J. KSIAM Vol.18, No.2, 143–156, 2014 http://dx.doi.org/10.12941/jksiam.2014.18.143

A REVIEW ON DENOISING

YOON MO JUNG[1]

[1]DEPARTMENT OF COMPUTATIONAL SCIENCE AND ENGINEERING, YONSEI UNIVERSITY, SEOUL 120-749, KOREA
E-mail address: ymjung@yonsei.ac.kr

ABSTRACT. This paper aims to give a quick view on denoising without comprehensive details. Denoising can be understood as removing unwanted parts in signals and images. Noise incorporates intrinsic random fluctuations in the data. Since noise is ubiquitous, denoising methods and models are diverse. Starting from what noise means, we briefly discuss a denoising model as maximum a posteriori estimation and relate it with a variational form or energy model. After that we present a few major branches in image and signal processing; filtering, shrinkage or thresholding, regularization and data adapted methods, although it may not be a general way of classifying denoising methods.

1. INTRODUCTION

In a broad sense, denoising is removing noise which may be unwanted parts in signals, images, measurements, data, and so on. Imaging or data acquisition devices, including diverse modalities from daily-life devices such as cellular phones and digital cameras to medical imaging devices such as CT and MRI output noisy measurements of incoming signals. Noise incorporates intrinsic random fluctuations in the data. Fundamentally, noise is *ubiquitous*. Usually an elementary denoising process is already installed in the device itself and further operations can be proceeded as post-process if the quality of the acquisition or reconstruction is not satisfactory.

Through various disciplines, tremendous techniques are developed and under developing; typing denoising related keywords in Google will show numerous articles in this topic. Some methods are very problem specific, since they exploit a given specific situation or incorporate physical and engineering circumstances. A general technique or principle such as total variation [20] and soft thresholding [9] still can improve the quality of results. However, finding a universal method which outperforms in overall cases is hopeless; there is no panacea at all, since the scope of denoising is too wide and diverse. Even if one narrows the focus, only on

Received by the editors April 24 2014; Accepted May 8 2014; Published online May 29 2014.

2010 *Mathematics Subject Classification.* 94A08.

Key words and phrases. Noise, Denoising, Filtering, Wavelet Shrinkage, Regularization, Machine learning.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (012R1A1A1015492).

specific sounds or images for example, they contain various situations as if zooming in fractals shows recurrent intricacies.

Although a unified theory might be a daydream of mathematician, developing common mathematical foundations and frameworks is still worthwhile to pursuit. Such efforts lead to speculating proper spaces for signals and images such as BV space and Besov space, or finding a best or sparse representation using wavelets and some orthonormal bases or a dictionary by learning.

This paper gives a short review on denoising and by no means do we intend to give a comprehensive review on it. Denoising belongs to applied science and engineering and the primal matter of concern in denoising is practicability. Theory should come in the second. Deng Xiaoping said that "It doesn't matter whether a cat is white or black, as long as it catches mice." The same tenet may go to denoising. Thus, classifying existing methods in a few categories might not be proper. We illustrate a few tracks of them in this paper and some important methods could be missing or misclassified by author's carelessness or ignorance.

Here is the organization of the paper; we start from what noise means in Section 2. In Section 3, we briefly discuss a denoising model as maximum a posteriori estimation and relate it with a variational form or energy model. Also we discuss some probabilistic foundations and measures of denoising quality are given. After that we present a few major branches in image and signal denoising; filtering in Section 4, shrinkage or thresholding in Section 5, regularization in Section 6 and data adapted techniques in Section 7, although it may not be a general way of classifying denoising methods. Since those methods are usually devised for images and signals, we discuss those methods assuming the given data is an image or signal. But their applications are not limited to those fields and there are also many mutations for other fields. We close this review by giving a brief remark in Section 8.

2. What is noise?

Often a denoising paper starts in this fashion: "From the observation u_o, we want to restore u under the assumption $u_o = u + n$ where n is a Gaussian white noise." Since it is already a well-established problem, possibly no more insight or explanation is necessary to justify the importance and soundness of the problem. Thus, researchers often add random numbers generated by a computer to whatever they already have, signals and images, for examples, and just enjoy games how the computed $\tilde{u}$ is close to u which they originally have. The term 'Gaussian' has mathematical definition; it is the most popular distribution in the probability theory. Then how about 'noise' or 'white'?

From this standpoint, we'd like to discuss noise first, before we discuss de-noising; before we want to remove 'something', we should understand what 'something' is. Merriam-Webster dictionary defines it as 'a loud or unpleasant sound' and 'unwanted electronic signals that harm the quality of something'. The first definition is subjective or psychological but the second one looks objective or scientific. Wikipedia does similarly; it starts with the sentence, "Noise means any unwanted sound." to define noise. However, 'Noise (electronics)' starts with that "In electronics, noise is a random fluctuation in an electrical signal, a characteristic of all electronic circuits." Now, what is noise? Can we define it mathematically?

Consider the following two scenarios: 1. You talk to a friend on a street and a bus passes by. 2. You wait for a bus in a bus station. Is the sound of bus noise? You may consider an opposite situation. Strangers talk next to you but you wait for a bus. Is their talk noise? Basically, one may split the heard into two part; meaningful + meaningless, wanted + unwanted, interesting + uninteresting, and important + unimportant. Noise is just the general name for the second category. Now how one can define the second category mathematically? It may be uninteresting because it is unstructured or unpredictable, i.e., random. In other words, we can define it mathematically as a random process. Then the first category should be structured or ordered; it should be regular mathematically. In short, 'signal + noise'.

We close this section by defining a noise model formally. From the original signal u, we observe noisy version u_o of u:

$$u_o = u + n, \tag{2.1}$$

where n is a random process and u belongs to a certain type of regular signals. It is called the *additive noise model*. As you can guess, one another possible model is *multiplicative* such as $u_o = u \cdot n$. By taking the logarithm, one may derive an additive formulation. Hence we consider the former exclusively.

3. What is denoising?

Removing noise from signals or images is called *denoising*. More specifically, the goal is to reconstruct the original signal u from noisy observation u_0 [6]. If u is an 1-dimensional function, it includes signals such as audio, if 2-dimensional, it includes images, and if 3-dimensional, it includes video (It could be even higher dimensional. what would it be?). It involves designing an operator $\mathcal{D}$ called a denoising processor: $\tilde{u} = \mathcal{D}u_o$. Ideally, if n is given or known, then u can be recovered by the subtraction. so, $\tilde{u} = u_o - n = u$. However, n is random. Thus, to remove noise, we require some prior information.

How much do we know about n? At least its distribution should be known. Usually *homogeneous white* noise is assumed. White noise means the randomness with a constant power spectral density. In other words, it has the same density along frequencies after Fourier transform. It is white as a light, otherwise it is *color* noise. For a discrete signal or image, it is used for independent random variables with zero mean and finite variance. *Homogeneity* implies same means and variances everywhere. Distribution doesn't change spatially, for example.

Gaussian white noise is often assumed; noise follows a normal distribution with zero mean. It is the standard or universal noise model. If the noise property is not known or you have no idea, just assume it. However, often quantities interested are nonnegative, sounds and photos, for examples. Photography records light by means of sensoring photons, which are positive. Noise should also consist of photons, so its mean cannot be zero except the case of no noise, and usually a different distribution is assumed for photons. But under the sunlight, huge amounts of photons are captured by each sensor of digital camera. Thus, the value returned by each sensor can be regarded as a number drawn from a normal distribution by the following theorem:

Theorem 3.1 (Central Limit Theorem (CLT)). *Let X_1, $X_2 \cdots$ be iid random variables with $EX_i = \mu$ and $var(X_i) = \sigma^2 < \infty$. Let $S_k = X_1 + \cdots + X_k$. Then $\frac{S_k - k\mu}{\sigma n^{1/2}}$ converges weakly to a standard normal distribution.*

When the light source is about constant, the number of photons received by each sensor fluctuates around its average by CLT. One may set the mean of noise zero if it is small enough, or by mean-shifting if considerable. Hence Gaussian white noise can be understood as an ideal noise.

We consider a perfect denoising shortly; how can we guarantee $\tilde{u} = \mathcal{D}u_o = u$? Assume we can generate multiples of data as many as possible:

$$u_o^{(1)} = u + n^{(1)},\ u_o^{(2)} = u + n^{(2)}, \ldots,$$

then the sample mean $1/k \sum_{i=1}^{k} u_o^{(i)}$ will recover the true u in the limit sense by the following theorem:

Theorem 3.2 (Law of Large Numbers (LLN)). *Let X_1, $X_2 \cdots$ be iid random variables with $E|X_i| < \infty$. Let $S_k = X_1 + \cdots + X_k$ and $EX_i = \mu$. Then,*

$$S_k/k \to \mu \text{ a.s. as } k \to \infty.$$

Thus, assuming white noise or mean-zero noise, if we sufficiently collect data, i.e., k is large enough,

$$\bar{u} = \frac{1}{k} \sum_{i=1}^{k} u_o^{(i)} \approx u. \tag{3.1}$$

Often in experimental situations, researchers measure multiple instances to average out errors. Humble human beings also use the same tactic through experiences in everyday life, and such strategies can be justified by Theorem 3.2 (LLN) under proper assumptions. In reality, the limitation is that the number of samples is always limited. For example, taking multiple shots of the exact same scene is fairly circumscribed, due to moving objects such as humans and cars, sudden illumination such as reflection by windows, variations by wind, etc. In some situations, obtaining multiple copies is very costly or impossible. But we note that LLN tells us the fundamental principle for denoising: *taking average*. We also note that mean is the Maximum Likelihood Estimator (MLE) of the normal distribution and it is closely related to least squares, which approximates the optimal solution of overdetermined systems with the most important application, data fitting.

We introduce a general denoising model by invoking statistical modeling; since noise itself is viewed as random phenomena, probabilistic/statistic interpretation and understanding are inevitable. The denoising processor $\mathcal{D}$ can be modeled as *Maximum A Posteriori* (MAP) estimation [7, 6]. By Bayes' formula, the posterior probability given observation u_o is

$$p(u|u_o) = \frac{p(u_o|u)p(u)}{p(u_o)}. \tag{3.2}$$

The *prior* model specifies how images are distributed *a priori*, or equivalently, which images occur more frequently than others. Probabilistically, it specifies the prior probability $p(u)$. Note that u_o denotes the noised data that are observed or measured. The *data* model is to model how u_o is generated from u, or to specify the conditional probability $p(u_o|u)$. Now, the denoising processor $\mathcal{D}$ is achieved by solving the MAP problem $\max_u p(u|u_o)$, which is equivalent to maximizing the product of the prior model and the data model, since the denominator is a fixed normalization constant once u_0 is given. In words, we seeks '*what is the most plausible u under the given observation u_o?*'.

An *variational form* or *energy model* can be driven from MAP (3.2). Under the notice that probability distributions are often expressed by the exponential functions, by taking the logarithm on the right hand side of (3.2), we have

$$\log p(u_o|u) + \log p(u) - \log p(u_o).$$

Since u_o is already observed, the last term may be dropped. By changing sign and replacing notations properly, we have the following variational form [7]:

$$\min_u E[u] + \frac{\lambda}{2} E[u_0|u], \tag{3.3}$$

where $\lambda/2$ is the Lagrange multiplier. The first term is called the *image prior* or the *regularity term* and the second term is called the *data-fitting term* or *data-fidelity term.* The Lagrange multiplier λ expresses the balance between prior and fitting. Due to the Lagrange multiplier, the variational form is closely related to the following constrained optimization problems:

$$\begin{cases} \min_u E[u] \\ \text{subject to } E[u_0|u] \leq C_1 \end{cases} \quad \text{and} \quad \begin{cases} \min_u E[u_0|u] \\ \text{subject to } E[u] \leq C_2 \end{cases}. \tag{3.4}$$

Designing a proper image prior and a data-fitting is called the image modeling and deriving a suitable image model on the given situation is very crucial in image and signal processing.

We close this section by introducing methodologies of measuring the signal quality. Signal-to-noise ratio (SNR) is often adopted for such a criterion, which compares the level of a desired signal to the level of background noise [23]. Although the perception of human beings on the quality may be a little different from SNR, it provides a value-neutral quantity. More specifically, SNR is defined as the ratio between the variance of a signal and that of noise:

$$\text{SNR} = \frac{\sigma_u^2}{\sigma_n^2} = \frac{E\|u\|^2}{E\|u_o - u\|^2}, \tag{3.5}$$

where σ_u and σ_n represent the standard deviations of the signal and noise, respectively. If u is replaced by $\mathcal{D}u = \tilde{u}$, it can be treated as the numerical value of the *risk* [16]. SNR is often expressed using the logarithmic scale $\text{SNR}_{\text{dB}} = 10\log_{10}\text{SNR}$, measured in decibels. One another available alternative is peak signal-to-noise ratio (PSNR). It measures the ratio between the maximum possible power of the signal and the power of noise. Its technical definition can be easily found in the literature, even in Wikipedia [22]. Lastly, *method noise* is recently introduced by Buades *et al* [3] to check which geometrical features or details are preserved and which are eliminated after image denoising. It is defined as the difference between an image u and its denoised version $\mathcal{D}u = \tilde{u}$:

$$n(\mathcal{D}, u) = u - \mathcal{D}u.$$

4. Filtering

In general, a *filter* is a device or process that removes some unwanted components or features from a signal [21]. This category includes the most traditional but the most widely used methods since the structure is simple, so a fast or real-time implementation is possible, although it may be inaccurate compared to other sophisticated methods.

It is reasonable to assume the *local homogeneity* in a neighborhood; if we assume some regularity as a function, values should be similar locally. For denoising purpose, one may take *local average* as a substitute or approximate for the denoising scheme (3.1) by LLN. For example, one may apply the following *averaging filter* or *lowpass filter* [13]:

$$\tilde{u}(i,j) = \frac{1}{|\mathcal{N}|} \sum_{(i',j')\in\mathcal{N}} u_o(i',j'),$$

where $\mathcal{N}$ is a neighborhood of (i,j). If $\mathcal{N} = \{(i+k, j+m) \mid k, m = 0 \text{ or } \pm 1\}$ is chosen for the neighborhood, 9 point average is taken. More generally, a *weighted average* can be taken:

$$\tilde{u}(i,j) = \sum_{(i',j')\in\mathcal{N}} w(i',j')u_o(i',j'),$$

where $\sum_{(i',j')\in\mathcal{N}} \omega(i',j') = 1$. Such time or space-invariant filter can be rewritten as a convolution:

$$\tilde{u}(x,y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} w(s,t)u_o(x-s, y-t) := \omega * u_o. \tag{4.1}$$

If the weight w is chosen from a normal distribution according to the distance from the origin, it can be regarded as the solution to the heat equation which performs averaging infinitesimally. If the explicit forward time method with the five-point stencil discrete Laplacian is applied to the standard heat equation, we have

$$\frac{u^{t+\triangle t}_{(i,j)} - u^{t}_{(i,j)}}{\triangle t} = \frac{u^{t}_{(i+1,j)} + u^{t}_{(i-1,j)} + u^{t}_{(i,j+1)} + u^{t}_{(i,j-1)} - 4u^{t}_{(i,j)}}{h^2}.$$

By a rearrangement, we derive the following iteration of the weighted mean filter:

$$u^{t+\triangle t}_{(i,j)} = (1-4\alpha)u^{t}_{(i,j)} + \alpha\left[u^{t}_{(i+1,j)} + u^{t}_{(i-1,j)} + u^{t}_{(i,j+1)} + u^{t}_{(i,j-1)}\right], \tag{4.2}$$

with $\alpha = \triangle t/h^2 < 1/4$. Actually one may glimpse the relation between random walk and diffusion as a continuum limit.

If the assumption on local homogeneity is broken, it will average out inhomogeneity. Jumps or edges in an image are crashed and the image becomes *blurry*. Note that the heat equation is also an isotropic diffusion. That is one clear drawback of the averaging filter. To fix it, many other filters are introduced using other statistics such as median filter.

If Fourier transform is taken on the convolution in (4.1),

$$\widehat{\tilde{u}}(\omega) = \hat{w}(\omega)\hat{u}_o(\omega).$$

Thus, it can be viewed as the modulation by $\hat{w}$ along the frequency spectrum of u_o, attenuating high frequency bands for example. Thus, one may also modulate the magnitude along the spectrum by attenuating or strengthening, which is called *frequency filtering*. The former is called *spatial filtering*.

If one considers discrete Fourier transform (DFT) on the one dimensional signal $u_o[m]$, $m = -M, \cdots, M$,

$$u_o[m] = \frac{1}{2M+1} \sum_{k=-M}^{+M} \hat{u}_o[k] \exp\left(\frac{i2\pi km}{2M+1}\right), \text{ where } \hat{u}_o[k] = \sum_{m=-M}^{+M} u_o[m] \exp\left(\frac{-i2\pi km}{2M+1}\right). \tag{4.3}$$

One may truncate the high frequency terms treating high oscillations as noise:

$$\tilde{u}[m] = \frac{1}{2M+1} \sum_{k=-\widetilde{M}}^{+\widetilde{M}} \hat{u}_o[k] \exp\left(\frac{i2\pi km}{2M+1}\right), \tag{4.4}$$

for some $0 \leq \widetilde{M} < M$. It is a crude denoising in the frequency domain and this truncated DFT is also related to the solution of the heat equation by separation of variables which leads to Fourier series expansion, since the high frequency terms exponentially decay.

This type of truncation is also related to the *lossy compression*. Assuming the evenness of the image, DFT is reduced to discrete cosine transform (DCT), and a proper selection and thresholding of the DCT coefficients is the main idea of the popular image compression technique *jpeg*. Such frequency filtering is crucial in many places, such as wireless communication and imaging areas including MRI and CT images.

5. Shrinkage or Thresholding

To reach the idea of wavelet shrinkage by Donoho and Johnstone [10], we follow the logic in [16]. For such purpose, we start from the *Bayesian decision theory*, which is originated from Bayes' formula (3.2). The *risk* of the denoiser $\mathcal{D}$ of u_o is the average *loss* with respect to the probability distribution of the noise n:

$$r(\mathcal{D}, u) = \mathbb{E}\{\|u - \mathcal{D}u_o\|^2\} \tag{5.1}$$

with $u_o[m] = u[m] + n[m]$, $m = 0, \cdots M-1$, similar to (2.1). If one assumes the *prior probability distribution* π, we can define *Bayes risk*, which is the expected risk with respect to π:

$$r(\mathcal{D}, \pi) = \mathbb{E}_\pi\{r(\mathcal{D}, u)\}. \tag{5.2}$$

Definitely we want to minimize Bayes risk, which yields *minimum Bayes risk*:

$$r(\pi) = \inf_{\mathcal{D} \in \mathcal{O}} r(\mathcal{D}, \pi),$$

where $\mathcal{O}$ is the set of all operators. If we restrict such $\mathcal{D}$ on the set $\mathcal{O}_l$ of all linear operators, the optimal operator is called the *Wiener filter* and it can be derived by using covariance matrices:

Theorem 5.1 (Wiener filter). *If the signal u and noise n are independent with the covariance matrices R_u and R_n, respectively, then the Wiener filter that minimizes $E\{\|\tilde{u} - u_o\|^2\}$ is*

$$\tilde{u} = R_u(R_u + R_n)^{-1}u_o. \tag{5.3}$$

Notice that if R_u and R_n are uncorrelated, and so are diagonal with the variances $\sigma_u^2[m]$ and $\sigma_n^2[m]$, the relation (5.3) is the following ratio:

$$\tilde{u}[m] = \frac{\sigma_u^2[m]}{\sigma_u^2[m] + \sigma_n^2[m]} u_o[m]. \tag{5.4}$$

We can clearly interpret it; if the variance of the signal is relatively larger than that of the noise, we trust the observed data $u_o[m]$ and don't shrink much. On the other hand, if that of the signal is relatively smaller, it is very possible to be noise so we shrink.

Since the covariance operator is symmetric and positive semi-definite, it can be diagonalized using eigenvectors with decreasing order of eigenvalues, called the *Karhunen-Loéve basis* or *Principal Component Analysis (PCA)*. If the covariance operators R_u and R_n are diagonalizable under the same Karhunen-Loéve basis, the equation (5.4) is accomplished. We also remark that the standard heat operator is also symmetric and positive semi-definite, it can be diagonalizable by sinusoidal waves. Its discrete version is given in (4.3), and then one may speculate the resemblance between (4.4) and (5.4).

It is generally not possible to compute the optimal Bayes estimator. To avoid such complexity, classical strategies choose a linear operator, although the minimum risk among linear estimators may be far beyond the minimum risk from all estimators. Thus we consider a particular class of nonlinear estimators that are diagonal in a basis $\mathcal{B}$.

In the basis $\mathcal{B} = \{v_m\}_{0\leq m<M}$, u_o has the following basis expansion:

$$u_o = \sum_{m=0}^{M-1} u_o^{\mathcal{B}}[m]v_m \quad \text{where} \quad u_o^{\mathcal{B}}[m] = \langle u_o, v_m\rangle.$$

A *diagonal operator* estimates each $u^{\mathcal{B}}[m]$ by multiplying $u_o^{\mathcal{B}}[m]$ by a factor $a_m(u_o^{\mathcal{B}}[m])$ independently:

$$\tilde{u} = \mathcal{D}u_o = \sum_{m=0}^{M-1} u_o^{\mathcal{B}}[m]a_m(u_o^{\mathcal{B}}[m])v_m. \tag{5.5}$$

For such $a_m(\cdot)$, one may choose the *hard thresholding HT_λ* with the parameter λ:

$$HT_\lambda(x) = \begin{cases} x & \text{if } x \geq \lambda, \\ 0 & \text{if } |x| < \lambda, \\ x & \text{if } x \leq -\lambda. \end{cases} \tag{5.6}$$

One may compare it with the truncation in (4.4) which is linear with fixed $\widetilde{M}$ and assume that low order or smooth terms are signal and high order or highly oscillatory terms are noise. The idea of hard thresholding is that if the amplitude $u_o^{\mathcal{B}}[m]$ is large enough, it is very possible to be a signal and we keep. Otherwise, it is likely to be noise, we throw away. But if the value is often just little above or below the thresholding level λ, is the decision by HT_λ proper? Furthermore, it is discontinuous.

With such concerns, one may consider the following alternative, which is called the *soft thresholding* ST_λ with the parameter λ [9]:

$$ST_\lambda(x) = \begin{cases} x - \lambda & \text{if } x \geq \lambda, \\ 0 & \text{if } |x| \leq \lambda, \\ x + \lambda & \text{if } x \leq -\lambda. \end{cases} \tag{5.7}$$

The operator is now continuous and still shrink even if the signal strength is more than thresholding level λ.

Now one big question is arising: How does one choose the thresholding level λ? What is optimal? Donoho and Johnstone [10] showed the fundamental result that the risk of thresholding estimators is close to that of the *oracle projector*, which is the diagonal estimator by some strong prior knowledge. For more details, we refer to [16].

Theorem 5.2 (Donoho, Johnstone)**.** *Let $\lambda = \sigma\sqrt{2\ln M}$. The risk $r_{th}(u)$ of a hard- or soft-thresholding estimator satisfies for all $M \geq 4$,*

$$r_{th}(u) \leq (2\ln M + 1)\big(\sigma^2 + r_{pr}(u)\big).$$

The factor $2\ln M$ is optimal among diagonal estimators in $\mathcal{B}$:

$$\lim_{M\to\infty} \inf_{\mathcal{D}\in\mathcal{O}_d} \sup_{u\in\mathbb{C}^M} \frac{\mathbb{E}\{\|u - \mathcal{D}u_o\|^2\}}{\sigma^2 + r_{pr}(u)} \frac{1}{2\ln M} = 1,$$

where $\mathcal{O}_d$ is the set of all diagonal operator and r_{pr} is the risk of the oracle projector.

6. Regularization

We recall the variational form (3.3):

$$\min_u E[u] + \frac{\lambda}{2} E[u_0|u].$$

Under this setting, we have to choose a regularity term $E[u]$ and a fitting term $E[u_0|u]$. If we assume the Gaussian white noise, naturally the least square term $\|u - u_0\|_2^2$ comes out for the fitting term $E[u_0|u]$. One popular choice of the regularity term $E[u]$ is also adding the L^2 criteria $\|u\|_2^2$, then it is called *Tikhonov regularization*, which is widely used in inverse problems. One may choose the L^1 norm $\|u\|_1$ which enforces sparsity of u, which will be discussed in the next section. Considering the regularity or differentiability of u, one may choose the L^2 norm or the L^1 norm of the gradient ∇u. If the L^2 norm of the gradient of u is chosen, we consider Sobolev space as a proper function space for the original u. From the energy

$$\min_u \|\nabla u\|_2^2 + \frac{\lambda}{2}\|u - u_0\|_2^2,$$

one may derive the following elliptic equation as Euler-Lagrange equation:

$$-\Delta u + \lambda(u - u_o) = 0.$$

Since the Laplacian Δu is isotropic, blur in the solution u is expected by crushing edges and jumps, similar to the heat equation case (4.2).

To overcome such drawbacks, Rudin, Osher, and Fatemi [20] introduce the following energy, called *total variation (TV)* denoising or *Rudin-Osher-Fatemi* model, assuming u belong to the space of bounded variations (BV):

$$\min_u \|\nabla u\|_1 + \frac{\lambda}{2}\|u - u_0\|_2^2.$$

Its Euler-Lagrange equation is

$$-\nabla \cdot \left[\frac{\nabla u}{|\nabla u|}\right] + \lambda(u - u_o) = 0. \tag{6.1}$$

In (6.1), the term $1/|\nabla u|$ can be understood as a diffusion coefficient or conductivity of heat; if the region is smooth or homogeneous, $|\nabla u|$ is small. Thus the diffusion coefficient $1/|\nabla u|$ is large, local fluctuations are averaged out by diffusion. Along edges and jumps, $|\nabla u|$ is very large and so $1/|\nabla u|$ is small, hence there is a little diffusion due to the low conductivity. Thus edges and jumps are kept. Actually, BV space allows discontinuity. BV space may be considered as the space of piecewise differentiable functions. Meanwhile, Sobolev space doesn't allow such discontinuity.

TV denoising model and the anisotropic diffusion by Perona and Malik [19] triggered the researches in image processing by mathematicians and image processing problems including denoising are popularized in applied mathematics, especially in the areas of partial differential equations and numerical analysis.

After introducing a discretization, solving a PDE such as (6.1) numerically becomes the iteration of a discrete filter, similar to (4.2). Then, how is it different from designing a discrete filter directly? Basically, PDE theory and numerical analysis answer the stability issues and the behavior of the limit by iterations. Even when we apply a discrete filter, if denoising is not satisfactory, we may try to apply the filter more than one time. To forecast what happens eventually, one may investigate PDE-related to the filter and invoke the theories for stabilities and limits.

Usually the solution of a PDE satisfies some types of smoothness. However, so-called *texture* such as patterns in furs, rocks and soils is not smooth at all. Due to the regularity assumptions, details and fine structures behave as noise in function aspects. Thus, the denoising scheme induced by PDE washes out them. Human beings also recognize them as meaningful or geometric objects. It is a typical defect in the variational PDE methods.

Starting from Rudin-Osher-Fatemi model, Yves Meyer considers the 'texture + noise' component as an 'oscillating pattern' which is defined by Besov norm estimates and develops so called *Meyer G norm* [17]. In other words, if one wants to analyze 'texture' mathematically, various Besov-type function spaces may be necessary. Combining TV model and oscillatory functions by Meyer, Osher, Sole and Vese [18] developed a model for image restoration and image decomposition into cartoon and texture by using the negative Sobolev space H^{-1}.

7. Data Adapted Methods

Buades, Coll and Morel introduced the following *nonlocal means* algorithm [3]. Shortly speaking, it estimates the value of x as an average of the values of all the pixels whose Gaussian neighborhood looks like the neighborhood of x:

$$u(x) = \mathrm{NL}(u_o)(x) = \frac{1}{C(x)} \int_\Omega \exp\left(-\frac{(G_a * |u_o(x+.) - u_o(y+.)|^2)(0)}{h^2} \right) u_o(y)dy, \tag{7.1}$$

where G_a is the Gaussian kernel with the standard deviation a, h acts as a filtering parameter, and $C(x) = \int_\Omega \exp\left(-\frac{(G_a*|u_o(x+.)-u_o(z+.)|^2)(0)}{h^2} \right) dz$ is the normalizing factor. Lastly,

$$G_a * |u_o(x+.) - u_o(y+.)|^2)(0) = \int_{\mathbb{R}^2} G_a |u_o(x+t) - u_o(y+t)|^2 dt,$$

which measures the distance between the neighborhoods of x and y under Gaussian weights.

They noticed that every small window in a natural image has many similar windows in the same image. In that sense it is highly *redundant*. Instead of acknowledging the given image as one instance, the given image can be regarded as a composition of local images, called *patches*. For example, one may consider the 5×5 neighboring patch for each pixel, then the given image consists of many small images. If there are many alike patches by *selfsimilarity* and we take the average among them, we obtain the ideal denoising method in (3.1). The locations of those patches may not be close together, that is why it is called nonlocal means.

The average filter and frequency filtering in Section 4 and variational PDE methods in Section 6 belong to local smoothing methods, which may reconstruct main geometrical configurations but fail to preserve the fine structure, details, and texture. But nonlocal means bypass such restriction by exploiting innate redundancy and selfsimilarity.

The truncation in (4.4), hard thresholding (5.6) and soft thresholding (5.7) try to make its coefficient *sparse* to achieve denoising under some orthogonal basis. If we have a redundant set consisting of clean patches or good prototypes, where the redundancy means spanning same set but possibly linearly dependent, than the chance for sparsity will increase. Such a code book is called a *dictionary* and those members are called *atoms*.

Let A be a given dictionary, where each column is an atom. Then we can consider the following optimization problem:

$$\min_x \|x\|_0 \quad \text{subject to } \|Ax - u_o\| < \epsilon, \tag{7.2}$$

where $\|x\|_0$ is the number of nonzeros in x. Since the dictionary A is redundant, if A is a $m \times p$ matrix, $m \leq p$ and thus the equation is underdetermined. Considering $Ax = u_o$, if one assumes at least one solution, there must be infinitely many; under the assumption that the matrix A have full rank, the dimension of the solution is $p - m$. Among them we seek the sparsest solution by penalizing the number of nonzero in x and $\tilde{u} = Ax$ is represented by a few columns in A. If the dictionary A consists of good and clean patches or prototypes, $\tilde{u} = Ax$ is a denoised version of u_o.

Due to nonconvexity and non-differentiability, $\|\cdot\|_0$ is often replaced by $\|\cdot\|_1$ and then it is called the *basis pursuit* denoising problem [8]:

$$\min_x \|x\|_1 \quad \text{subject to } \|Ax - u_o\| < \sigma^2, \tag{7.3}$$

which is closely related to compressed sensing [4, 5]. Their general form is described in (3.3) and (3.4). Note that soft thresholding (5.7) is a closed-form solution for $\min_x \lambda|y| + \frac{1}{2}|y-x|^2$, and one can quickly figure out the closed-form solution for

$$\min_x \lambda\|x\|_1 + \frac{1}{2}\|y-x\|_2^2.$$

With the Lagrange multiplier or thresholding level λ, soft thresholding is a basic tool to solve L^1 problems including (7.3).

Although pre-constructed dictionaries consisting of existing bases such as DFT, DCT, and wavelets lead to fast transforms of the complexity $O(m \log m)$, they are typically limited to sparsify the signals and images of interest. Thus, generating a dictionary to sparsify only a certain type of signals or images, called *dictionary learning* is developed in the machine learning point of view:

$$\min_{A,\{x_i\}_{i=1}^M} \|x_i\|_0 \quad \text{subject to } \|y_i - Ax_i\| < \epsilon,\ 1 \le i \le M, \tag{7.4}$$

where $\{y\}_{i=1}^M$ is a *training database* consisting of typical signals/images of interest. After running a computational algorithm, The dictionary A should consists of good and clean patches or prototypes by learning or experiencing through the data base $\{y\}_{i=1}^M$. K-SVD algorithm [1] is such an instance, which is closely related to k-means clustering and singular value decomposition or PCA. For more details, we refer to [12].

8. CONCLUSION

In this paper, we review noise, denoising, and its major branches in somewhat unorganized and messy fashion, which is the nature of noise. We emphasize again the fundamental principle for denoising; *taking average*. Actually all methods try proper averaging without obtaining several copies. The problem of denoising will long live, since most algorithms are away from a desirable level of applicability. Furthermore, more and more new problems flow into this enterprise.

We remind that we don't restrict the dimension of the problem. Now what is that data with the dimension more than 3? They may be data from internet, social networks, medical clinics, financial transactions, to name a few; we live in a deluge of data. Now, *big data* is an overused buzzword, the stuff people don't understand but want to sell. It may be described as tremendous items in a higher dimensional space and there could be some weird phenomena which can not be observed in low dimensional spaces we got used to, such as *curse of dimensionality*, or *concentration of measure* [15]. However, they may be divided into two categories; 'information + non-information', like 'signal + noise'. Thus, we may denoise them. Furthermore, we also observe that denoising is closely related to other problems such as lossy compression and representation. They are all related to the *dimensionality reduction* of data.

Turning back to images, denoising is a part of the image restoration problem, since noise is a part of image degradation:

$$u_o = Ku + n,$$

where K is a blur kernel explaining blurring in the image, defocusing and motion blur, for examples. Thus, image restoration consists of denoising and deblurring. For more reading, we recommend Gonzalez and Woods [13] for general and engineering aspects. For mathematical point of view, Chan and Shen [6] and Aubert and Kornprobst [2] are recommended. Also, Chan, Shen, and Vese's review paper [7] is also an excellent introductory. For wavelets and related signal/image processing, one has to consult with Mallat [16], and for sparse and redundant representation, Elad [12] is excellent. Toward machine learning and data analysis, Duda, Hart and Stork [11] is popular, and Hastie, Tibshirani, and Friedman [14] is advanced. Actually, there must be many many excellent books in this huge world, which the author may neither know nor explore.

ACKNOWLEDGMENTS

The author would like to thank his former advisor Jianhong Shen, who opened a new door for him. Through the door, there was a new world consisting of harmony and cacophony, sound and noise, conviction and skepticism, determinism and accidentalism . . . .

REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein. k -svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, Nov 2006.

[2] Gilles Aubert and Pierre Kornprobst. *Mathematical problems in image processing*, volume 147 of *Applied Mathematical Sciences*. Springer, New York, second edition, 2006. Partial differential equations and the calculus of variations, With a foreword by Olivier Faugeras.

[3] A. Buades, B. Coll, and J. M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.*, 4(2):490–530, 2005.

[4] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[5] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[6] Tony F. Chan and Jianhong Shen. *Image processing and analysis*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.

[7] Tony F. Chan, Jianhong Shen, and Luminita Vese. Variational PDE models in image processing. *Notices Amer. Math. Soc.*, 50(1):14–26, 2003.

[8] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[9] David L. Donoho. De-noising by soft-thresholding. *IEEE Trans. Inform. Theory*, 41(3):613–627, 1995.

[10] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[11] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[12] Michael Elad. *Sparse and redundant representations*. Springer, New York, 2010. From theory to applications in signal and image processing, With a foreword by Alfred M. Bruckstein.

[13] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice Hall, 3 edition, August 2007.

[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009.

[15] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[16] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier/Academic Press, Amsterdam, third edition, 2009.

[17] Yves Meyer. Oscillating patterns in some nonlinear evolution equations. In *Mathematical foundation of turbulent viscous flows*, volume 1871 of *Lecture Notes in Math.*, pages 101–187. Springer, Berlin, 2006.

[18] Stanley Osher, Andrés Solé, and Luminita Vese. Image decomposition and restoration using total variation minimization and the H^{-1} norm. *Multiscale Model. Simul.*, 1(3):349–370 (electronic), 2003.

[19] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:629–639, 1990.

[20] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, November 1992.

[21] Wikipedia. Filter (signal processing). `http://en.wikipedia.org/wiki/Filter_(signal_processing)`.

[22] Wikipedia. Peak signal-to-noise ratio — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio`.

[23] Wikipedia. Signal-to-noise ratio — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Signal-to-noise_ratio`.