Pattern Recognition 43 (2010) 2609–2619

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Total variation, adaptive total variation and nonconvex smoothly clipped absolute deviation penalty for denoising blocky images

Aditya Chopra [a], Heng Lian [b,*]

[a] School of Computing Sciences, VIT University, Vellore, TN, India
[b] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore

ARTICLE INFO

Article history:
Received 19 August 2009
Received in revised form
23 March 2010
Accepted 25 March 2010

Keywords:
MM algorithm
SCAD penalty
Total variation denoising

ABSTRACT

The total variation-based image denoising model has been generalized and extended in numerous ways, improving its performance in different contexts. We propose a new penalty function motivated by the recent progress in the statistical literature on high-dimensional variable selection. Using a particular instantiation of the majorization-minimization algorithm, the optimization problem can be efficiently solved and the computational procedure realized is similar to the spatially adaptive total variation model. Our two-pixel image model shows theoretically that the new penalty function solves the bias problem inherent in the total variation model. The superior performance of the new penalty function is demonstrated through several experiments. Our investigation is limited to "blocky" images which have small total variation.

1. Introduction

Denoising is probably the most common and most studied problem in image processing. Approaches developed so far include many methods arising from the field of engineering, computer science, statistics and applied mathematics. There are several popular classes of existing denoising algorithms, from simple linear neighborhood filtering to mathematically more involved wavelet methods based on solid statistical foundations [1–3]. The PDE-based methods first proposed in [4] are unique in their formulation of images as functions in a suitable function space. Relatively few comparison studies exist among different methods, which is quite understandable due to (i) there are a large number of existing denoising approaches with many different modifications and extensions; (ii) the success or failure of different approaches depends largely on the characteristics exhibited by different types of images, whether cartoon or natural scene images, grayscale or colored, textured or solid objects. One exception is the work [5] which compared the standard total variation (TV) model with wavelet denoising and finds that TV is inferior for some standard test images. With different fine tunings and extensions available in both the class of PDE-based and wavelet-based methods, such as using higher order derivatives or correlated wavelet coefficients, it is still hard to judge from their results the relative merits of these two approaches, although it seems to be the prevailing mindset that the wavelet-based methods work better for general images.

Denoting the unobserved original noiseless image by u, the goal of denoising is to recover this original image given an observed noisy image $f=u+n$, where n denotes the noise. In traditional filtering as well as wavelet-based approaches, we either think of images as $m \times l$ matrices or $N=ml$-dimensional vectors, while the PDE-based method will generally treat images as bivariate functions defined on the unit square $\Omega=[0,1]\times[0,1]$. Introduced in [4], the standard total variation (TV) image denoising method estimates the original image by solving the following minimization problem

$$\hat{u}=\arg\min_{u}\|f-u\|^2+\lambda TV(u), \tag{1}$$

where $\|.\|$ is the L_2 norm of the function and $TV(u)=\int_{\Omega}|\nabla u|$ is the total variation norm of u [4]. The regularization parameter λ controls the tradeoff between the fidelity to the observed image and smoothness of the recovered image. Actually the paper [4] used the somewhat equivalent formulation of minimizing the total variation with constraints on the noise level, which is assumed to be known. But the penalized L_2 version stated above is more convenient when the level of the noise is unknown and we will adopt this formulation in our study. Both practically and theoretically, this model is the best understood one among PDE-based methods as of today, where the images are considered as belonging to the space of functions of bounded variation (BV) and the existence and uniqueness of solution is well-established [6–8]. Discrete version of the TV model is considered in [5], arguing that all approaches have to go through the discretization

* Corresponding author. Tel.: +65 65137175.
E-mail address: henglian@ntu.edu.sg (H. Lian).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.03.022

procedure when implemented anyway. Our point of view is that using either the continuous or discrete formulation for the PDE-based methods makes little difference in practice.

Although the standard TV model above might not be competitive for general image denoising tasks, it is believed to be ideal for blocky images, i.e., images that are nearly piece-wise constant. From a statistical point of view, this can be simply seen by the fact that it penalizes the first order partial derivatives (or, in discrete version, first order differences) and thus shrinks them towards zero. Such images are interesting for at least two reasons. First, examples of blocky images abound in real life, such as vehicle registration plates, traffic signs, postal code on envelopes, etc. A more complicated example in medical imaging is found in [9]. Second, studying of such relatively simple images can usually lead to deeper insights into different denoising approaches. [10] noted the inherent bias in the TV model and proposed the spatially adaptive total variation (SATV) model that applies less smoothing near significant edges by utilizing a spatially varying weight function that is inversely proportional to the magnitude of image derivatives. SATV is a two-step procedure where the weight function obtained from the first step using standard TV is then used to guide smoothing in the second step. The authors showed that with a modest increase in computation, SATV is superior to standard TV in restoring piece-wise constant image features.

Curiously, there is an almost parallel development in the statistical literature in the context of high-dimensional linear regression with variable selection. As explained in the next section, these studies focus on the regression problem where although there exist numerous covariates a priori, most of the regression coefficients are exactly zero, implying that the corresponding covariates have no effects on the response variable. Thus shrinking most regression coefficients to zero is a viable strategy for efficient estimation. For piece-wise constant images, with first derivatives in most locations exactly equal to zero, shrinking them to zero is thus also a reasonable approach. Taking advantage of this observation, we propose to adapt the smoothly clipped absolute deviation (SCAD) penalty [11,12] that has become extremely popular in the statistical community for our image denoising task. Although in the case of TV model the correspondence between the functional-analytical approach and the statistical approach seems to be well-known, and some have studied in detail the properties of total variation from a statistical point of view [13,14], these statistical works are only restricted to the one-dimensional case. Besides, as far as we know, the parallelism stated above has not been fully utilized and in particular the SCAD penalty has not been applied to penalize the first order differences even in the one-dimensional case. Besides its superior performance in practice, there are several advantages of SCAD penalty compared to SATV, most notably getting rid of the extra parameter that a user needs to tune for SATV in implementation. As mentioned before, we think using either discrete or continuous formulation formally makes little difference, but we choose to use the continuous formulation since it can simplify description and notation significantly. The only problem is that the objective functional using the SCAD penalty being nonconvex, existence of solution is not guaranteed. The theoretically inclined reader might want to think in discrete terms so that such technical point does not arise. Our computational experiments show that SCAD is superior to SATV in terms of mean squared error (MSE). Although MSE is notorious for describing the visual quality of an image, it is arguably less so for blocky images where MSE can describe the accuracy of restoration rather faithfully.

The rest of the paper is organized as follows. In the next section, we briefly review the TV and the SATV model and point out the almost trivial connection to Lasso and the adaptive Lasso developed in the statistical literature so that we hope readers from both fields can follow the motivation and development of the current paper. In Section 3, we adapt the SCAD penalty for our image denoising problem and discuss some properties in detail in this context. We also developed a majorization-minimization procedure using first order Taylor expansion so that the computation involved simply reduces to that similar to the SATV model, although with a different weight function. In Section 4, we will briefly review a method called Monte-Carlo SURE [15] for regularization parameter selection which is used in our study when required. In Section 5, several computational experiments are used to show the superiority of the proposed method in denoising blocky images. In these experiments, we also intentionally emphasize the difficulty encountered with SATV model in tuning its performance. We conclude the paper with a discussion in Section 6.

2. Review of the TV and SATV model

The TV model proposed by [4] and presented above in Eq. (1) has received a great deal of attention in the last decade. In [10], the authors argued that it is desirable that less smoothing is carried out where there is more detail in the image. This motivated the replacement of TV norm by the following more general weighted TV functional

$$TV_w(u) = \int_\Omega w(x,y)|\nabla u(x,y)|\,dx\,dy. \tag{2}$$

The weight w should be small in the presence of an edge so that less smoothing is performed near an edge. [10] used a weight function inversely proportional to the partial derivatives, with a parameter $e>0$ added both to avoid dividing by zero and to be used as a tuning parameter to control the amount of adaptivity. Thus in their proposal of the spatially adaptive total variation (SATV) model $w=1/(|u_x|+e)+1/(|u_y|+e)$ where u_x and u_y are the partial derivatives. [10] used a two-step method. In the first step the standard TV model (1) is used to estimate u based on which the partial derivatives (first order differences) are computed. Then the derivatives are used in (2) to compute the final restored image. If e is chosen sufficiently large, SATV basically reduces to the standard TV. On the other hand, if e is too small, artificial edges will appear and the algorithm will be numerically unstable as well. We will see in our simulations that the result is somewhat sensitive to the choice of e and the appropriate amount of adaptivity is not universal to all images, which makes it difficult to choose e in practice, or leads to a sizable increase on the amount of computation required to say the least.

As we mentioned in the introduction, there is an almost parallel line of development in the statistical literature that uses the same idea of SATV in a different context. Consider a linear regression problem $y_i=\mathbf{x}_i^T\beta+\varepsilon_i$ based on independent and identically distributed (i.i.d.) data $\{y_i,\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i=(x_{i\,1},\ldots,x_{ip})^T$ are the covariates, $\beta=(\beta_1,\ldots,\beta_p)^T$ are the regression coefficients, and ε_i is a zero mean noise. Sometimes one has good reasons to believe that only a few of the x_{iq}'s are related to y_i, i.e., many of the β_q's are exactly zero. In these situations it is desirable to design an approach that shrinks many regression coefficients to zero automatically. Lasso [16] does exactly that and IS formulated as the minimization of the following objective function:

$$\sum_{i=1}^{n}\|y_i-\mathbf{x}_i^T\beta\|^2+\lambda\sum_{i=1}^{p}|\beta_i|.$$

It is now well-known that this algorithm encourages many coefficients to be exactly zero as desired due to the use of L_1

norm penalty for β. Later [17] proposes the adaptive Lasso, which possesses better theoretical properties than Lasso and also proves to be superior in practice, that solves the following minimization problem

$$\sum_{i=1}^{n} \|y_i - \mathbf{x}_i^T \beta\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|/|\hat{\beta}_i|,$$

where $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ is the standard least square estimate. Any other reasonable estimate can be used (to be more rigorous, $\hat{\beta}$ must be *consistent* in statistical terms in order to enjoy the theoretical properties stated in that paper).

The readers can immediately see the parallel developments in statistics and TV-based image processing. When it is desirable to shrink the first order differences in an image towards zero, the same arguments that lead to Lasso and adaptive Lasso now assume the form of TV and SATV, respectively. In the statistical literature, [13,14] studied the TV problem in its discrete form, but we have not seen any mention of utilizing adaptive Lasso to penalize the first order differences.

Historically, before the appearance of adaptive Lasso, to address the shortcomings of Lasso (which in general cannot identify the non-zero coefficients in the linear model with high probability), [11] proposed the smoothly clipped absolute deviation (SCAD) penalty which is motivated by the desire to achieve several properties of the estimator such as continuity, asymptotic unbiasedness, etc. We discuss these properties in more detail in the next section. They also show that the resulting estimator possesses the so-called oracle property, i.e. it can identify non-zero coefficients with high probability and behaves the same as when the positions of the non-zero coefficients are known in advance. In the next section, we adapt the SCAD penalty for image processing tasks. Using SCAD penalty gets rid of the clumsiness of having to choose the parameter e in SATV and our experiments show that its performance is superior to SATV.

3. Image denoising with the SCAD penalty

In linear regression, using the SCAD penalty amounts to minimizing the following functional

$$\sum_{i=1}^{n} \|y_i - \mathbf{x}_i^T \beta\|^2 + \sum_{i=1}^{p} p_\lambda(|\beta_i|), \tag{3}$$

where $p_\lambda(.)$ is more conveniently defined by its derivative

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\} \quad \text{for } \theta > 0 \quad \text{with } a > 2,$$

and $p_\lambda(0) = 0$, where $I\{\cdot\}$ denotes the indicator function. As usual, $a=3.7$ is used. The recommendation of this value for a in [11] is based on some Bayesian arguments which also show that the estimate is not very sensitive to the value of a. Although it is possible to optimize over a for each specific problem, such an implementation can be computationally expensive and thus we use the recommended value in this paper.

We plot the function p_λ in Fig. 1(a) for $\lambda = 1$ and its derivative in Fig. 1(b). As seen in (3) we only use p_λ and its derivative with a nonnegative functional argument. We plot both functions in Fig. 1 as even functions for convenience, although the derivative should be an odd function if p_λ is defined as an even function. Note that this penalty function, unlike the L_1 penalty used in Lasso, is not convex. To use the SCAD penalty for image denoising, we formally write down the functional

$$\|f - u\|^2 + \int_\Omega p_\lambda(|\nabla u|). \tag{4}$$

Some readers will have the objection that p_λ is nonconvex and thus the existence of solution to the above functional is in question. Even the definition of $p_\lambda(|\nabla u|)$ itself seems to be a difficult task, if not impossible. Note [7] only defined $\phi(|\nabla u|)$ when ϕ is convex and u is a BV function. Due to this problem we encourage the reader to change to a discrete formulation which is straightforward from (4). The expression (4) in the continuous form is so much cleaner so we prefer to keep it. This should hopefully be just a minor nuisance for practitioners.

As argued in [11], the SCAD penalty function is motivated by the following desired properties.

- (P1) (Continuity) The estimates should smoothly depend on the data to avoid instability in estimation.
- (P2) (Sparsity) In linear regression, it should correctly identify the zero linear coefficients. Adapting this property in our context, for image denoising tasks, it should correctly identify clusters of pixels with identical intensities.
- (P3) (Unbiasedness) The effective shrinkage applied to large coefficients should decrease to zero.

These properties can be easily seen by considering a one-component model

$$\tfrac{1}{2}(y-\theta)^2 + p_\lambda(|\theta|),$$

where y is the single noisy observation with unknown mean θ. In this case, the minimizer $\hat{\theta}$ has a closed-form expression as shown in Eq. (2.8) in [11]. If instead of the SCAD penalty, we use the Lasso penalty $p_\lambda(|\theta|) = \lambda|\theta|$, the minimizer becomes $\hat{\theta} = (y-\lambda)_+$, where

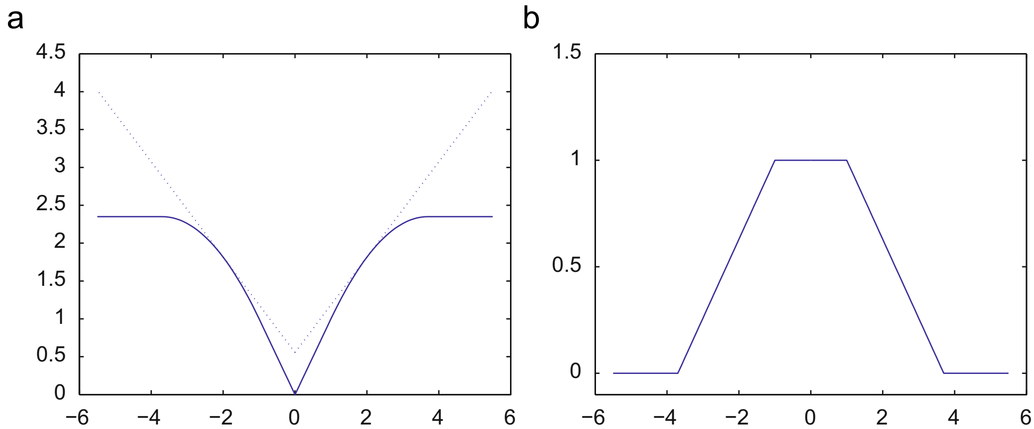

Fig. 1. (a) The SCAD penalty function and its linear approximation. (b) The derivative of the SCAD penalty function.

$(a)_+=a$ if $a>0$ and zero otherwise. Thus we see the Lasso penalty results in soft thresholding of the observation. On the other hand, if $p_\lambda(\theta)=\lambda^2-(|\theta|-\lambda)^2 I\{|\theta|<\lambda\}$, we have $\hat{\theta}=yI\{|y|>\lambda\}$, obtaining the hard-thresholding rule. The soft threshoding rule does not satisfy the unbiasedness property, the hard-thresholding rule does not satisfy the continuity property, while the estimate obtained using the SCAD penalty satisfies all three properties.

Before we present further properties of the SCAD penalty in the context of image denoising, it should be mentioned that [18] has used SCAD penalty for wavelet denoising. They focus on one-dimensional denoising problems and application to 2D denoising is straightforward although no examples are provided for image denoising in that paper. In [19], nonlinear shrinkage estimation in the wavelet domain is proposed based on multiple differences of sigmoid functions. In [20,21], a new sigmoid-based shrinkage function motivated from similar considerations as discussed in the previous paragraph is used also in the wavelet domain. Different from all these works, we apply the SCAD penalty directly on the image domain. The choice of penalty function is based on its simplicity in form and that there is only one tuning parameter in the penalty function to be chosen. The penalty function proposed in [20] could be used but involves more tuning parameters. Our main message here is that a carefully chosen penalty function on the image domain can improve the simple total variation penalty-based image denoising.

To see clearly the effect of SCAD compared to TV, we consider the following simple discrete problem instead:

$$\arg\min_{\theta_1,\theta_2}(y_1-\theta_1)^2+(y_2-\theta_2)^2+p_\lambda(|\theta_1-\theta_2|), \tag{5}$$

i.e., we consider an "image" with only two pixels. We have the following property of the minimizer comparing the SCAD penalty with the TV and SATV penalties, the proof of which is deferred to the Appendix A.

Proposition 1. *Suppose without loss of generality that $y_1\geq y_2$.*

(a) *If $y_1-y_2>a\lambda$, the minimizer of* (5) *is $\theta_1=y_1,\theta_2=y_2$.*
(b) *If $y_1-y_2<\lambda=\min_{\xi\in R}(|\xi|+p'_\lambda(|\xi|))$, the minimizer of* (5) *is $\theta_1=\theta_2=(y_1+y_2)/2$.*
If instead the TV norm is used, i.e. $p_\lambda(|\theta_1-\theta_2|)$ is replaced by $\lambda|\theta_1-\theta_2|$ in (5), *then*
(c) *if $y_1-y_2\geq\lambda$, the minimizer is $\theta_1=y_1-\lambda/2,\theta_2=y_2+\lambda/2$.*
(d) *if $y_1-y_2<\lambda$,the minimizer is $\theta_1=\theta_2=(y_1+y_2)/2$.*
When using the SATV penalty $\lambda|\theta_1-\theta_2|/(|\theta_1-\theta_2|+e)$, we have
(e) *if $(y_1-y_2)(y_1-y_2+e)\geq\lambda$, the minimizer is $\theta_1=y_1-\lambda/(2(y_1-y_2+e)),\theta_2=y_2+\lambda/(2(y_1-y_2+e))$.*
(f) *if $(y_1-y_2)(y_1-y_2+e)<\lambda$, the minimizer is $\theta_1=\theta_2=(y_1+y_2)/2$.*

From the proposition, we see that for this simple two-pixel image model, although both penalties have the effect of shrinking θ_1 and θ_2 to be exactly equal to each other, the SCAD penalty has the additional desired property that when the difference $|y_1-y_2|$ is large enough, no shrinkage is applied. This has the intuitive appeal that when it is suspected an "edge" exists between two pixels, no "borrowing of information" occurs across the edge. From part (c) of the proposition the TV model is implicitly biased, which is already known in more general contexts as shown in [22,23]. Our experiments later also demonstrated this effect. From the proof in the Appendix A it can be seen that this difference arises basically from the fact that $p'_\lambda(\theta)=0$ when θ is big enough. Finally, although SATV has less bias for larger $|y_1-y_2|$, this bias only vanishes asymptotically as $|y_1-y_2|$ goes to infinity. One can

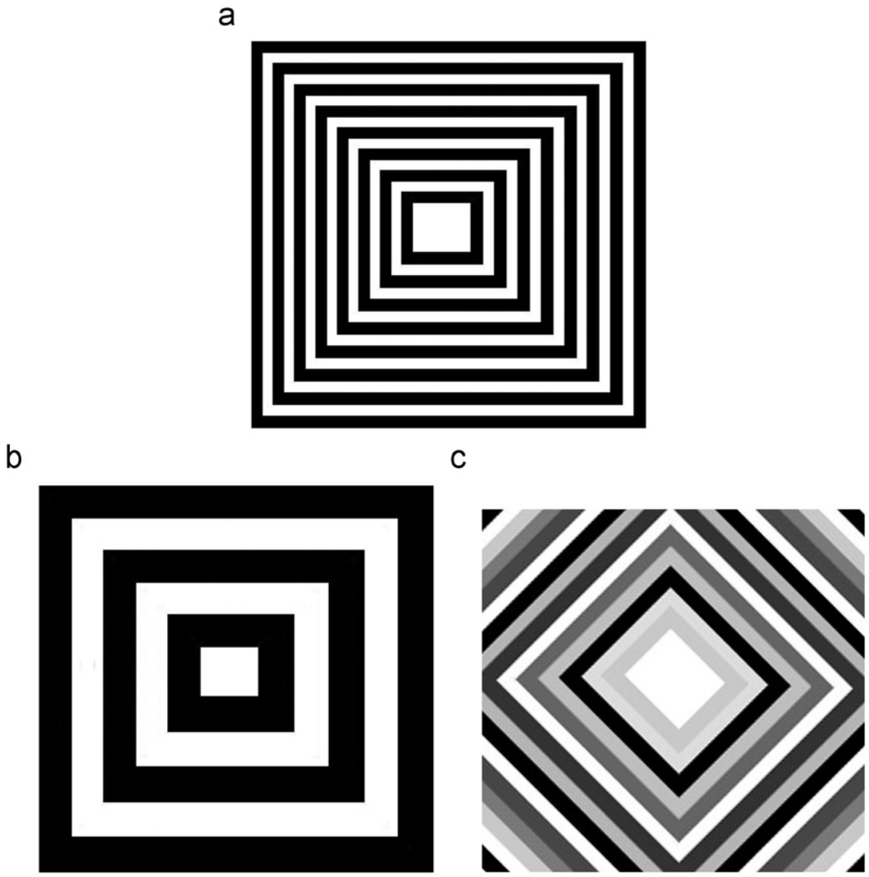

Fig. 2. Several simple grayscale images used in the experiments.

also see the effect of the parameter e: larger e reduces the bias as shown in part (e) of the proposition, but at the same time the condition $(y_1-y_2)(y_1-y_2+e)<\lambda$ in part (f) is only satisfied for smaller y_1-y_2, thus the parameter e represents a bias-sparsity trade-off.

Compared to TV or SATV, optimization of the functional (4) is more complicated since the functional is nonconvex and using time evolution of the corresponding Euler–Lagrange equation (i.e., gradient descent) is potentially problematic. Thus we use the following majorization-minimization (MM) algorithm instead. Note that [5] also proposed an MM algorithm for standard TV image denoising.

MM algorithm is a general technique for finding the minimizer (or maximizer) of a function, say $g(\theta)$. Suppose $\theta^{(k)}$ represents the current estimate of θ in search of the minimizer, and let $Q(\theta|\theta^{(k)})$ denote a real-valued function of θ whose form depends on $\theta^{(k)}$. The function $Q(\theta|\theta^{(k)})$ is said to majorize g at the point $\theta^{(k)}$ if

$$Q(\theta|\theta^{(k)}) \geq g(\theta) \text{ for all } \theta \quad \text{and} \quad Q(\theta^{(k)}|\theta^{(k)}) = g(\theta^{(k)}).$$

If $\theta^{(k+1)}$ is the minimizer of $Q(\theta|\theta^{(k)})$, then we have

$$g(\theta^{(k+1)}) \leq Q(\theta^{(k+1)}|\theta^{(k)}) \leq Q(\theta^{(k)}|\theta^{(k)}) = g(\theta^{(k)}).$$

This descent property is the key to the celebrated numerical stability of the MM algorithm. Since the well-known EM algorithm is a special case of MM algorithm, many results from the EM algorithm literature, such as those contained in [24], carry over without change to MM algorithms. See also Theorems A.1 and A.3 in [25].

For our problem, we first majorize the SCAD penalty function using its first order Taylor expansion at an initial estimated image $u^{(0)}$ (we could simply set $u^{(0)}=f$ for example):

$$p_\lambda(|\nabla u|) \leq p_\lambda(|\nabla u^{(0)}|) + p'_\lambda(|\nabla u^{(0)}|)(|\nabla u| - |\nabla u^{(0)}|),$$

which is illustrated in Fig. 1(a) as the dotted line. Using this approximation, we can repeatedly solve the problem:

$$\begin{aligned} u^{(k)} = \arg\min_u \|f-u\|^2 + \int & p_\lambda(|\nabla u^{(k-1)}|) \\ & + p'_\lambda(|\nabla u^{(k-1)}|)(|\nabla u| - |\nabla u^{(k-1)}|), \quad k=1,2,\ldots,K, \end{aligned}$$

i.e., replacing the SCAD penalty by its upper bound and then solving the new optimization problem. Getting rid of terms that are independent of u, we are actually minimizing the following functional

$$u^{(k)} = \arg\min_u \|f-u\|^2 + \int p'_\lambda(|\nabla u^{(k-1)}|)|\nabla u|, \quad k=1,2,\ldots,K, \tag{6}$$

which is in the same form as the functional with SATV penalty (2) with a weight function $w = p'_\lambda(|\nabla u^{(k-1)}|)$ that is different for each iteration k. Thus the computation involved is almost identical to SATV, with an extra outer loop that modifies the weight function in each iteration. Formally, each inner loop will use the evolutionary PDE derived from the Euler–Lagrange equation to solve (6):

$$u_t = \nabla \cdot \left\{ p'_\lambda(|\nabla u^{(k-1)}|)\frac{\nabla u}{|\nabla u|} \right\} - (u-f). \tag{7}$$

From this analogy with SATV, we can also see the advantage of SCAD from another point of view: the weight function $w = p'_\lambda(|\nabla u^{(k-1)}|)$ is bounded and thus there is no instability

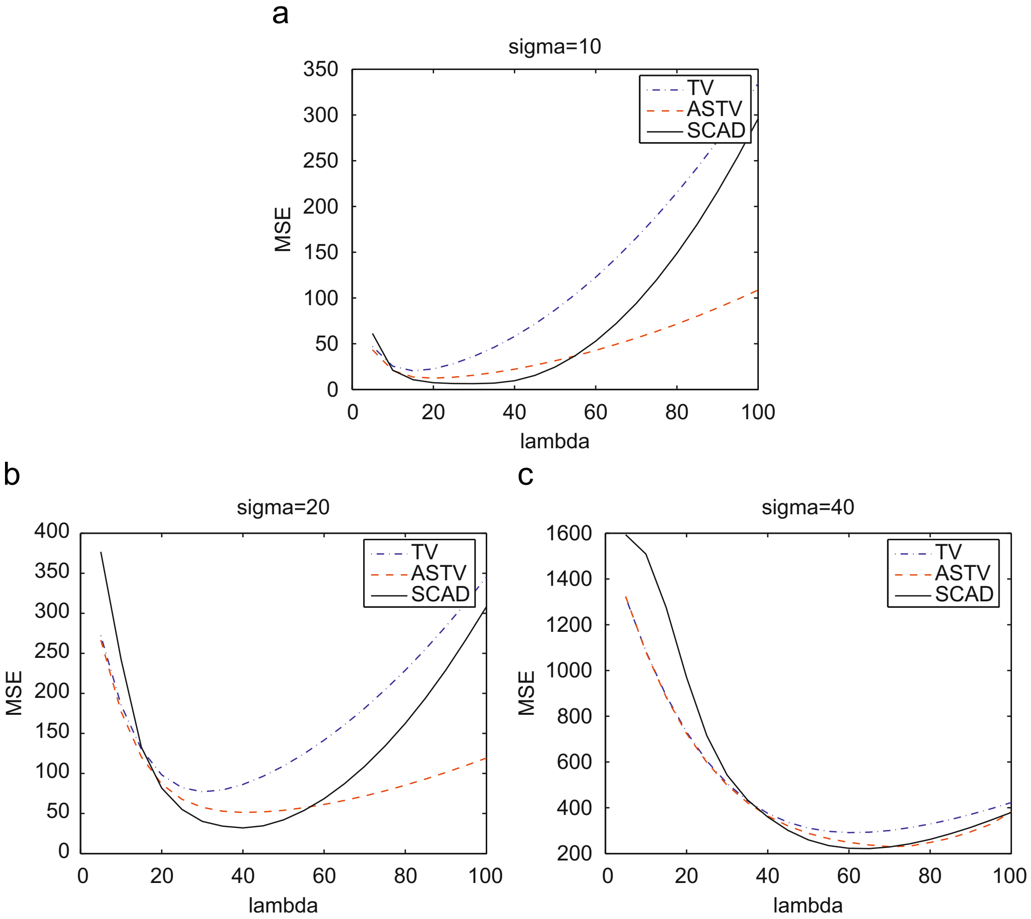

Fig. 3. Comparison of MSE for the three methods for the image shown in Fig. 2(a), with different noise levels: (a) $\sigma=10$; (b) $\sigma=20$; (c) $\sigma=40$.

problem as when w is inversely proportional to the first derivative, which makes an extra tuning parameter e unnecessary in the SCAD model.

In summary, the MM algorithm for minimizing (4) goes as follows:

- Set $u^{(0)}=f$
- for $k=1{:}K$
  Minimize the convex function (6) by evolving the Euler–Lagrange Eq. (7). The gradient $\nabla=(\partial_x,\partial_y)$ is simply approximated by the difference of intensities of neighboring pixels and $|\nabla u|$ in the denominator is approximated by $|\nabla u|\approx\sqrt{u_x^2+u_y^2+\beta^2}$ with $\beta=0.001$.
- end for

From the general property of the MM algorithm, it produces a sequence of monotonically decreasing values of the objective functional (4) which makes the algorithm very stable. In practice for our experiments, we find that the number of iterations K can be taken as small as $K=2$, thus the running time of the algorithm is comparable to both standard TV and SATV.

4. Monte-Carlo SURE for regularization parameter selection

In all the above methods the value of the regularization parameter chosen largely determines the quality of the denoised image. We use MSE as the criterion for judging the relative merits of different methods in this paper, which is defined by

$$\frac{1}{N}\|u-\hat{u}\|^2,$$

where we take the original image u as a N-dimensional vector and $\hat{u}$ is the restored image. Note that we will consider the discrete formulation in this section. To calculate MSE we need to have the

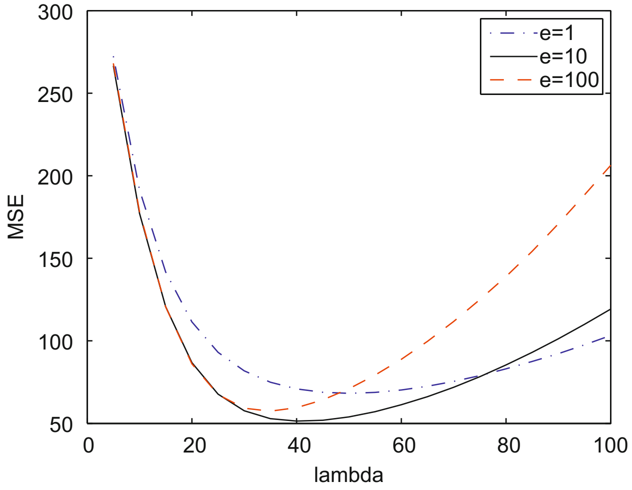

Fig. 5. Comparison of MSE for the SATV model when different values for e are chosen, with noise level $\sigma=20$.

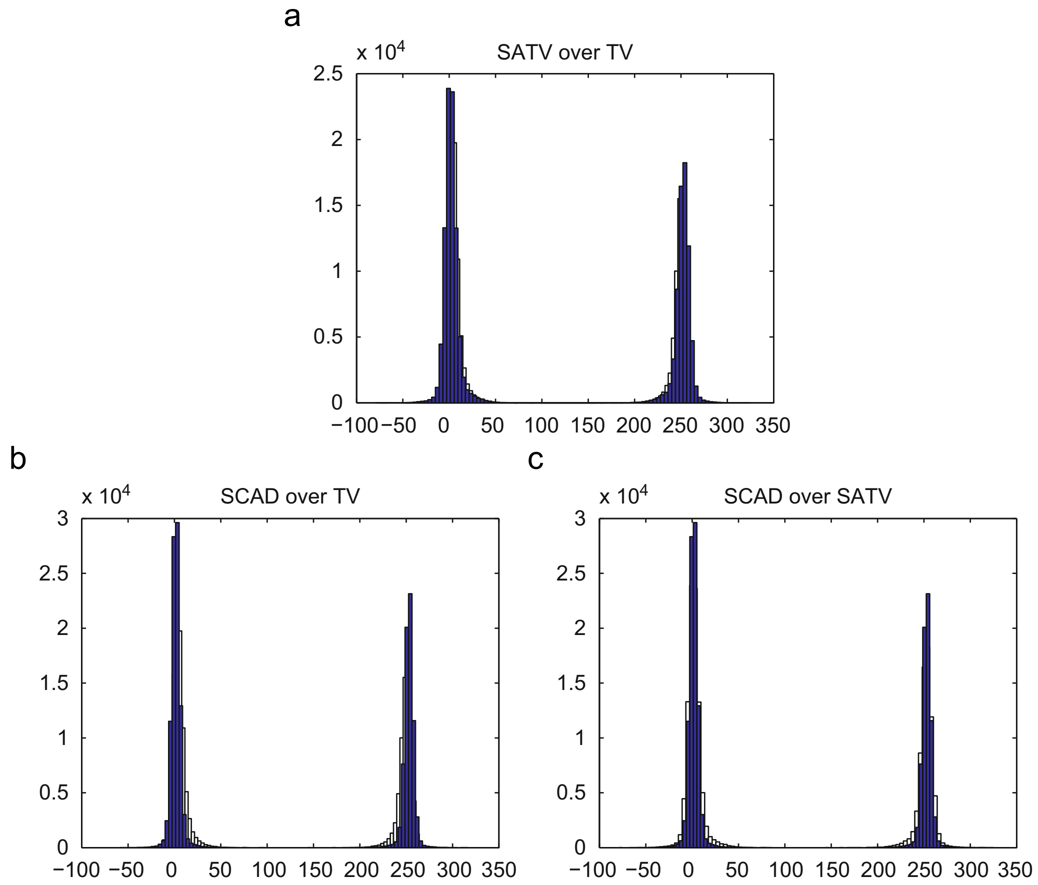

Fig. 4. The histogram of restored image intensities overlaid on top of each other. (a) Histogram of restored image intensities obtained by SATV over that obtained by TV model. (b) Histogram of restored image intensities obtained by SCAD over that obtained by TV model. (c) Histogram of restored image intensities obtained by SCAD over that obtained by SATV model.

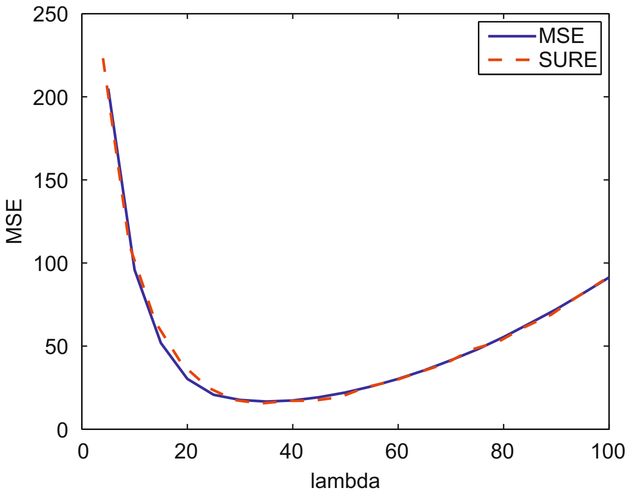

Fig. 6. MSE and SURE estimate for the SCAD method.

Table 1
MSE of using different methods on the image shown in Fig. 2(b).

Noise level	TV	SATV	SCAD
$\sigma=20$	31.97	24.96(e=100)	17.13
$\sigma=40$	114.71	95.76(e=100)	92.00
$\sigma=80$	415.11	387.10(e=100)	383.77

Table 2
MSE of using different methods on the image shown in Fig. 2(c).

Noise level	TV	SATV	SCAD
$\sigma=10$	37.02	34.10 (e=10)	29.10
$\sigma=20$	99.37	92.46(e=10)	77.39
$\sigma=40$	370.68	275.08(e=10)	266.65
$\sigma=80$	886.95	858.32(e=100)	805.66

prior knowledge of the noise-free image which in most realistic scenarios is unavailable. When the noise is Gaussian, [15] proposed a technique called Monte-Carlo SURE, which does not require any prior knowledge of the noise-free image or the specific nature of the denoising algorithm. For the purpose of presenting this method, we now should change to a discrete formulation. For a noisy image $f=u+n$, formulated in the discrete domain, and a denoising algorithm considered abstractly as a mapping $\hat{u}=M(f)$ that returns a restored image $\hat{u}$ with f as the input, [15] proved that

$$\frac{1}{N}\|f-M(f)\|^2-\sigma^2+\frac{2\sigma^2}{N}\mathrm{div}_f M(f) \tag{8}$$

is an unbiased estimator of the true MSE, where σ is the standard deviation of the Gaussian noise and $\mathrm{div}_{fM}(f)$ is the divergence of the multivariate function M. Note in our context the mapping M implicitly depends on the regularization parameter λ. Direct calculation of $\mathrm{div}_{fM}(f)$ is not feasible except for simple linear filtering operation, and [15] used the Monte-Carlo approximation

$$\mathrm{div}_f M(f)\approx \mathbf{b}^T(M(f+\varepsilon\mathbf{b})-M(f)),$$

where $\mathbf{b}$ is a N-dimensional vector with i.i.d. standard normal random components, and ε is a small positive constant. That is, we artificially add more noise to the observed image and run the same denoising algorithm again and then approximate the divergence based on the differences of the two recovered images. We will use Monte-Carlo SURE to choose the regularization parameter whenever required in the next section. Since the noise level is assumed to be unknown in our experiments, some pilot estimate of σ should be plugged into Eq. (8). In all our experiments, we used the following simple estimate that is quite robust empirically for blocky images:

$$\hat{\sigma}=\mathrm{median}\{|f_i-f_j|\}/0.954, \tag{9}$$

where $f=(f_1,\ldots,f_N)$ is the observed image and the differences f_i-f_j are taken over all neighboring pixels (four neighbors for each pixel). This estimate is based on the fact that with a normal random variable $X\sim N(0,2\sigma^2)$, $\mathrm{median}(|X|)\approx 0.954\sigma$. Related to

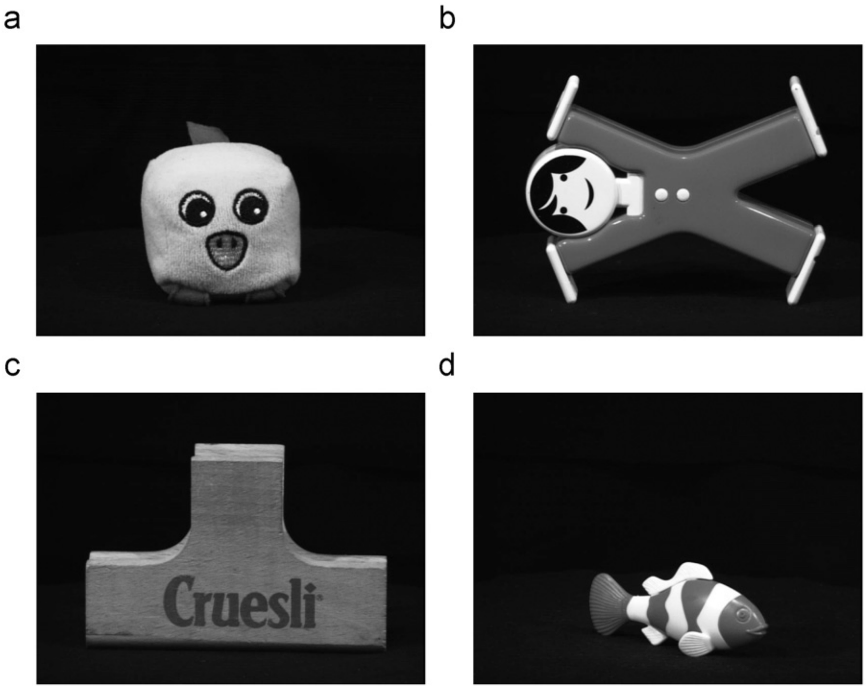

Fig. 7. Four images obtained from ALOI used for testing the performances of different methods.

this proposal, median-based noise estimate in the wavelet domain has been adopted by [26].

5. Experiments

First we compare the performances of the three approaches TV, SATV and SCAD using a simple black-and-white image shown in Fig. 2(a). In this first experiment, we do not choose any single regularization parameter but compare the performance over a whole wide range of regularization parameters. Independent Gaussian noise with standard deviations $\sigma=10,20$ and 40 are added to the original image and taken as the observed noisy input. For the initial step of SATV, we use TV with optimal parameter λ to estimate the weight function. We also search for a good value of e in the second step (based on minimization of the true MSE) for $e \in \{1,10,100,500\}$, it turns out for all three different noise levels for this image $e=10$ gives the best result. Note that we consider the intensity values of an image to be in the range of [0,255]. Both choices actually make the results more favorable for SATV, but we will see that even so it is being outperformed by SCAD. Fig. 3 shows the evolution of the true MSE using different regularization parameters for the three methods, with different subfigures illustrating the observed image with different noise levels. From these figures, it is clearly seen that SCAD performs better than SATV, while both are significantly better than TV. To get some insights into the effect of the different penalties, the image histograms for the recovered images are shown in Fig. 4 for the case of $\sigma=20$. One can see from the histograms of the TV-based restoration that the TV estimate is biased, in that black colored pixel intensities (with original intensity value of zero) are generally shifted up while white colored pixel intensities (with original intensity of 255) are shifted down, consistent with the proposition stated previously. While SATV only partially addresses this, SCAD seems to be more efficient in solving this bias problem. Besides, Fig. 4(c) demonstrates that for the recovered image using the SCAD penalty, the histogram is more peaked and thus resulting in smaller MSE.

Table 3
MSE for different methods applied to four object images obtained from ALOI when $\sigma=40$.

	TV	SATV	SCAD
duck	77.20	75.80 (e=100)	69.70
person	93.22	84.89(e=100)	79.35
board	82.58	74.95(e=100)	68.87
fish	70.99	63.69(e=100)	55.58

Using this experiment, we can also see the effect of e on the result. As stated above $e=10$ is optimal for SATV for this image. We see from Fig. 5 that using $e=1$ or $e=100$ makes the MSE bigger. Specifically, using $e=1$ enlarged the minimum MSE from 51.40 to 68.21, or by 34%, while using $e=100$ enlarged MSE by 13%. Based on the suggestion from one referee, for this experiment, we further search for optimal e using a finer grid $(5,10,15,\ldots)$. For three noise levels, the optimal MSEs from the SATV model are 11.81, 50.03 and 226.05, respectively. In comparison the MSEs from the SCAD method are 6.41, 31.01 and 221.57, respectively. Unfortunately there is no universally best value for e, and our later experiments demonstrate that for different images the optimal e is difficult to predict. Choosing a wrong value for e makes the performance of SATV more unpredictable. Although e could be selected by similar methods that have been developed for selecting λ, for example using Monte-Carlo SURE, this at least increases significantly the computational burden of the algorithm. And even with a good estimate of e, our result here shows that it is still worse than SCAD in terms of the MSE criterion.

We also use this test image of size 400×400 to give the readers some idea of the running time of different algorithms.

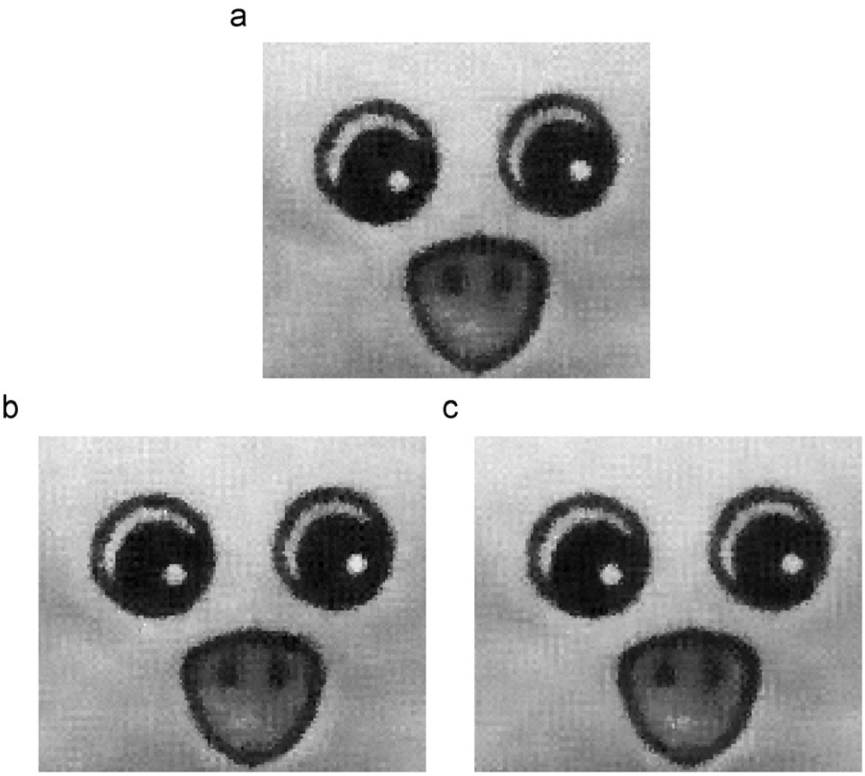

Fig. 8. Zooming into part of the image from the denoising results on the duck image using (a) TV (b) SATV (c) SCAD.

Exact comparison is difficult since each algorithm is iterative in nature and the number of iterations used determine the running time. In our implementation, we always use 100 iterations when performing gradient descent evolution. The SATV uses the TV result to compute the weight function. The SCAD model with $K=2$ also has 200 iterations in total. Thus we expect the running times of TV and SCAD are similar. Indeed, the time used in denoising for TV, SATV and SCAD are 6.35, 14.00 and 15.37 s, respectively.

Our second experiment uses images as shown in Fig. 2(b) and (c). The former is still a black-and-white image with thicker nested squares. The latter is an image similar in structure to Fig. 2(a) but with different grayscale levels and also rotated by 45°. Image Fig. 2(b) is clearly easier to denoise due to the larger scale of its features, thus we choose to add Gaussian noise with standard deviations $\sigma=20,40,80$. For image (c) we use four different levels $\sigma=10,20,40,80$. The regularization parameters now are selected using Monte-Carlo SURE as briefly described previously with σ assumed unknown and estimated using (9). The effectiveness of Monte-Carlo SURE in general has been demonstrated for some methods including the TV model in [15]. We additionally verified its performance in our SCAD model under several situations and found it to be quite accurate for our proposed model. As an illustration, for denoising the image shown in Fig. 2(b) with $\sigma=20$, we demonstrate that Monte-Carlo SURE accurately predicts the true MSE in Fig. 6. The MSE of the restoration results for the two images are presented in Tables 1 and 2, respectively. For the SATV method, the optimal values of e in each situation is also indicated in the table. Note that the optimal e is found from the true MSE and thus the results presented are favorable for the SATV method. The reader can now see that different situations require different choices of e and there seems to be no universal way of specifying a good value a priori. The conclusion is the same as before: SCAD is superior to SATV.

Table 4
MSE for different methods applied to five test images when $\sigma=20$.

	TV	SATV	SCAD
barbara	174.52	171.13 (e=10)	168.53
lena	92.08	89.83(e=10)	89.11
mandrill	209.54	209.03(e=100)	209.71
boat	115.59	111.82(e=10)	112.80
house	78.10	74.73(e=100)	74.12

Next, we use some slightly more complicated images to test the performances. Amsterdam Library of Object Images (ALOI, http://staff.science.uva.nl/simaloi/) is a color image collection of one-thousand small objects, recorded for scientific purposes. We pick four images as shown in Fig. 7 and transform them to grayscale images, which looks close to piece-wise constant visually. Gaussian noises with standard deviation of 40 are added to each image and different methods are applied. The results in terms of MSE are presented in Table 3, and the method

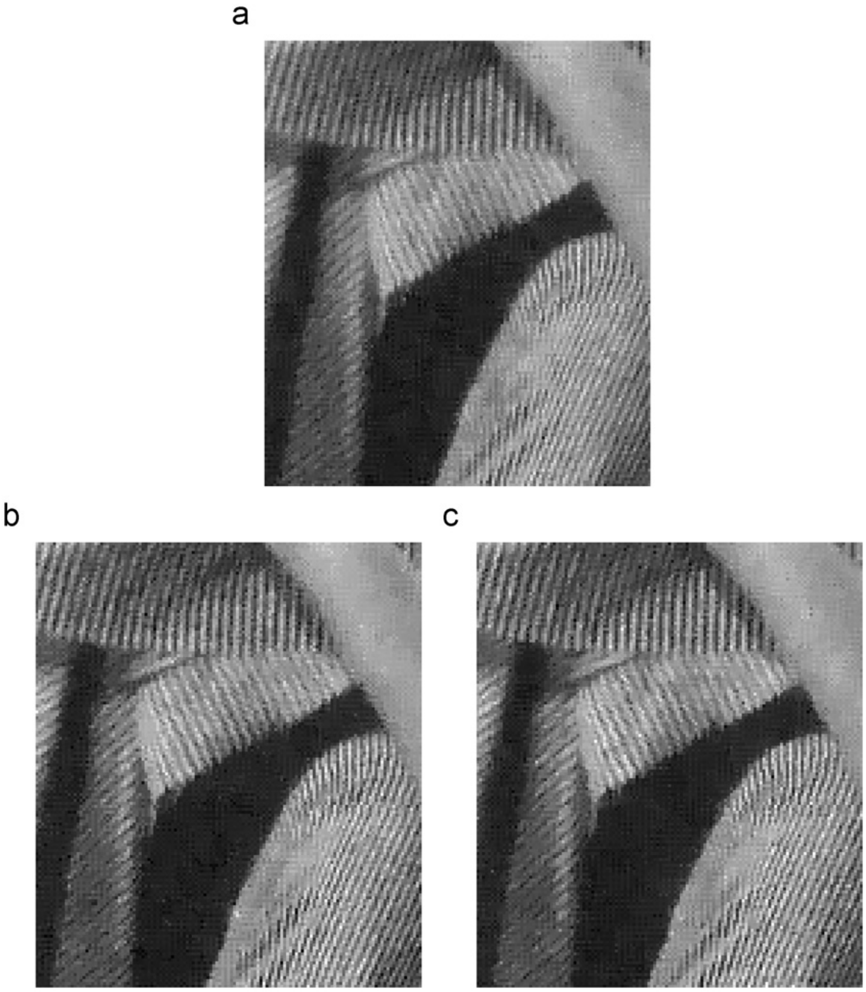

Fig. 9. Zooming into part of the image from the denoising results on the Barbara image with $\sigma=20$ using (a) TV (b) SATV (c) SCAD.

using the SCAD penalty is still the best even for these more complicated images. The denoising results for the duck image are shown in Fig. 8. Since it is visually difficult to distinguish the restored images in print using different methods, we choose not to show the restored images here, but the images are available from http://www3.ntu.edu.sg/home/henglian/SCADcode/ in MATLAB's .fig format, where the reader can also find the MATLAB codes used in our implementations.

Finally, we perform some experiments on five standard test images: Barbara, Lena, Mandrill, Boat and House. The MSE are presented in Table 4. Both SATV and SCAD produce comparable or better results than simple TV model. However, note that our implementation is biased favorably towards SATV in that the optimal parameter e is used based on minimization of the MSE. The results for the Barbara image are shown in Fig. 9 while the rest are provided on the same website mentioned above.

6. Conclusion

In this paper, we proposed a new penalization functional for image denoising. The penalty function is directly motivated by the well-known oracle property of the SCAD penalty from the statistical literature originally proposed for high-dimensional statistical regression problems. Using a simple argument in a maybe overly simplistic situation, i.e., our two-pixel image model (5), we show that the functional with SCAD penalty solves the bias problem inherent in TV regularization, which is also verified by our experimental results. Compared to spatially adaptive TV, the newly proposed method gets rid of the headache of choosing an extra parameter that controls the stability and adaptivity of the algorithm, and achieves better mean squared error at the same time. Our goal in this paper is not to propose a general image denoising method to compete with the state-of-the-art such as the wavelet-based method or the nonlocal mean [27] which has become very popular recently, but to show that a carefully designed penalty function can improve existing PDE-based approaches without extra computational burden. Due to its shrinkage to zero of the first order differences, the method is most suitable for recovering blocky images. For TV regularization, some extensions using higher order derivatives have been proposed, but this is outside the scope of the current paper.

Acknowledgment

The authors would like to thank the anonymous referees whose comments and suggestions greatly improve the manuscript. The research of the second author is supported by Singapore MOE Tier 1 Grant.

Appendix A

Proof of the Proposition. We only prove the proposition for parts (a) and (b), the proofs for parts (c)–(f) are similar and slightly simpler. Let $Q(\theta_1,\theta_2)=(y_1-\theta_1)^2+(y_2-\theta_2)^2+p_\lambda(|\theta_1-\theta_2|)$. Obviously the minimizer satisfies $\theta_1 \geq \theta_2$ when $y_1 \geq y_2$ (otherwise exchanging the values of θ_1 and θ_2 makes the functional smaller). The partial derivatives are (for $\theta_1 > \theta_2$)

$$\frac{\partial Q}{\partial \theta_1}=2(\theta_1-y_1)+p'_\lambda(|\theta_1-\theta_2|), \quad \frac{\partial Q}{\partial \theta_2}=2(\theta_2-y_2)-p'_\lambda(|\theta_1-\theta_2|).$$

The complication only comes from nondifferentiability when $\theta_1=\theta_2$. When constrained to $\theta_1=\theta_2$, it is easy to see from the quadratic form of Q that the only potential minimizer is $\theta_1=\theta_2=(y_1+y_2)/2$. Meanwhile, when $y_1-y_2>a\lambda$, we have $Q((y_1+y_2)/2,(y_1+y_2)/2)=(y_1-y_2)^2/2>(a+1)\lambda^2/2=p_\lambda(|y_1-y_2\ \ |)=Q(y_1,y_2)$. Here the equality $(a+1)\lambda^2/2=p_\lambda(|y_1-y_2|)$ follows directly from the definition of the SCAD penalty which is also explicitly written out in [28]. Thus the minimizer must satisfy $\theta_1 \neq \theta_2$ and the functional is differentiable near the minimizer, which in turn implies that both partial derivatives are equal to zero. Adding and subtracting the two partial derivatives, we get

$$\theta_1+\theta_2=y_1+y_2, \tag{10}$$

$$\theta_1-\theta_2=y_1-y_2-p'_\lambda(|\theta_1-\theta_2|). \tag{11}$$

From (11), $\theta_1-\theta_2$ is a solution to the equation $x+p'_\lambda(x)=y_1-y_2$. The function on the left hand side, when written down explicitly, is

$$x+p'_\lambda(x)=\begin{cases}\lambda+x & x<\lambda\\ \dfrac{a\lambda}{a-1}+\left(1-\dfrac{1}{a-1}\right)x & \lambda \leq x \leq a\lambda\\ x & x>a\lambda,\end{cases} \tag{12}$$

which is strictly increasing for $x>0$ and the equation $x+p'_\lambda(x)=y_1-y_2$ obviously has a unique solution $x=y_1-y_2$ when $y_1-y_2>a\lambda$. Combining this with (10), we get $\theta_1=y_1,\theta_2=y_2$, and part (a) is proved. □

For part (b), if the minimizer satisfies $\theta_1 \neq \theta_2$ so that the minimizer is a stationary point, then $\theta_1-\theta_2>0$ is a solution to the equation $x+p'_\lambda(x)=y_1-y_2$ by exactly the same arguments as before. From (12), it is easy to see that the left hand side is bounded below by $\lambda>0$ and thus there exists no solution when $y_1-y_2<\lambda$, leading to a contradiction. Now with the constraint $\theta_1=\theta_2$, it is immediate from the form of the functional $Q(\theta_1,\theta_2)$ that $\theta_1=\theta_2=(y_1+y_2)/2$.

References

[1] D.L. Donoho, I.M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, Journal of the American Statistical Association 90 (432) (1995) 1200–1224.

[2] M.A.T. Figueiredo, R.D. Nowak, Wavelet-based image estimation: an empirical Bayes approach using Jeffreys' noninformative prior, IEEE Transactions on Image Processing 10 (9) (2001) 1322–1331.

[3] J. Portilla, V. Strela, M.J. Wainwright, E.P. Simoncelli, Image denoising using scale mixtures of Gaussians in the wavelet domain, IEEE Transactions on Image Processing 12 (11) (2003) 1338–1351.

[4] L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, in: 11th Annual International Conference of the Center for Nonlinear Studies on Experimental Mathematics : Computational Issues in Nonlinear Science, Los Alamos, Nm, 1991, pp. 259–268.

[5] M.A.T. Figueiredo, J. Bioucas Dias, J.P. Oliveira, R.D. Nowak, On total variation denoising: a new majorization-minimization algorithm and an experimental comparison with wavelet denoising, in: IEEE International Conference on Image Processing (ICIP 2006), Atlanta, GA, 2006, pp. 2633–2636.

[6] A. Chambolle, P.L. Lions, Image recovery via total variation minimization and related problems, Numerische Mathematik 76 (2) (1997) 167–188.

[7] L. Vese, A study in the BV space of a denoising-deblurring variational problem, Applied Mathematics and Optimization 44 (2) (2001) 131–161.

[8] D.C. Dobson, F. Santosa, Recovery of blocky images from noisy and blurred data, SIAM Journal on Applied Mathematics 56 (4) (1996) 1181–1198.

[9] D. Dobson, Recovery of blocky images in electrical impedence tomography, in: H. Engl, A. Louis, W. Rundell (Eds.), Inverse Problems in Medical Imaging and Nondestructive Testing, Springer-Verlag, Wien, 1997, pp. 43–64.

[10] D.M. Strong, P. Blomgren, T.F. Chan, Spatially adaptive local feature-driven total variation minimizing image restoration, in: F. Preteux, J.L. Davidson, E.R. Dougherty (Eds.), Conference on Statistical and Stochastic Methods in Image Processing II, San Diego, CA, 1997, pp. 222–233.

[11] J.Q. Fan, R.Z. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association 96 (456) (2001) 1348–1360.

[12] J.Q. Fan, H. Peng, Nonconcave penalized likelihood with a diverging number of parameters, Annals of Statistics 32 (3) (2004) 928–961.

[13] E. Mammen, S. van de Geer, Locally adaptive regression splines, Annals of Statistics 25 (1) (1997) 387–413.

[14] P.L. Davies, A. Kovac, Local extremes, runs, strings and multiresolution, Annals of Statistics 29 (1) (2001) 1–48.

[15] S. Ramani, T. Blu, M. Unser, Monte-Carlo sure: a black-box optimization of regularization parameters for general denoising algorithms, IEEE Transactions on Image Processing 17 (9) (2008) 1540–1554.

[16] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Series B-Methodological 58 (1) (1996) 267–288.

[17] H. Zou, The adaptive lasso and its oracle properties, Journal of the American Statistical Association 101 (476) (2006) 1418–1429.

[18] A.A.J. Fang, Regularization of wavelet approximations, Journal of the American Statistical Association 96 (455) (2001) 939–967.

[19] J.C. Pesquet, D. Leporini, A new wavelet for image denoising, in: Sixth International Conference on Image Processing and Its Applications, Dublin, Ireland, 1997, pp. 249–253.

[20] A.M. Atto, D. Pastor, G. Mercier, Smooth sigmoid wavelet shrinkage for nonparametric estimation (2008) 3265–3268.

[21] A.M. Atto, D. Pastor, G. Mercier, Smooth adaptation by sigmoid shrinkage, EURASIP Journal on Image and Video Processing (2009) doi:10.1155/2009/532312.

[22] D. Strong, T. Chan, Edge-preserving and scale-dependent properties of total variation regularization, Inverse Problems 19 (6) (2003) S165–S187.

[23] T.F. Chan, S. Esedoglu, Aspects of total variation regularized L_1 function approximation, SIAM Journal on Applied Mathematics 65 (5) (2005) 1817–1837.

[24] C.F. Wu, On the convergence properties of the EM algorithm, Annals of Statistics 11 (1983) 95–103.

[25] E.D. Schifano, R.L. Strawderman, M.T. Wells, MM algorithms for minimizing nonsmoothly penalized objective functions, Arxiv (2010) 〈http://arxiv.org/PS_cache/arxiv/pdf/1001/1001.4776v1.pdf〉.

[26] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, D. Picard, Wavelet shrinkage—asymptopia, Journal of the Royal Statistical Society Series B-Methodological 57 (2) (1995) 301–337.

[27] A. Buades, B. Coll, J. Morel, A non-local algorithm for image denoising, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2005.

[28] J. Huang, H. Xie, Asymptotic oracle properties of SCAD-penalized least squares estimators, IMS Lecture NotesCMonograph Series 55 (2007) 149C–166.

About the Author—ADITYA CHOPRA is an undergraduate researcher at the School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore. He completed his Bachelor of Technology in Information Technology in July 2009 from VIT University, India. He has been involved in different projects related to face recognition, image restoration, and cryptography. His areas of interests include applied statistics, probability theory and computational optimization.

About the Author—HENG LIAN received the B.S. degrees in Mathematics and Computer Science from the University of Science and Technology of China in 2000, the M.S. degree in Computer Science, M.A. degree in Economics, both from Brown University in 2005, and the Ph.D. degree in Applied Mathematics from Brown University in 2007. He has been with the Division of Mathematical Sciences at Nanyang Technological University since July 2007, where he is currently an assistant professor. His research interests include statistical computation, pattern recognition and image analysis.