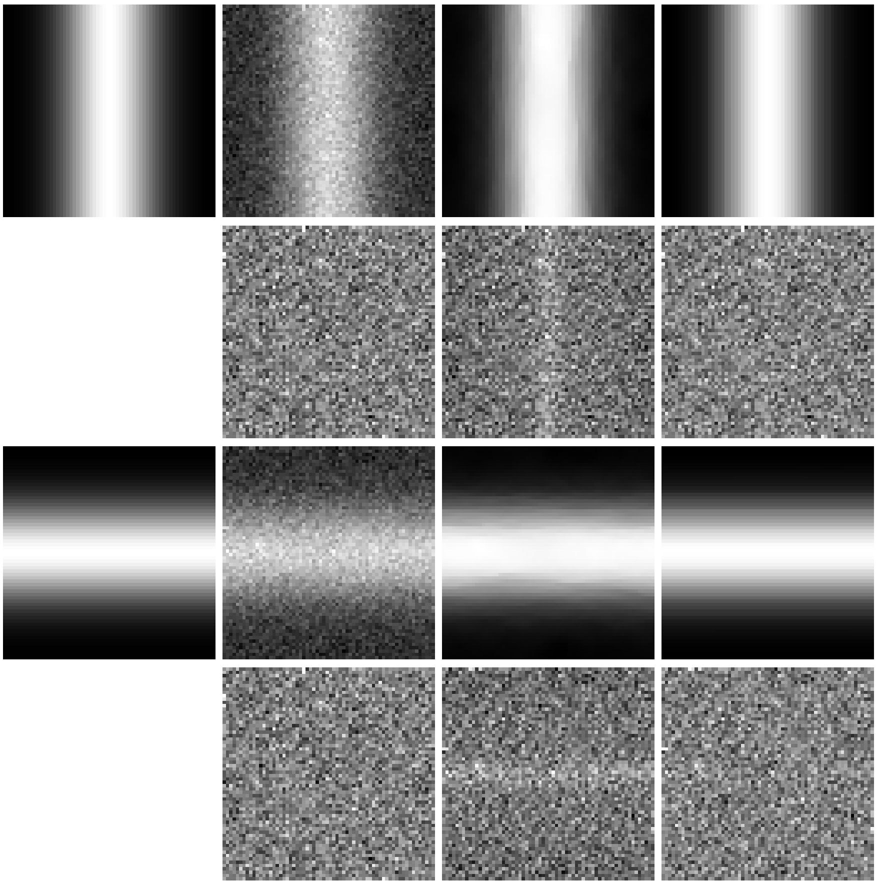

Fig. 3. Denoising results for a toy example. First and third rows, from left to right: original image, noisy image (PSNR=14.7dB), denoised with TV (PSNR=27.9dB), denoised with the bilevel approach (PSNR=34.9dB). Second and fourth rows: residuals associated to the top images.

4.2 A real-world denoising experiment

We now turn to a real denoising example of Lena. Similarly to [5], the transforms R_i are set as convolution products with the 25 elements of the discrete cosine transform basis on 5×5 windows. The number of elements of the Fourier domain partition is 50. The results are presented in Figure 5. The bilevel denoising result is significantly better (1.5dB) than the standard TV result. The z-test indicates that the TV residual is extremely unlikely to correspond to white Gaussian noise. It also indicates that the bilevel residual is unlikely. This result suggests that much better denoising results could be expected by considering different parameterized denoising models.

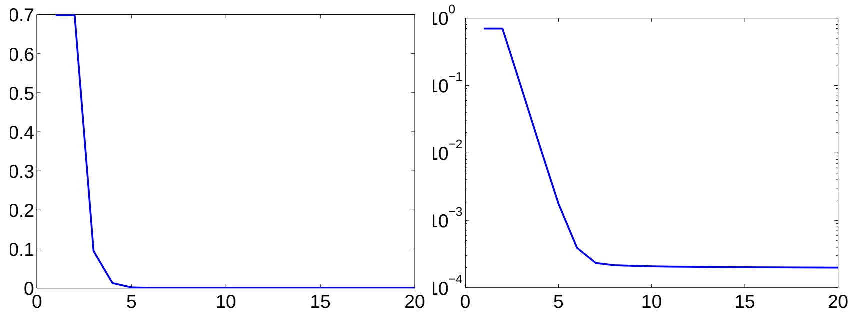

Fig. 4. Function $g(\alpha_k)$ with respect to k. Left: standard scale. Right: log_{10} scale. The cost function reaches a plateau after 6 iterations.

5 Conclusion & Outlook

In this work, we explored the use of bilevel programming to choose an optimal parameterized denoising model by measuring the Gaussianity of the residuals. The results are encouraging and provide significantly better results than standard variational models. They are probably not comparable to state-of-the-art methods based on nonlocal means or BM3D for instance both in terms of restoratin quality and computing times.

We still believe that this approach has a great potential in applications since i) the method can be adapted to arbitrary inverse problems and ii) the method is capable of automatically finding the class of regularity of the considered signals. This is a very nice feature that is absent in most current approaches.

Finally, let us mention that the considered parameterized denoising models can probably be improved significantly by considering not only adapting to the global regularity of signals, but also to the local regularity. To achieve this, the operators R_i should be localized in space. We plan to investigate this issue in our forthcoming work.

References

1. F. Baus, M. Nikolova, and G. Steidl. Smooth objectives composed of asymptotically affine data-fidelity and regularization: Bounds for the minimizers and parameter choice. *Journal of Mathematical Imaging and Vision*, 48(2):295–307, 2013.
2. Emmanuel Candes, Laurent Demanet, David Donoho, and Lexing Ying. Fast discrete curvelet transforms. *Multiscale Modeling & Simulation*, 5(3):861–899, 2006.
3. Stephan Dempe. *Foundations of Bilevel Programming*. Springer, 2002.
4. Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.
5. Karl Kunisch and Thomas Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013.

6. Demetrio Labate, Wang-Q Lim, Gitta Kutyniok, and Guido Weiss. Sparse multidimensional representation using shearlets. In Manos Papadakis, Andrew F. Laine, and Michael A. Unser, editors, *Proceedings of Wavelets XI*, volume 5914 of *Proc. SPIE*, San Diego, 2005.
7. Florian Luisier, Thierry Blu, and Michael Unser. A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding. *IEEE Transactions on Image Processing*, 16(3):593–606, 2007.
8. Vladimir Alekseevich Morozov and Michael Stessin. *Regularization Methods for Ill-posed Problems.* CRC Press Boca Raton, FL:, 1993.
9. David Mumford, Agnès Desolneux, et al. *Pattern theory: The Stochastic Analysis of Real-world Signals.* 2010.
10. Mila Nikolova. Model distortions in Bayesian MAP reconstruction. *Inverse Problems and Imaging*, 1(2):399, 2007.
11. Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.
12. Pierre Weiss, Laure Blanc-Féraud, and Gilles Aubert. Efficient schemes for total variation minimization under constraints in image processing. *SIAM Journal on Scientific Computing*, 31(3):2047–2080, 2009.

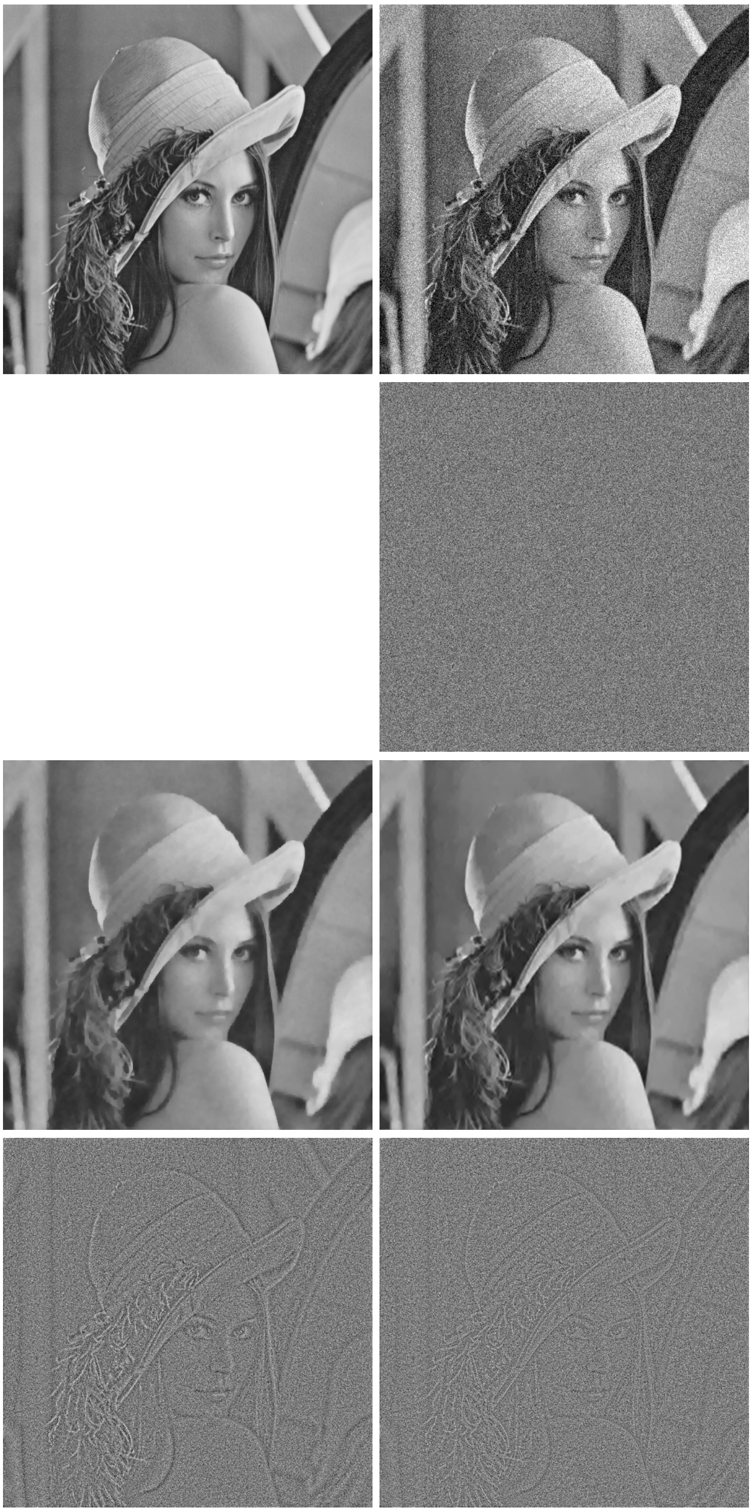

Fig. 5. Denoising results for a true example. First row, from left to right: original image, noisy image (SNR=15.4dB), denoised with TV and a regularization parameter maximizing the SNR (SNR=23.1dB, worst z-score: 69.9), denoised with the bilevel approach (SNR=24.5dB, worst z-score: 7.7). Second row: residuals associated to the top images. (same scale for the gray-level).