Image Denoising Based on Nonlocal Bayesian Singular Value Thresholding and Stein's Unbiased Risk Estimator

Caoyuan Li, Hong-Bo Xie, *Member, IEEE*, Xuhui Fan, Richard Yi Da Xu, Sabine Van Huffel, *Fellow, IEEE*, Scott A. Sisson, and Kerrie Mengersen

***Abstract*— Singular value thresholding (SVT)- or nuclear norm minimization (NNM)-based nonlocal image denoising methods often rely on the precise estimation of the noise variance. However, most existing methods either assume that the noise variance is known or require an extra step to estimate it. Under the iterative regularization framework, the error in the noise variance estimate propagates and accumulates with each iteration, ultimately degrading the overall denoising performance. In addition, the essence of these methods is still least squares estimation, which can cause a very high mean-squared error (MSE) and is inadequate for handling missing data or outliers. In order to address these deficiencies, we present a hybrid denoising model based on variational Bayesian inference and Stein's unbiased risk estimator (SURE), which consists of two complementary steps. In the first step, the variational Bayesian SVT performs a low-rank approximation of the nonlocal image patch matrix to simultaneously remove the noise and estimate the noise variance. In the second step, we modify the conventional SURE full-rank SVT and its divergence formulas for rank-reduced eigen-triplets to remove the residual artifacts. The proposed hybrid BSSVT method achieves better performance in recovering the true image compared with state-of-the-art methods.**

***Index Terms*— Image denoising, noise variance estimation, singular value thresholding, variational Bayesian inference, Stein's unbiased risk estimator.**

I. Introduction

SINGULAR value thresholding (SVT) aims to recover an approximately low-rank data matrix $\boldsymbol{X}$ from a noisy observation matrix $\boldsymbol{Y}$ by shrinking its singular values (SV). SVT has been widely applied in signal and image processing, computer vision, and pattern recognition. It is well known that, if $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \sum_{i=1}^{\min(m,n)} \lambda_i u_i v_i^\top$ is a singular value decomposition (SVD) for $\boldsymbol{Y}$, the hard thresholding estimator simply truncates the singular spectrum by setting some of the SV to zero. The level of the SV truncating can be determined by cross-validation; however, this approach can be unstable and computationally expensive [1], [2]. Donoho and Gavish [3] proposed an optimal hard threshold of $4/\sqrt{3}/\sqrt{m}\sigma$ for an $m \times m$ square matrix with known noise variance σ^2. Under the framework of nonlocal image denoising, the representative hard threshold algorithms include [4], [5] and the very recent [6].

In contrast to hard thresholding, soft thresholding aims to shrink each SV using the function

$$\hat{\lambda}_i = \lambda_i(1 - \frac{\tau}{\lambda_i})_+, \tag{1}$$

where $x_+ = \max(x, 0)$ for $x \in \mathbb{R}$. Candes *et al.* [7] provided a closed-form expression of Stein's unbiased risk estimate (SURE) to select the threshold $\tau > 0$. Dong *et al.* [8] extended the principle of wavelet BayesShrink to determine the soft threshold. Their spatially adaptive iterative singular value thresholding (SAIST) method estimates the threshold corresponding to each SV based on the locally estimated signal variance and overall noise variance. To exploit the low-rank structure of the patch matrix, substantial effort has been expended on rank-penalized methods and convex relaxation or, for computational reasons, penalization of the nuclear norm of the matrix. Gu *et al.* [9] assumed that the noise energy is evenly distributed over each subspace spanned by the eigen-triplets. Specific thresholds are then determined by the individual SV and noise variance. Although this method is termed the weighted nuclear norm minimization (WNNM), it lies in the category of SV soft thresholding methods. Since the algorithms in [8], [9] consider the relative importance of different SVs, the quality of the recovered image is very competitive in terms of the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). Several variants of SAIST and WNNM have been developed [10]–[13].

Recently, Josse and Sardy [14] defined a two-parameter threshold function

$$\hat{\lambda}_i = \lambda_i(1 - \frac{\tau^\eta}{\lambda_i^\eta})_+, \tag{2}$$

Manuscript received July 18, 2018; revised March 7, 2019; accepted April 15, 2019. Date of publication April 26, 2019; date of current version August 1, 2019. This work is funded by the Australian Research Council (ARC) Laureate Program and the ARC Centre of Excellence Program. S. A. Sisson is also supported by the ARC Discovery Project DP160102544. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Keigo Hirakawa. *(Corresponding author: Hong-Bo Xie.)*

C. Li is with the School of Computer Science and Technology, Beijing Institute of Technology (BIT), Beijing 100081, China, and also with the Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Ultimo, NSW 2007, Australia (e-mail: licaoyuan@bit.edu.cn).

H.-B. Xie and K. Mengersen are with the ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology, Brisbane, QLD 4001, Australia (e-mail: hongbo.xie@qut.edu.au; k.mengersen@qut.edu.au).

X. Fan and S. A. Sisson are with the School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2033, Australia (e-mail: xuhui.fan@unsw.edu.au; scott.sisson@unsw.edu.au).

R. Y. D. Xu is with the Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Ultimo, NSW 2007, Australia (e-mail: YiDa.Xu@uts.edu.au).

S. Van Huffel is with the ESAT-Stadius Division, Department of Electrical Engineering, KU Leuven, 3001 Leuven, Belgium (e-mail: sabine.vanhuffel@esat.kuleuven.be).

Digital Object Identifier 10.1109/TIP.2019.2912292

1057-7149 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

which encompasses hard thresholding for $\eta \to \infty$ and soft thresholding when $\eta = 1$. Their Monte Carlo simulation revealed that such a trade-off between soft and hard thresholding yields the best performance in terms of MSE on both low-rank and general signal matrices across different signal-to-noise ratio regimes. Following the same principle, Verbanck *et al.* [15] suggested a regularized version of PCA (rPCA) that essentially selects a certain value for the rank and shrinks the corresponding SVs. Jia *et al.* [16] defined this problem as rank constrained nuclear norm minimization (RNNM), in which the rank and the extent of thresholding are controlled separately. The thresholding function is accordingly denoted as:

$$\hat{\lambda}_i = \begin{cases} \lambda_i(1 - \frac{\tau^\eta}{\lambda_i^\eta})_+ & i = 1, \ldots, r \\ 0 & i = r+1, \ldots, \min(m, n), \end{cases} \tag{3}$$

where r is the selected rank with $r < \min(m, n)$.

These methods not only aim to better approximate the original low-rank structure of the patch matrix, but also differentiate the importance of each rank component. Due to this balance between the reduced rank and threshold, these algorithms can achieve superior results compared with benchmark methods such as nonlocal means and BM3D [17].

However, there are a number of issues shared by these existing methods. Firstly, almost all of the aforementioned methods and their variants require the noise variance σ^2 to be known [3]–[8], [10]–[16], which is not realistic in practice. An extra step is therefore required to pre-determine the noise variance. While the denoising performance of these methods can be substantially degraded when using poor estimates of the noise variance, the numerically impressive results of a number of approaches including BM3D, SAIST, WNNM, and RNNM are obtained simply because of the assumption that the exact noise variance is known [7], [15], [16], [18]–[22]. The impact of the error in the estimation of the noise variance on recovered images has not been examined, which casts doubt on the actual performance of these approaches. There are also other free parameters that need to be empirically determined. For example, two extra constants control the weights in WNNM and the pre-specified order in low-rank approximation methods [8], [15], [16]. Furthermore, in order to thoroughly remove the noise, the iterative regularization scheme is frequently adopted in these methods. The variance of the residual noise for the next iteration is estimated from the difference between the initial variance and that of the filtered noise at the previous iteration. The initial error in the estimation of the noise variance therefore propagates and accumulates at each iteration, ultimately degrading the quality indexes in real-world applications. Another issue is that these approaches are largely based on the conventional singular value decomposition (SVD) in the least squares sense. For high-dimensional parameter spaces, the MSE of a least-squares method is often larger than that of a Bayesian estimator [23]. It is also highly susceptible to outlier values in the data. Finally, low-rank approximation-based algorithms, as well as BM3D and WNNM, tend to produce a weak noise-like pattern in low contrast areas of the image when the noise level is moderate or high [8], [24]. This is because the noise in similar patches is partially correlated, which can lead to the incorrect estimation of low-rank patterns as the output of these algorithms [18].

Contribution of This Work: To address the many issues identified above, we propose a unified nonlocal image denoising framework based on variational Bayesian inference and Stein's unbiased risk estimator (BSSVT). This generic nonlocal denoising framework consists of two complementary steps. In the first step, the variational Bayesian model performs a low-rank approximation of the noisy patch matrix. This is functionally equivalent to other low-rank approximation or nuclear norm minimization methods. More importantly, the noise variance is a latent parameter which is automatically inferred, so does not need to be provided beforehand. The SURE criterion has been employed in a variety of denoising problems to optimize regularization parameters for minimizing the estimation risk or MSE [25]–[29]. With the noise variance obtained via the Bayesian model, the second step carries out the SURE-based singular value thresholding on the rank-reduced eigen-triplets to optimally refine the SVs. This further attenuates the very weak noise-like pattern in low contrast areas of the image and reduces artifacts around edges, overcoming the shortcomings of low-rank approximations [18].

Our main contributions are summarized as follows: (a) We formulate a hybrid nonlocal image blind denoising framework which exploits both Bayesian low-rank approximation and Stein's unbiased risk estimation; (b) We adopt a variational Bayesian model to approximate the low-rank structure of the patch matrix, which simultaneously performs the noise removal and noise variance estimation. This Bayesian model was first developed in [30], with a focus on general principal component analysis. In this study, we apply and extend its construction for image processing applications. Since the original model in [30] needs to try out all possible values of the rank to determine the reduced rank, the huge computational burden makes the model non-viable for patch-based image restoration tasks. We employ the automatic relevance determination principle [31] to automatically prune the rank, which significantly relieves the computational cost; (c) We modify the full-rank Stein's unbiased risk estimator and its divergence formulas for use in reduced-rank singular value thresholding. This modified SSVT algorithm directly maximizes the PSNR by refining the optimal threshold that minimizes the MSE estimation of rank-reduced eigen-triplets; (d) We apply the modified SURE model on the rank-reduced eigen-triplets to enhance the initial low-rank approximation and to produce a more precise estimate of the original image.

The experimental results demonstrate that our proposed BSSVT approach has superior performance in comparison with the state-of-the-art methods in terms of both PSNR and SSIM. The rest of this paper is organized as follows. Section II reviews Bayesian matrix orthogonal factorization, image noise variance estimation and the SURE criterion. In Section III we elaborate on the details of the Bayesian model for low-rank patch recovery in the presence of noise. The nonlocal Stein's unbiased risk estimator is also described in this section.

Experimental results, comparison with the state-of-the-art methods and objective assessments are presented in Section IV. Finally, the Section V discussion concludes the paper.

II. Related Work

The first part of our method is related to Bayesian approaches for orthogonal matrix low-rank approximation and for orthogonal nonnegative matrix factorization [23], [30], [32]–[37]. Hoff [23] presented a full Bayesian singular value decomposition model. However, using Gibbs sampling to estimate the parameters makes it unsuitable for nonlocal image denoising because of its huge computational cost. The singular value may also be negative in this model, leading to further issues. The Bayesian inference on the unknown parameters in [32] was also carried out using Markov chain Monte Carlo (MCMC), while the variational Bayesian PCA algorithm in [30] focused on feature extraction and reduction. Variational inference was employed in [33] to perform SVD; however, this model only considered a prior of a singular vector and omitted singular values. Although [34]–[37] emphasize orthogonality in their models, the basic framework of these models is nonnegative matrix decomposition.

As we have reviewed above, most image denoising approaches are developed based on the assumption of a known noise variance [3]–[8], [10]–[16]. This largely restricts them in terms of practical use. Consequently, the first step of image denoising is often dedicated to estimating the noise variance using the same available image that needs to be denoised. The most well-known noise variance estimator is the scaled Median Absolute Deviation (MAD) method in wavelet denoising [38]. The noise variance is roughly approximated by the median of the absolute value of the wavelet coefficients at the finest decomposition level, which is employed in [8] and its variants. Other methods to estimate the noise variance are mainly based on residual principal components or singular values [39]–[43]. It is very common for the noise variance or precision to be estimated via generative models [23]. However, this has not attracted enough attention to be exploited in the image processing community. To the best of our knowledge, this is the first paper to present a variational Bayesian model to shrink SVs for nonlocal denoising, and in particular, simultaneous noise removal and noise variance estimation.

Low-rank approximations tend to produce a very weak noise-like pattern in flat areas of the image when the noise level is moderate or high [18]. This arises from the fact that the noise in a group of overlapping similar patches is partially correlated, which can incorrectly lead to the reconstruction of a low-rank approximation. The SURE criterion has been well developed to optimally adjust the parameters of a variety of denoising algorithms for edge-preserving filtering and artifact removal [7], [14], [25]–[27]. However, the existing SURE is only applicable to shrinking the full rank eigen-triplets. Here we modify the existing SURE and its divergence formulas to accommodate the rank-reduced eigen-triplets obtained by Bayesian low-rank approximation.

III. Methods

Our proposed BSSVT method consists of two successive and complementary steps: Bayesian singular value thresholding (BSVT) for low-rank approximation representation of nonlocal similarities; and the singular value thresholding based on SURE (SSVT) with respect to the rank-reduced representation. Fig. 1 shows a schematic diagram of BSSVT, in which the leftmost component is the graphic model of the Bayesian SVD and the rightmost component represents the SURE-based shrinker. The details of these steps are presented below.

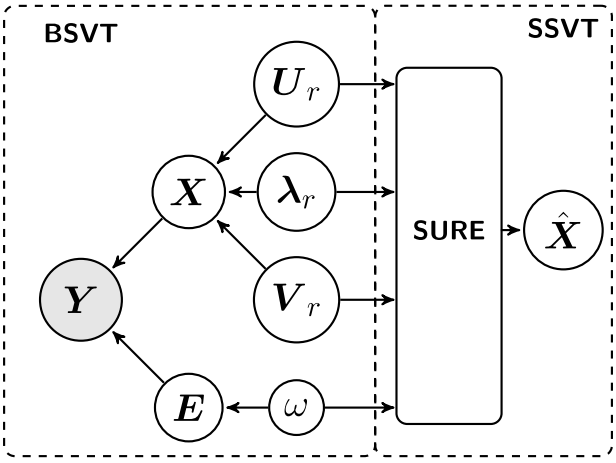


Fig. 1. Schematic diagram of BSSVT to denoise a patch matrix using variational Bayesian inference and SURE criterion.

A. Variational Bayesian Singular Value Thresholding

Under the nonlocal framework, an image is divided into small square blocks, i.e. patches. A patch group matrix is constructed by the vectorization of each patch and its nonlocal neighbors. The final output image is formed by reassembling the individually processed patches. The purpose of variational Bayesian singular value thresholding is to learn this low-rank subspace, while simultaneously providing the noise variance and eigen-triplets for refinement at the second stage.

1) Model Specification: Without loss of generality, we assume that the noisy patch matrix is $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{E}$, where $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$ is composed of n vectorized similar patches with size $\sqrt{m} \times \sqrt{m}$ from a noisy image and $\boldsymbol{E}$ denotes the noise matrix with i.i.d. entries $\boldsymbol{E}_{i,j} \sim \mathcal{N}(0, \omega^{-1})$, where $\mathcal{N}(0, \omega^{-1})$ denotes a Gaussian distribution with mean 0 and precision ω. A natural way to represent the low-rank subspace is to truncate the singular values of the observed matrix $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \sum_{i=1}^{\min(m,n)} \lambda_i u_i v_i^\top$ to $\hat{\boldsymbol{X}} = \boldsymbol{U}_r\boldsymbol{D}_r\boldsymbol{V}_r^\top$ (for $r < \min(m,n)$), which satisfies $\boldsymbol{U}_r \in \mathbb{R}^{n\times r}$, $\boldsymbol{V}_r \in \mathbb{R}^{m\times r}$ and $\boldsymbol{U}_r^\top\boldsymbol{U}_r = \boldsymbol{I}_r$, $\boldsymbol{V}_r^\top\boldsymbol{V}_r = \boldsymbol{I}_r$. Here, $\boldsymbol{D}_r = diag(\boldsymbol{\lambda}_r) \in \mathbb{R}^{r\times r}$ is a diagonal matrix with non-zero singular values in descending order and r is the rank of the low-rank approximation. From the point of view of Bayesian inference, the task is to infer the posterior eigen-triplets of $\boldsymbol{U}_r$, $\boldsymbol{D}_r$ and $\boldsymbol{V}_r$ from their prior distributions and the observed patch matrix. The likelihood of the noisy patch matrix is denoted as:

$$p(\boldsymbol{Y}|\boldsymbol{U}_r, \boldsymbol{D}_r, \boldsymbol{V}_r, \omega, r) = \mathcal{N}(\boldsymbol{U}_r\boldsymbol{D}_r\boldsymbol{V}_r^\top, \omega^{-1}\boldsymbol{I}_n \otimes \boldsymbol{I}_m), \quad (4)$$

where $\boldsymbol{I}_n$ denotes an $n \times n$ identity matrix and $\boldsymbol{I}_n \otimes \boldsymbol{I}_m \in \mathbb{R}^{nm\times nm}$ denotes the Kronecker product of matrices $\boldsymbol{I}_n$ and $\boldsymbol{I}_m$.

Since $\boldsymbol{U}_r$ has orthonormal columns, it is constrained to the Stiefel manifold $\mathcal{S}_{n,r}$ [23]. Therefore, both the prior and posterior distributions of $\boldsymbol{U}_r$ have a support confined to $\mathcal{S}_{n,r}$. The finite area $C(n,r)$ of $\mathcal{S}_{n,r}$ is given by [30], [32], [33]

$$C(n,r) = \frac{2^r \pi^{(1/2)nr}}{\pi^{(1/4)r(r-1)} \prod_{j=1}^{r} \Gamma((1/2)(n-j+1))}, \quad (5)$$

where $\Gamma(\cdot)$ is the gamma function. Similarly, $\boldsymbol{V}_r$ is constrained to the manifold $\mathcal{S}_{m,r}$. We adopt the priors on $\boldsymbol{U}_r$ and $\boldsymbol{V}_r$ to be the least informative, i.e. uniform on $\mathcal{S}_{n,r}$ and $\mathcal{S}_{m,r}$, respectively.

$$p(\boldsymbol{U}_r) = C(n,r)^{-1} \chi(\mathcal{S}_{n,r}), \quad (6)$$
$$p(\boldsymbol{V}_r) = C(m,r)^{-1} \chi(\mathcal{S}_{m,r}), \quad (7)$$

where $\chi()$ denotes the indicator function on the argument set.

In the absence of specific prior knowledge on ω, we use Jeffreys' prior for the precision parameter so that

$$p(\omega) \propto \omega^{-1}, \quad (8)$$

which corresponds to an improper gamma distribution attained when both shape and scale parameters approach zero [44]. The above uninformative priors can be modified in obvious ways if relevant information is available.

We assume that the prior knowledge of $\boldsymbol{D}_r$ can be expressed by an upper bound on the norm of $\boldsymbol{\lambda}_r$:

$$\sum_{i=1}^{r} \lambda_i^2 \leq 1, \quad (9)$$

together with the descending order constraint, so that $\boldsymbol{\lambda}_r$ is confined to the space

$$\mathcal{L}_r = \{\boldsymbol{\lambda}_r | \lambda_1 > \lambda_2 > \cdots > \lambda_r > 0, \sum_{i=1}^{r} \lambda_i^2 \leq 1\}, \quad (10)$$

which is a segment of the unit hyperball $\mathcal{H}_r$. The volume of $\mathcal{L}_r$ is:

$$\mathcal{V}_r = h_r \frac{1}{2^r (r!)} = \frac{\pi^{r/2}}{\Gamma(r/2+1) 2^r (r!)}. \quad (11)$$

where h_r is the volume of $\mathcal{H}_r$. The prior distribution on $\boldsymbol{\lambda}_r$ is then chosen to be uniform on $\mathcal{L}_r$ [30], [44]:

$$p(\boldsymbol{\lambda}_r) = \mathcal{U}(\mathcal{L}_r) = \mathcal{V}_r^{-1} \chi(\mathcal{L}_r). \quad (12)$$

The resulting probabilistic graphical model is shown in the leftmost part of Fig. 1. For notational simplicity, all unknown parameters are collectively denoted by $\Theta = \{\boldsymbol{U}_r, \boldsymbol{\lambda}_r, \boldsymbol{V}_r, \omega\}$. Therefore, the joint distribution of the parameters and data is given by

$$p(\boldsymbol{Y}, \Theta | r) = p(\boldsymbol{Y} | \boldsymbol{U}_r, \boldsymbol{\lambda}_r, \boldsymbol{V}_r, \omega, r) p(\boldsymbol{U}_r) p(\boldsymbol{\lambda}_r) p(\boldsymbol{V}_r) p(\omega). \quad (13)$$

2) Model Learning via Variational Bayesian Inference: We take advantage of a fast variational Bayesian (VB) approximation method to infer the posterior distribution of Eq. (13) [45]–[47]. In particular, we construct a distribution $q(\Theta)$ to approximate the true posterior distribution $p(\Theta|\boldsymbol{Y})$ by minimizing the Kullback-Leibler (KL) divergence:

$$KL(q(\Theta) \parallel p(\Theta|\boldsymbol{Y})) = -\int q(\Theta) \log \frac{p(\Theta|\boldsymbol{Y})}{q(\Theta)} d\Theta \geq 0. \quad (14)$$

The KL divergence is equal to 0 iff $p(\Theta|\boldsymbol{Y})$ is identical to $q(\Theta)$ [48].

Based on the mean field approximation, we assume that the proposed posterior approximation can be factorized as

$$q(\Theta) = q(\boldsymbol{U}_r|\boldsymbol{Y}, r) q(\boldsymbol{\lambda}_r|\boldsymbol{Y}, r) q(\boldsymbol{V}_r|\boldsymbol{Y}, r) q(\omega|\boldsymbol{Y}, r). \quad (15)$$

Applying the VB theorem to Eq. (13), we obtain the following approximate posterior distributions:

$$q(\boldsymbol{U}_r|\boldsymbol{Y}, r) \sim \upsilon MF(\boldsymbol{F}_U), \quad (16)$$
$$q(\boldsymbol{V}_r|\boldsymbol{Y}, r) \sim \upsilon MF(\boldsymbol{F}_V), \quad (17)$$
$$q(\boldsymbol{\lambda}_r|\boldsymbol{Y}, r) \sim t\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_r; \mathcal{L}_r), \quad (18)$$
$$q(\omega|\boldsymbol{Y}, r) \sim \text{Gamma}(\alpha, \beta). \quad (19)$$

Here, $\upsilon MF(\cdot)$ denotes the von Mises-Fisher distribution [49], $t\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_r; \mathcal{L}_r)$ is the truncated normal distribution with support $\mathcal{L}_r$ and Gamma(α, β) denotes the gamma distribution with shape α and rate β. The analytical forms of the above distributions are provided in the Appendix.

The parameters of Eqs. (16)-(19) are given by

$$\boldsymbol{F}_U = \hat{\omega} \boldsymbol{Y} \widehat{\boldsymbol{V}_r} \widehat{\boldsymbol{D}_r}, \quad (20)$$
$$\boldsymbol{F}_V = \hat{\omega} \boldsymbol{Y}^\top \widehat{\boldsymbol{U}_r} \widehat{\boldsymbol{D}_r}, \quad (21)$$
$$\boldsymbol{\mu} = \text{diag}(\widehat{\boldsymbol{V}_r}^\top \boldsymbol{Y}^\top \widehat{\boldsymbol{U}_r}), \quad (22)$$
$$\sigma^2 = \hat{\omega}^{-1}, \quad (23)$$
$$\alpha = \frac{nm}{2}, \quad (24)$$
$$\beta = \frac{1}{2}(\widehat{\boldsymbol{\lambda}_r^\top \boldsymbol{\lambda}_r} + tr(\boldsymbol{Y}\boldsymbol{Y}^\top - 2\boldsymbol{Y} \widehat{\boldsymbol{V}_r} \widehat{\boldsymbol{D}_r} \widehat{\boldsymbol{U}_r}^\top)), \quad (25)$$

where $\widehat{\omega}$ denotes the expectation of ω with respect to $q(\omega)$ and similarly for the other variables.

The VB algorithm requires iteration of Eqs. (20)-(25) until convergence, which in turn requires iterative evaluation of the moments of the distributions (16)-(19):

$$\widehat{\boldsymbol{U}_r} = \boldsymbol{U}_{F_U} G(n, \boldsymbol{D}_{\boldsymbol{F}_U}) \boldsymbol{V}_{\boldsymbol{F}_U}^\top, \quad (26)$$
$$\widehat{\boldsymbol{V}_r} = \boldsymbol{U}_{\boldsymbol{F}_V} G(m, \boldsymbol{D}_{\boldsymbol{F}_V}) \boldsymbol{V}_{\boldsymbol{F}_V}^\top, \quad (27)$$
$$\widehat{\boldsymbol{\lambda}_r} = \boldsymbol{\mu} + \sigma \zeta(\boldsymbol{\mu}, \sigma), \quad (28)$$
$$\widehat{\boldsymbol{\lambda}_r^\top \boldsymbol{\lambda}_r} = r\sigma^2 + \boldsymbol{\mu}^\top \widehat{\boldsymbol{l}_r} - \sigma \rho(\boldsymbol{\mu}, \sigma), \quad (29)$$
$$\hat{\omega} = \frac{\alpha}{\beta}, \quad (30)$$

where $\boldsymbol{U}_{\boldsymbol{F}_U}$, $\boldsymbol{D}_{\boldsymbol{F}_U}$, $\boldsymbol{V}_{\boldsymbol{F}_U}$, $\boldsymbol{U}_{\boldsymbol{F}_V}$, $\boldsymbol{D}_{\boldsymbol{F}_V}$ and $\boldsymbol{V}_{\boldsymbol{F}_V}$ are the SVD parameters of $\boldsymbol{F}_U$ and $\boldsymbol{F}_V$ respectively and the definition of the functions $\zeta(\cdot,\cdot)$, $\rho(\cdot,\cdot)$ and $G(\cdot,\cdot)$ are provided in the Appendix.

If $\widehat{\boldsymbol{U}_r}$ and $\widehat{\boldsymbol{V}_r}$ are formed from scaled singular vectors of the noisy patch matrix $\boldsymbol{Y}$, so that

$$\widehat{\boldsymbol{U}_r} = \boldsymbol{U}_{;r}\boldsymbol{K}_{\boldsymbol{U}}, \tag{31}$$

$$\widehat{\boldsymbol{V}_r} = \boldsymbol{V}_{;r}\boldsymbol{K}_{\boldsymbol{V}}, \tag{32}$$

where $\boldsymbol{U}_{;r}$ and $\boldsymbol{V}_{;r}$ denote the first r columns of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively, $\boldsymbol{K}_{\boldsymbol{U}} = diag(\boldsymbol{k}_{\boldsymbol{U}}) \in \mathbb{R}^{r\times r}$ and $\boldsymbol{K}_{\boldsymbol{V}} = diag(\boldsymbol{k}_{\boldsymbol{V}}) \in \mathbb{R}^{r\times r}$ are the proportionality constants, then Eqs. (20)-(25) and (26)-(30) can be greatly simplified and each iteration using these equations satisfies (31) and (32). Detailed derivations of these equations are given in [30]. For the above inference, it is assumed that the rank r was known. One popular method to determine the rank r is to infer the posterior $p(r|\boldsymbol{Y})$ [23], [30]. This method requires trying out all possible values of the rank, i.e. from order 1 to $n-1$ for each patch group matrix, resulting in a huge computational burden in patch-based image processing. Here, we resort to the automatic relevance determination principle in Bayesian sparse learning to determine the rank r [31]. We initialize a relatively large value for r, e.g. $r = n-1$. During iterations, most of the values of $\boldsymbol{k}_{\boldsymbol{U}}$ and $\boldsymbol{k}_{\boldsymbol{V}}$ are driven to very small values, which forces the posterior means of most rows of $\boldsymbol{U}$ and $\boldsymbol{V}$ as well as most SVs to approach zero. The rank is therefore effectively reduced by removing those items from the model. The inferential framework of BSVT is outlined in Algorithm 1.

Algorithm 1 Variational Bayesian Singular Value Thresholding

Input: Noisy patch matrix $\boldsymbol{Y}$ of size $n \times m$

1: Perform SVD on $\boldsymbol{Y}$: $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, $\boldsymbol{D} = diag(\boldsymbol{\lambda}_Y)$

2: Initialize: $\boldsymbol{k}_{\boldsymbol{U}}^{(0)} = \boldsymbol{k}_{\boldsymbol{V}}^{(0)} = \boldsymbol{1}_{r,1}$, which is a $r \times 1$ matrix with all entries equal to 1, $\widehat{\boldsymbol{\lambda}_r}^{(0)} = \boldsymbol{\lambda}_{\boldsymbol{Y};r}$, $\widehat{w}^{(0)} = mn/\sum_{i=r+1}^{n}\lambda_{\boldsymbol{Y},i}^2$, $t = 1$.

3: **repeat** evaluate the following equations:

$$\boldsymbol{k}_{\boldsymbol{U}}^{(t)} = G(n, \widehat{w}^{(t-1)}\boldsymbol{\lambda}_{\boldsymbol{Y};r} \circ \boldsymbol{k}_{\boldsymbol{V}}^{(t-1)} \circ \widehat{\boldsymbol{\lambda}_r}^{(t-1)}), \tag{33}$$

$$\boldsymbol{k}_{\boldsymbol{V}}^{(t)} = G(m, \widehat{w}^{(t-1)}\boldsymbol{\lambda}_{\boldsymbol{Y};r} \circ \boldsymbol{k}_{\boldsymbol{U}}^{(t-1)} \circ \widehat{\boldsymbol{\lambda}_r}^{(t-1)}), \tag{34}$$

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{k}_{\boldsymbol{V}}^{(t-1)} \circ \boldsymbol{\lambda}_{\boldsymbol{Y};r} \circ \boldsymbol{k}_{\boldsymbol{U}}^{(t-1)}, \tag{35}$$

$$\sigma^{(t)} = (\widehat{w}^{(t-1)})^{-1/2}, \tag{36}$$

$$\widehat{\boldsymbol{\lambda}_r}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \sigma^{(t-1)}\zeta(\boldsymbol{\mu}^{(t-1)}, \sigma^{(t-1)}), \tag{37}$$

$$\begin{aligned}\widehat{\boldsymbol{\lambda}_r^\top\boldsymbol{\lambda}_r}^{(t)} =& (\boldsymbol{\mu}^{(t-1)})^\top\widehat{\boldsymbol{\lambda}_r}^{(t-1)} + r(\sigma^{(t-1)})^2 \\ &- \sigma^{(t-1)}\rho(\boldsymbol{\mu}^{(t-1)}, \sigma^{(t-1)})^\top\boldsymbol{1}_{r,1},\end{aligned} \tag{38}$$

$$\begin{aligned}\widehat{w}^{(t)} =& mn[\boldsymbol{\lambda}_{\boldsymbol{Y}}^\top\boldsymbol{\lambda}_{\boldsymbol{Y}} + \widehat{\boldsymbol{l}_r^\top\boldsymbol{l}_r}^{(t-1)} \\ &- 2(\boldsymbol{k}_{\boldsymbol{V}}^{(t-1)} \circ \widehat{\boldsymbol{\lambda}_r}^{(t-1)} \circ \boldsymbol{k}_{\boldsymbol{U}}^{(t-1)})^\top\boldsymbol{\lambda}_{\boldsymbol{Y};r}]^{-1}.\end{aligned} \tag{39}$$

4: Set $t = t + 1$

5: **until** convergence is reached with reduced r

6: Set $\boldsymbol{U}_r = \boldsymbol{U}_{;r}\boldsymbol{k}_{\boldsymbol{U};r}$, $\boldsymbol{D}_r = \boldsymbol{D}_{;r}$, $\boldsymbol{V}_r = \boldsymbol{V}_{;r}\boldsymbol{k}_{\boldsymbol{V};r}$

Output: $\boldsymbol{U}_r$, $\boldsymbol{D}_r$, $\boldsymbol{V}_r$

B. SURE-Based Singular Value Thresholding

As noted above, if the image is reconstructed using the low-rank approximation directly, it tends to produce a very weak noise-like pattern in flat areas and around edges, particularly in the case of moderate or high noise levels [18]. SURE is an unbiased statistical estimate of the MSE between an original unknown data source and a processed version of its noisy observation. This estimate depends only on the observed data and does not require any prior assumption on the noise-free source. Various studies have demonstrated that SURE is particularly powerful for tuning the regularization parameters for high-quality edge-preserving image filtering [28], [29]. In order to suppress artifacts in smooth areas and around edges, we thus employ SURE to refine the singular values $\boldsymbol{D}_r$ ($r < \min(m,n)$) with respect to minimizing the estimation risk or the MSE between the actual data $\boldsymbol{X}$ and the approximation $\hat{\boldsymbol{X}}$. This can be performed by selecting a parameter τ to shrink the singular values $\boldsymbol{D}_r$:

$$\begin{aligned}\mathrm{MSE}(\tau) &= E\left\|\boldsymbol{X} - \mathrm{SVT}_\tau(\hat{\boldsymbol{X}})\right\|_F^2 \\ &= E\left\|\boldsymbol{X} - \mathrm{SVT}_\tau(\boldsymbol{U}_r\boldsymbol{D}_r\boldsymbol{V}_r^\top)\right\|_F^2,\end{aligned} \tag{40}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Similar to the case for full-rank SVD, the expectation in Eq. (40) depends on the true $\boldsymbol{X}$ which is not available. Determination of τ based on minimizing MSE thus cannot be achieved directly. However, it is feasible to construct an unbiased estimate of the MSE, namely, Stein's Unbiased Risk Estimator. Assuming $m > n > r$, we can derive the unbiased risk estimator for the rank reduced eigen-triplets:

$$\begin{aligned}\mathrm{SURE}_s(\mathrm{SVT}_\tau(\boldsymbol{U}_r\boldsymbol{D}_r\boldsymbol{V}_r^\top)) =& -mn\sigma_s^2 + \sum_{i=1}^{\min(m,r)}\min(\tau^2, \lambda_i^2) \\ &+ 2\sigma_s^2\mathrm{div}_s(\mathrm{SVT}_\tau(\boldsymbol{U}_r\boldsymbol{D}_r\boldsymbol{V}_r^\top)).\end{aligned} \tag{41}$$

In comparison with full-rank SURE [7], [50], please note that σ_s^2 here is the residual noise variance of the rank-reduced $\boldsymbol{X}$ and $\min(m,n)$ degrades to $\min(m,r)$. Considering the soft threshold function of $f(\lambda_i)$ in Eq. (1), the divergence for rank-reduced eigen-triplets is modified to:

$$\begin{aligned}\mathrm{div}_s(\mathrm{SVT}_\tau(\boldsymbol{U}_r\boldsymbol{D}_r\boldsymbol{V}_r)) =& 2\sum_{i\neq j,i,j=1}^{\min(m,r)}\frac{\lambda_i(\lambda_i-\tau)_+}{\lambda_i^2-\lambda_j^2} \\ &+ (|m-r|)\sum_{i=1}^{\min(m,r)}(1-\frac{\tau}{\lambda_i})_+ \\ &+ \sum_{i=1}^{\min(m,r)}\mathbb{I}_{\lambda_i>\tau},\end{aligned} \tag{42}$$

where $\mathbb{I}$ denotes the indicator function.

Algorithm 2 SURE-Based Singular Value Thresholding

Input: Rank-reduced eigen-triplets $\boldsymbol{U}_r$, $\boldsymbol{D}_r$, $\boldsymbol{V}_r$, and interval $[\tau_{\min}, \tau_{\max}]$
1: **for** τ from $\tau_{\min}$ to $\tau_{\max}$ **do**
2: Compute $\text{div}_s(\text{SVT}_\tau(\boldsymbol{U}_r\boldsymbol{D}_r\boldsymbol{V}_r^\top))$ using Eq. (42)
3: Compute $\text{SURE}_s(\text{SVT}_\tau)(\boldsymbol{U}_r\boldsymbol{D}_r\boldsymbol{V}_r^\top)$ using Eq. (41)
4: **end for**
5: Find the τ_0 with minimal $SURE_s$
Output: $\hat{\boldsymbol{X}} = \sum_{i=1}^{r}(\lambda_i - \tau_0)_+ \boldsymbol{u}_i \boldsymbol{v}_i^\top$

We assume that $\hat{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{E}_s$, and σ_s can be estimated from the difference between the noisy observation $\boldsymbol{Y}$ and the estimation $\hat{\boldsymbol{X}}$

$$\Delta = \boldsymbol{Y} - \hat{\boldsymbol{X}} = (\boldsymbol{X} + \boldsymbol{E}) - (\boldsymbol{X} + \boldsymbol{E}_s) = \boldsymbol{E} - \boldsymbol{E}_s. \quad (43)$$

The expectation of Eq. (43) is denoted as

$$\begin{aligned}\langle \Delta^2 \rangle &= \langle \boldsymbol{E}^2 \rangle + \langle \boldsymbol{E}_s^2 \rangle - 2\langle \boldsymbol{E} \cdot \boldsymbol{E}_s \rangle \\ &= \sigma^2 + \sigma_s^2 - 2\langle \boldsymbol{E} \cdot \boldsymbol{E}_s \rangle, \end{aligned} \quad (44)$$

where $\langle \cdot \rangle$ is the expectation operator.

Since $\boldsymbol{E}_s$ can be viewed as the smoothed version of noise $\boldsymbol{E}$, we have $\langle \boldsymbol{E} \cdot \boldsymbol{E}_s \rangle = \langle (\boldsymbol{E}_s + \Delta) \cdot \boldsymbol{E}_s \rangle = \langle \Delta \cdot \boldsymbol{E}_s \rangle + \langle \boldsymbol{E}_s^2 \rangle$. It is well known that the high-frequency component Δ is much smaller than $\boldsymbol{E}_s$, which results in $\langle \boldsymbol{E} \cdot \boldsymbol{E}_s \rangle \approx \langle \boldsymbol{E}_s^2 \rangle = \sigma_s^2$. Therefore, Eq. (44) can be written as

$$\langle \Delta^2 \rangle = \sigma^2 + \sigma_s^2 - 2\sigma_s^2 = \sigma^2 - \sigma_s^2, \quad (45)$$

where σ^2 is the noise variance in the observation $\boldsymbol{Y}$ which has been estimated using BSVT in the first step. Eq. (45) is thus equivalent to

$$\sigma_s^2 = \sigma^2 - \langle \Delta^2 \rangle = \sigma^2 - \frac{1}{mn} \left\| \boldsymbol{Y} - \hat{\boldsymbol{X}} \right\|_F^2. \quad (46)$$

Considering that $\boldsymbol{E}_s$ contains not only the noise residual but also the estimation error of the noiseless image, a scaling factor γ controlling the depth of filtering is required. That is

$$\sigma_s = \gamma \sqrt{(\sigma^2 - \frac{1}{mn} \left\| \boldsymbol{Y} - \hat{\boldsymbol{X}} \right\|_F^2)}. \quad (47)$$

It is recommended to set γ around 0.55 to 0.65 to produce satisfactory results for natural image denoising [51], [52].

The outline of the SSVT method is presented in Algorithm 2.

C. The Hybrid BSSVT Algorithm

For the complete image BSSVT denoising method, we first cluster patches with similar spatial structure to form a patch matrix. BSVT and SSVT are then applied in succession on each patch matrix. The denoised patches are aggregated to reconstruct the whole noise-free image. In practice, iterative regularization is often adopted by mapping the filtered noise back to the denoised image, which has been demonstrated to be effective in improving the denoising performance [9]. This scheme is implemented as

$$\boldsymbol{Y}^{(k+1)} = \hat{\boldsymbol{X}}^{(k)} + \delta(\boldsymbol{Y} - \hat{\boldsymbol{X}}^{(k)}), \quad (48)$$

Algorithm 3 Image Denoising by BSSVT

Input: Noisy image $\boldsymbol{y}$
1: Initialize: $\hat{\boldsymbol{x}}^{(0)} = \boldsymbol{y}$, $\boldsymbol{y}^{(0)} = \boldsymbol{y}$;
2: **for** $k = 1 : K$ **do**
3: Iterative regularization using Eq. (48)
4: **for** each patch $\boldsymbol{y}_i$ in $\boldsymbol{y}^{(k)}$ **do**
5: Cluster similar patch to matrix $\boldsymbol{Y}_i$
6: Apply **Algorithm 1** on $\boldsymbol{Y}_i$ to obtain $\boldsymbol{U}_r$, $\boldsymbol{D}_r$ and $\boldsymbol{V}_r$
7: Apply **Algorithm 2** on $\boldsymbol{U}_r$, $\boldsymbol{D}_r$ and $\boldsymbol{V}_r$ to estimate $\hat{\boldsymbol{X}}_i$
8: **end for**
9: Aggregate $\hat{\boldsymbol{X}}_i$ to form the denoised image $\hat{\boldsymbol{x}}^{(k)}$
10: **end for**
Output: clean image $\hat{\boldsymbol{x}}^{(K)}$

where k denotes algorithm iteration and $0 < \delta < 1$ is a relaxation parameter. As reviewed in the Introduction, most existing approaches require an extra step to update the estimation of the noise variance due to the feedback of filtered noise, where the original noise variance propagates in each iteration. BSVT performs the low-rank approximation and infers the noise variance from $\boldsymbol{Y}^{(k+1)}$ itself without needing any prior knowledge of the original observation $\boldsymbol{Y}$ as well as the estimators from $\boldsymbol{Y}^{(1)}$ to $\boldsymbol{Y}^{(k)}$. The complete procedure for the image BSSVT denoising algorithm is summarized in Algorithm 3.

Algorithm 1 proceeds by iteratively estimating one variable while holding the others fixed. By the properties of the variational Bayesian method, the algorithm is guaranteed to converge to a local minimum of the variational bound [53]. Employing Algorithm 2 with a low threshold τ fails to remove noise, while a high threshold removes noise but also induces both spatial blurring and contrast loss. Due to the convex behavior of SURE/MSE, the searching scheme in Algorithm 2 can guarantee to obtain the optimal SURE threshold [54]. Therefore the BSSVT algorithm converges to a local minimum after a number of successive approximation iterations, resulting in an ideal balance offering strong noise reduction while maintaining important image features.

IV. Experiments

A. Parameter Settings and Performance Evaluation

The performance of BSSVT is evaluated on twelve benchmark grayscale images, shown in Fig. 2. The sizes of the first 10 images are 256×256 with the size of Baboon and Barbara being 512×512. Noisy images are produced by adding zero mean white Gaussian noise with standard deviation $\sigma = 20,\ 50,\ 70$ and 100. We adopted the setting of patch size and the number of similar patches recommended in previous studies [9], [17]: the former is set to 6×6, 7×7, 8×8 and 9×9, and the latter is set to 70, 90, 120 and 140 for $\sigma = 20$, 50, 70 and 100 respectively. Throughout this study, the scaling factor γ is fixed as 0.55.

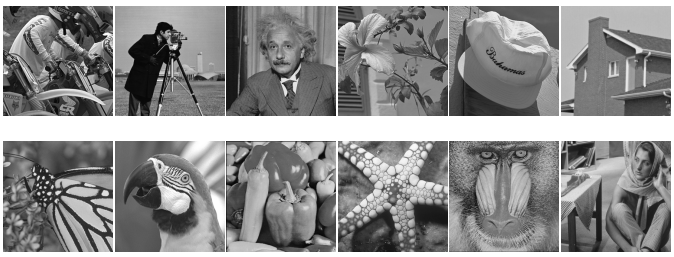


Fig. 2. The 12 test images used in image denoising experiments.

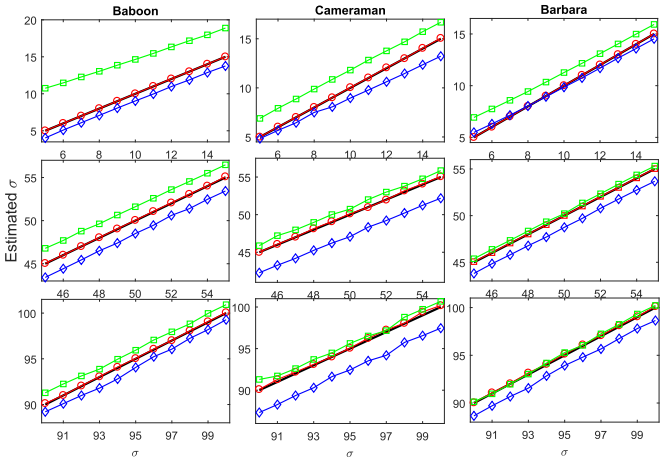


Fig. 3. Columns from left to right depict the comparison of the noise estimation results for the Baboon, Cameraman and Barbara images, respectively. Rows from top to bottom describe the comparison of noise estimation results for low ($5 \leq \sigma \leq 15$), moderate ($45 \leq \sigma \leq 55$) and severe ($90 \leq \sigma \leq 100$) levels of noise, respectively. The results of BSVT, MAD and SVK are represented by the circles, squares and diamonds, respectively. The truth is illustrated by the solid black line.

We evaluate the performance of BSSVT in terms of PSNR and SSIM [40]. Given a ground truth grayscale image $\boldsymbol{X}$, the PSNR of the recovered image $\hat{\boldsymbol{X}}$ is estimated by:

$$PSNR(\boldsymbol{X}, \hat{\boldsymbol{X}}) = 10 \cdot \log_{10}\left(\frac{255^2}{\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}\right\|_2^2}\right). \tag{49}$$

Assuming an image patch $\boldsymbol{G}$ from $\boldsymbol{X}$ as well as the patch $\boldsymbol{H}$ from the corresponding recovery $\hat{\boldsymbol{X}}$, the SSIM index between $\boldsymbol{G}$ and $\boldsymbol{H}$ is defined by:

$$SSIM(\boldsymbol{G}, \boldsymbol{H}) = \frac{2(\mu_{\boldsymbol{G}}\mu_{\boldsymbol{H}} + C_1)(2v_{\boldsymbol{GH}} + C_2)}{(\mu_{\boldsymbol{G}}^2 + \mu_{\boldsymbol{H}}^2 + C_1)(v_{\boldsymbol{G}}^2 + v_{\boldsymbol{H}}^2 + C_2)}, \tag{50}$$

where $\mu_{\boldsymbol{G}}$ and $v_{\boldsymbol{H}}$ are the average intensity and standard deviation of $\boldsymbol{G}$ and $\boldsymbol{H}$, respectively. $v_{\boldsymbol{GH}}$ denotes the cross correlation between $\boldsymbol{G}$ and $\boldsymbol{H}$, and the small constants C_1 and C_2 are used to avoid numerical instability. The SSIM of the entire image is estimated by averaging the local SSIM indices using a sliding window [55].

B. Effect on Noise Variance Estimation

We first demonstrate the effectiveness of our variational Bayesian model to estimate the noise variance in the BSVT step. We choose three patch group matrices, i.e. one with structure from Baboon, one with texture from Cameraman, and one with both structure and texture from Barbara. Fig. 3 shows the average noise variance of 20 noisy samples for each of these three representative patch group matrices. Two other popular methods based on the wavelet MAD and the scale variance of kurtosis (SVK) are also plotted for comparison [38], [43].

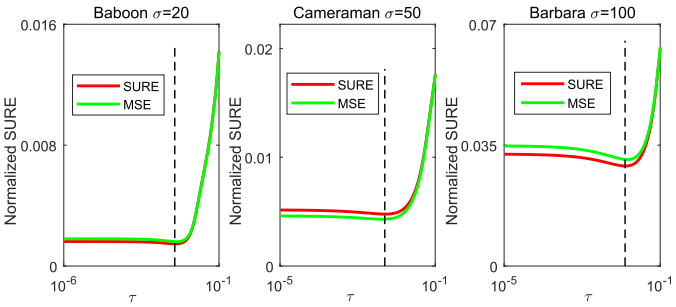


Fig. 4. SURE and MSE as a function of threshold value for Baboon ($\sigma = 20$), Cameraman ($\sigma = 50$) and Barbara ($\sigma = 100$). Columns from left to right correspond to noise level $\sigma = 20,\ 50$ and 100.

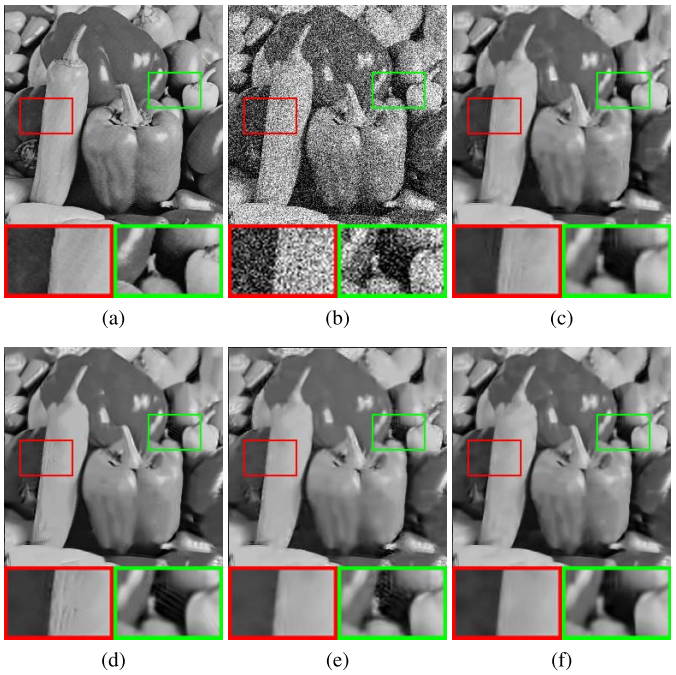


Fig. 5. Comparison of denoising results on the Peppers image contaminated by Gaussian white noise with $\sigma = 50$. (a) Original image, (b) noisy image (PSNR = 14.12 dB), (c) BM3D (PSNR = 26.16 dB), (d) WNNM (PSNR = 26.23 dB), (e) RMMM (PSNR = 25.87 dB), and (f) BSSVT (PSNR = **26.40** dB).

It is apparent that the variational Bayesian model accurately tracks the actual noise variance in the cases of low, moderate and severe noise contamination for each image. The difference between the true noise variance (in black) and that estimated by BSVT (in red) is almost unrecognizable in most cases. MAD has been broadly applied to assess different kinds of image denoising algorithms. However, it can be rather problematic when MAD is applied to images containing a considerable component at the HH_1 level in the wavelet domain [56]. Therefore, using the noisy version of these coefficients at this level to estimate the noise variance can

TABLE I
DENOISING RESULTS (PSNR) BY COMPETING METHODS ON THE 12 TEST IMAGES. THE BEST RESULTS ARE IN BOLD

σ	20				50			
schemes	BM3D	WNNM	RNNM	BSSVT	BM3D	WNNM	RNNM	BSSVT
Bike	28.24	28.70	27.99	**28.74**	22.42	22.50	22.47	**22.83**
Cameraman	30.36	**30.68**	30.08	30.43	24.99	25.16	24.93	**25.49**
Einstein	31.29	31.47	30.97	**31.48**	27.11	27.19	26.65	**27.35**
Flower	29.99	30.42	29.73	**30.37**	25.12	25.33	24.87	**25.65**
Hat	31.55	**32.05**	31.35	31.86	27.14	27.23	26.55	**27.59**
House	33.88	**34.14**	33.64	34.12	29.39	**29.87**	28.65	29.41
Monarch	30.52	31.34	29.25	**31.42**	25.46	25.56	25.37	**25.97**
Parrot	29.88	**30.03**	29.68	29.66	24.76	24.69	24.72	**25.08**
Peppers	31.28	31.59	31.07	**31.62**	26.16	26.23	25.87	**26.40**
Starfish	29.45	30.20	29.59	**30.23**	24.29	24.41	24.32	**24.67**
Baboon	25.58	**25.67**	25.49	25.59	21.83	22.15	22.13	**22.51**
Barbara	31.23	**31.68**	31.35	31.58	26.24	26.72	26.61	**26.73**
Average	30.27	**30.66**	30.02	30.59	25.41	25.59	25.26	**25.81**
σ	70				100			
schemes	BM3D	WNNM	RNNM	BSSVT	BM3D	WNNM	RNNM	BSSVT
Bike	20.46	20.08	20.36	**20.87**	18.38	17.83	18.25	**18.63**
Cameraman	22.56	22.72	22.59	**23.30**	19.86	20.25	20.08	**20.90**
Einstein	25.23	24.97	24.39	**25.56**	22.63	21.79	21.99	**22.68**
Flower	23.20	23.47	22.99	**23.77**	20.59	21.60	21.15	**22.32**
Hat	25.46	25.23	24.73	**25.80**	22.90	22.59	22.34	**23.24**
House	26.98	27.15	26.63	**27.58**	23.71	23.27	22.88	**24.15**
Monarch	22.99	23.40	23.14	**23.90**	19.85	20.82	20.35	**21.31**
Parrot	22.15	22.39	22.29	**22.87**	19.17	19.70	19.61	**20.45**
Peppers	23.97	23.63	23.55	**24.21**	21.52	20.82	21.12	**21.63**
Starfish	22.35	21.83	22.19	**22.53**	20.00	19.05	20.01	**20.41**
Baboon	20.58	20.87	20.32	**21.09**	19.17	19.39	19.46	**20.22**
Barbara	24.56	24.89	24.74	**25.25**	23.34	23.18	23.06	**23.98**
Average	23.37	23.39	23.16	**23.89**	20.93	20.87	20.86	**21.66**

result in considerable errors. The error according to MAD in our simulation is the largest across the three images and three noise level intervals. This result is consistent with the findings in [39]. Similar to MAD, the performance of SVK varies significantly across the images and noise levels. Although it is better than MAD, it is much worse than the Bayesian model, particularly for the Cameraman image. Recall that the major purpose of the first step in BSSVT is to remove noise through the Bayesian low-rank approximation. The precise noise variance obtained in this step is a by-product of this procedure, although it is required in the second step of BSSVT as well as in other denoising methods.

C. Effect of SURE on Eigen-Triplets Thresholding

The second step of BSSVT employs SURE to optimally tune the rank-reduced eigen-triplets in terms of minimizing the estimation risk or MSE. Here, we evaluate the effectiveness of SURE for rank-reduced eigen-triplets thresholding. Fig. 4 shows the SURE and MSE for the rank-reduced SVs as a function of the threshold τ for the representative patch group matrices used in Fig. 3 with $\sigma = 20$ for Baboon, 50 for Cameraman, and 100 for Barbara, respectively. For each case, with increasing threshold τ, the estimated risk and MSE first have a relatively high plateau, and then descend to reach the minimum. They increase dramatically from this point with increasing τ. Due to the error of the estimated variance, there is a minor offset between SURE and the actual MSE. However, it was found that the sensitivity of SURE to the estimated values of variance is small, and the locations of the minima of the MSE and SURE are almost the same. These findings are consistent with the results in [54], [57] of SURE for the full rank matrix with estimated variance. These plots thus indicate that SURE can converge to a minimum and approximate the true patch with minimal estimation risk.

D. Numerical Results

There have been a large number of nonlocal algorithms developed in the past decade. BM3D [17] is the benchmark algorithm in image nonlocal denoising. WNNM [9] is always ranked as one of the most competitive methods in comparative studies while RNNM [16] shares similar principles to BSSVT in aiming to balance between soft and hard thresholding. We thus compare the performance of BSSVT with BM3D, WNNM and RNNM. BSSVT and RNNM are implemented

TABLE II
DENOISING RESULTS (SSIM) BY COMPETING METHODS ON THE 12 TEST IMAGES. THE BEST RESULTS ARE IN BOLD

σ	20				50			
schemes	BM3D	WNNM	RNNM	BSSVT	BM3D	WNNM	RNNM	BSSVT
Bike	0.887	0.893	0.895	**0.896**	0.688	0.687	0.705	**0.711**
Cameraman	0.872	**0.877**	0.854	0.875	0.747	0.755	0.685	**0.760**
Einstein	0.801	**0.807**	0.806	0.802	0.696	0.699	0.648	**0.701**
Flower	0.874	**0.885**	0.885	0.880	0.716	0.724	0.687	**0.732**
Hat	0.876	0.883	0.856	**0.884**	0.767	0.776	0.666	**0.780**
House	**0.869**	0.871	0.863	0.864	0.812	0.826	0.734	**0.828**
Monarch	0.923	**0.930**	0.912	0.922	0.824	0.829	0.792	**0.831**
Parrot	0.867	0.868	0.857	**0.871**	0.757	0.750	0.697	**0.758**
Peppers	0.890	**0.894**	0.878	0.891	0.786	0.788	0.735	**0.790**
Starfish	0.870	0.885	0.872	**0.887**	0.725	0.720	0.713	**0.737**
Baboon	0.722	**0.730**	0.728	0.726	0.469	0.508	0.485	**0.513**
Barbara	0.909	**0.915**	0.910	0.912	0.762	0.785	0.784	**0.785**
Average	0.863	**0.870**	0.860	0.868	0.729	0.737	0.694	**0.744**
σ	70				100			
schemes	BM3D	WNNM	RNNM	BSSVT	BM3D	WNNM	RNNM	BSSVT
Bike	0.588	0.553	0.598	**0.613**	0.468	0.399	0.475	**0.495**
Cameraman	0.677	0.679	0.583	**0.695**	0.592	0.617	0.488	**0.620**
Einstein	0.646	0.637	0.563	**0.653**	0.592	0.569	0.468	**0.591**
Flower	0.623	0.640	0.584	**0.647**	0.505	0.552	0.465	**0.558**
Hat	0.732	0.738	0.589	**0.743**	0.689	0.683	0.489	**0.689**
House	0.778	**0.795**	0.648	0.791	**0.729**	0.726	0.649	0.726
Monarch	0.758	0.766	0.699	**0.770**	0.649	0.684	0.589	**0.695**
Parrot	0.685	0.693	0.617	**0.702**	0.599	0.624	0.524	**0.632**
Peppers	**0.739**	0.730	0.652	0.735	0.673	0.657	0.562	**0.681**
Starfish	0.652	0.623	0.619	**0.669**	0.556	0.484	0.513	**0.565**
Baboon	0.440	0.467	0.460	**0.470**	0.406	0.448	0.428	**0.452**
Barbara	0.685	0.701	0.694	**0.708**	0.658	0.683	0.669	**0.685**
Average	0.667	0.669	0.609	**0.683**	0.593	0.584	0.527	**0.616**

in MATLAB, while BM3D and WNNM are tested using the executables and source codes provided by the authors. Because the exact noise variance is not available in real applications, we feed the algorithms of BM3D, WNNM and RNNM with the noise variance estimated using SVK [43]. This is fair and reasonable and represents their implementation in practice. We estimate the PSNR and SSIM over 20 realizations for each scheme with $\sigma = 20,\ 50,\ 70$ and 100 dB. The PSNR and SSIM values are displayed in Tables I and II respectively, where the best results are bolded. It is apparent that for low noise levels, the performance of BSSVT is, in general, equivalent to WNNM. This is reasonable because less iterations are required to estimate the noise variance. With the increase of the noise level, our algorithm performs increasingly better than the other algorithms. In particular, compared with WNNM, the improvement in the PSNR values is greater than 0.6 dB for all images at $\sigma = 100$. As for RNNM, BSSVT outperforms it in almost every case. This may be due to its sensitivity to the error of the noise variance and the fact that the low-rank parameter r was set empirically. In addition, BSSVT outperforms BM3D in all cases in terms of PSNR. The SSIM result of BSSVT is also highly competitive against the other methods.

In Figs. 5 and 6, we compare the visual quality of the denoising results on the four methods. In Fig. 5, we compare the Peppers picture under a noise level of $\sigma = 50$. BSSVT restores the edges with fewer artifacts. However, BM3D and RNNM suffer from artifacts in smooth areas and around edges. In Fig. 6, we compare the Monarch image under a noise level of $\sigma = 100$, where BSSVT achieves a visually satisfactory result with the least artifacts. In such extreme noise contamination, it is evident that the other three methods are less able to preserve the edge structures and smooth features of the image. Overall, both quantitative assessment and visual inspection demonstrate that BSSVT yields better performance in comparison to the state-of-the-art methods.

The combination of BSVT with SSVT leads to superior performance compared with the state-of-the-art methods. A natural question to ask is whether such a combination can extend to BSVT together with other methods to take advantage of the estimated noise variance. We further tested the performance of BSVT-BM3D and BSVT-WNNM. It was found that both BSVT-BM3D and BSVT-WNNM generate over-smoothed images with performance scores lower than BSSVT. Fig. 7 shows a typical example of denoised Monarch images using BSVT followed by BM3D and WNNM for the image contaminated by noise with $\sigma = 50$ and $\sigma = 100$, respectively. Many previous studies have indicated that BM3D and WNNM, as well as some other low-rank approximation-based methods, tend to over-smooth images [8], [18], [24]. The consecutive use of BSVT followed by BM3D or WNNM

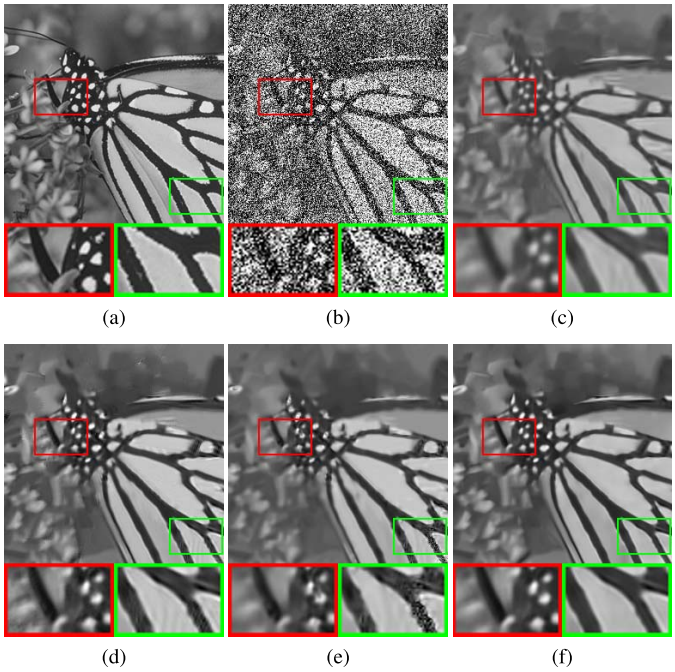


Fig. 6. Comparison of denoising results on the Monarch image contaminated by the Gaussian white noise with $\sigma = 100$. (a) Original image, (b) noisy image (PSNR = 8.10 dB), (c) BM3D (PSNR = 19.85 dB), (d) WNNM (PSNR = 20.82 dB), (e) RMMM (PSNR = 20.35 dB), and (f) BSSVT (PSNR = **21.31** dB).

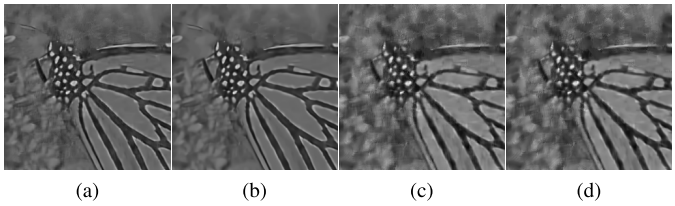


Fig. 7. The effect of BSVT-BM3D and BSVT-WNNM to denoise image Monarch contaminated by the Gaussian white noise with $\sigma = 50$ (a, b) and $\sigma = 100$ (c, d). (a) BSVT-BM3D (PSNR = 20.90 dB), (b) BSVT-WNNM (PSNR = 20.97 dB), (c) BSVT-BM3D (PSNR = 17.81 dB), (d) BSVT-WNNM (PSNR = 17.85 dB).

can remove the noise artifacts. However, this also smears out details, which results in over-smoothed images with relatively low PSNR. In the second step of the proposed method, SSVT, complementary to BSVT, directly maximizes the PSNR by refining the optimal threshold that minimizes the MSE estimation of rank-reduced eigen-triplets, avoiding over-smoothing the image.

In terms of computational efficiency, we use a desktop with a recent 2.2 GHz CPU to execute the code in Matlab 2017b (Mathworks, Massachusetts, US). BSSVT requires around 20 minutes to denoise an image for varying noise levels, while the times for BM3D, WNNM, and RNNM vary from one minute to around ten minutes. Although BSSVT is relatively slow in its current form, the computational efficiency can be significantly improved via parallel computing techniques and optimization of the search interval.

V. Discussion and Conclusion

In this study, we have proposed a hybrid nonlocal variational Bayesian image denoising framework. The proposed BSSVT approach is closely related to nuclear norm minimization. It can be interpreted as performing a weighted nuclear norm factorization or low-rank approximation using variational Bayesian inference. The noise variance is a crucial factor that impacts on the denoising quality. Most existing nonlocal image denoising methods either resort to an extra step to pre-determine the noise variance or simply assume the true value is known. However, the error in the noise variance accumulates in any iterative regularization scheme which can further worsen the denoising quality. In contrast to these existing methods, BSSVT simultaneously removes noise and infers the latent parameters including the noise variance and the rank. It adaptively adjusts the noise variance without incurring error propagation between iterations. This is the primary reason that the proposed method outperforms these competitive algorithms in this study. BSSVT further refines the rank-reduced SVs based on the SURE criterion in the second step to improve edge-preservation and artifact removal. SURE has an explicit mathematical mechanism to approximate the true image by minimizing the risk or MSE, therefore maximizing PSNR. This provides another indispensable element of BSSVT to enhance the denoised image quality. Since BSVT, the first step of BSSVT, can accurately approximate the noise variance, it can also be separately applied to improve other image denoising of segmentation methods.

In this work, only Gaussian noise was considered in the model. Both Bayesian inference and the SURE criterion are able to handle non-Gaussian noise [57]–[60]. BSSVT can naturally extend to the models to remove Poisson, Gamma, Rician as well as hybrid noise in the image for real-world applications. We will also investigate the combination of the current model with deep neural network to form a hybrid deep Bayesian learning scheme for improved image denoising, deblurring, and completion.

Appendix

A. Von Mises-Fisher Distribution

For a matrix random variable $\boldsymbol{D} \in \mathbb{R}^{p\times q}$ with restriction $p \geq q$ and $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{I}_q$, the von Mises-Fisher distribution of $\boldsymbol{D}$ is given by

$$f(\boldsymbol{D}|\boldsymbol{F}) = \text{vMF}(\boldsymbol{F}) = \frac{1}{\kappa(p, \boldsymbol{F}'\boldsymbol{F})} exp(tr(\boldsymbol{F}'\boldsymbol{D})), \quad (51)$$

$$\kappa(p, \boldsymbol{F}\boldsymbol{F}') = {}_0F_1(\frac{1}{2}p, \frac{1}{4}\boldsymbol{F}'\boldsymbol{F})C(p, q), \quad (52)$$

where $\boldsymbol{F} \in \mathbb{R}^{p\times q}$ is a matrix parameter of the same dimensions as $\boldsymbol{D}$ and $\kappa(p, \boldsymbol{F}'\boldsymbol{F})$ is the normalizing constant. ${}_0F_1(\cdot)$ denotes a hypergeometric function of matrix argument $\boldsymbol{F}'\boldsymbol{F}$. $C(p, q)$ denotes the area of the relevant Stiefel manifold $\mathfrak{F}$.

B. Truncated Normal Distribution

The probability density function of the truncated normal distribution $f(x)$ for $x \in (a, b)$ is given by

$$f(x|\mu, \sigma, a, b) = \frac{\sqrt{2}exp(-(1/2)((x-\mu)/\sigma)^2)}{\sigma\sqrt{\pi}(\text{erf}(\beta) - \text{erf}(\alpha))} \chi((a, b]), \quad (53)$$

where $\alpha = (a-\mu)/\sigma\sqrt{2}$, $\beta = (b-\mu)/\sigma\sqrt{2}$. The first two moments of Eq. (53) are $\widehat{x} = \mu - s\zeta(\mu, s)$ and $\widehat{x^2} = s^2 + \mu\widehat{x} - s\rho(\mu, s)$, which depend on the auxiliary functions

$$\zeta(\mu,\sigma) = \frac{\sqrt{2}[\exp(-\beta^2) - \exp(-\alpha^2)]}{\sqrt{\pi}(\mathrm{erf}(\beta) - \mathrm{erf}(\alpha))}, \tag{54}$$

$$\rho(\mu,\sigma) = \frac{\sqrt{2}[b\exp(-\beta^2) - a\exp(-\alpha^2)]}{\sqrt{\pi}(erf(\beta) - erf(\alpha))}. \tag{55}$$

Here $\mathrm{erf}(x)$ denotes the error function.

C. Gamma Distribution

The probability density function of the gamma distribution with shape parameter α and rate parameter β is denoted as

$$f(x|a,b) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{56}$$

where $x > 0$ and $\alpha, \beta > 0$. $\Gamma(\cdot)$ is the gamma function. The first moment of Eq. (56) is $\hat{x} = \alpha/\beta$.

ACKNOWLEDGMENT

The authors would like to thank the authors of BM3D [17], WNNM [9] for providing their source codes or executables for performance comparison. The code associated with this paper will be available online.

REFERENCES

[1] A. B. Owen *et al.*, "Bi-cross-validation of the SVD and the nonnegative matrix factorization," *Ann. Appl. Statist.*, vol. 3, no. 2, pp. 564–594, 2009.

[2] J. Josse and F. Husson, "Selecting the number of components in principal component analysis using cross-validation approximations," *Comput. Statist. Data Anal.*, vol. 56, no. 6, pp. 1869–1879, 2012.

[3] M. Gavish and D. L. Donoho, "The optimal hard threshold for singular values is $4/\sqrt{3}$," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 5040–5053, Jun. 2014.

[4] Q. Guo, C. Zhang, Y. Zhang, and H. Liu, "An efficient SVD-based method for image denoising," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 5, pp. 868–880, May 2016.

[5] Y. Zhang, J. Liu, M. Li, and Z. Guo, "Joint image denoising using adaptive principal component analysis and self-similarity," *Inf. Sci.*, vol. 259, pp. 128–141, Feb. 2014.

[6] Q. Guo, S. Gao, X. Zhang, Y. Yin, and C. Zhang, "Patch-based image inpainting via two-stage low rank approximation," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 6, pp. 2023–2036, Jun. 2018.

[7] E. J. Candès, C. A. Sing-Long, and J. D. Trzasko, "Unbiased risk estimates for singular value thresholding and spectral estimators," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4643–4657, Oct. 2012.

[8] W. Dong, G. Shi, and X. Li, "Nonlocal image restoration with bilateral variance estimation: A low-rank approach," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 700–711, Feb. 2013.

[9] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2862–2869.

[10] S. F. Yeganli, H. Demirel, and R. Yu, "Noise removal from MR images via iterative regularization based on higher-order singular value decomposition," *Signal, Image Video Process.*, vol. 11, no. 8, pp. 1477–1484, 2017.

[11] Z. Huang, Q. Li, H. Fang, T. Zhang, and N. Sang, "Iterative weighted nuclear norm for X-ray cardiovascular angiogram image denoising," *Signal, Image Video Process.*, vol. 11, no. 8, pp. 1445–1452, 2017.

[12] X. M. Luo, Z. Y. Suo, Q. G. Liu, and X. F. Wang, "Efficient noise reduction for interferometric phase image via non-local non-convex low-rank regularisation," *IET Signal Process.*, vol. 10, no. 7, pp. 815–824, 2016.

[13] Y. Xie, S. Gu, Y. Liu, W. Zuo, W. Zhang, and L. Zhang, "Weighted schatten p-norm minimization for image denoising and background subtraction," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4842–4857, Oct. 2016.

[14] J. Josse and S. Sardy, "Adaptive shrinkage of singular values," *Statist. Comput.*, vol. 26, no. 3, pp. 715–724, 2016.

[15] M. Verbanck, J. Josse, and F. Husson, "Regularised PCA to denoise and visualise data," *Statist. Comput.*, vol. 25, no. 2, pp. 471–486, 2015.

[16] X. Jia, X. Feng, and W. Wang, "Rank constrained nuclear norm minimization with application to image denoising," *Signal Process.*, vol. 129, pp. 1–11, Dec. 2016.

[17] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[18] M. Nejati, S. Samavi, H. Derksen, and K. Najarian, "Denoising by low-rank and sparse representations," *J. Vis. Commun. Image Represent.*, vol. 36, no. C, pp. 28–39, Apr. 2016.

[19] W. He, H. Zhang, L. Zhang, and H. Shen, "Hyperspectral image denoising via noise-adjusted iterative low-rank matrix approximation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3050–3061, Jun. 2015.

[20] C. Zhang, W. Hu, T. Jin, and Z. Mei, "Nonlocal image denoising via adaptive tensor nuclear norm minimization," *Neural Comput. Appl.*, vol. 29, no. 1, pp. 3–19, 2015.

[21] X. Liu, X.-Y. Jing, G. Tang, F. Wu, and Q. Ge, "Image denoising using weighted nuclear norm minimization with multiple strategies," *Signal Process.*, vol. 135, pp. 239–252, Jan. 2017.

[22] Z. Wu, Q. Wang, J. Jin, and Y. Shen, "Structure tensor total variation-regularized weighted nuclear norm minimization for hyperspectral image mixed denoising," *Signal Process.*, vol. 131, pp. 202–219, Feb. 2017.

[23] P. D. Hoff, "Model averaging and dimension selection for the singular value decomposition," *J. Amer. Stat. Assoc.*, vol. 102, no. 478, pp. 674–685, 2007.

[24] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[25] M. O. Ulfarsson and V. Solo, "Dimension estimation in noisy PCA with SURE and random matrix theory," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 5804–5816, Dec. 2008.

[26] M. O. Ulfarsson and V. Solo, "Selecting the number of principal components with SURE," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 239–243, Feb. 2015.

[27] N. R. Hansen, "On Stein's unbiased risk estimate for reduced rank estimators," *Statist. Probab. Lett.*, vol. 135, pp. 76–82, Apr. 2018.

[28] S. Ramani, Z. Liu, J. Rosen, J.-F. Nielsen, and J. A. Fessler, "Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and sure-based methods," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3659–3672, Aug. 2012.

[29] T. Qiu, A. Wang, N. Yu, and A. Song, "LLSURE: Local linear sure-based edge-preserving image filtering," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 80–90, Jan. 2013.

[30] V. Šmídl and A. Quinn, "On Bayesian principal component analysis," *Comput. Statist. Data Anal.*, vol. 51, no. 9, pp. 4101–4123, 2007.

[31] R. M. Neal, *Bayesian Learning for Neural Networks*. Berlin, Germany: Springer-Verlag, 1996.

[32] N. Dobigeon and J.-Y. Tourneret, "Bayesian orthogonal component analysis for sparse representation," *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2675–2685, May 2010.

[33] Y. J. Lim and Y. W. Teh, "Variational Bayesian approach to movie rating prediction," in *Proc. KDD Cup Workshop*, vol. 7, 2007, pp. 15–21.

[34] A. Holbrook, A. Vandenberg-Rodes, and B. Shahbaba. (2016). "Bayesian inference on matrix manifolds for linear dimensionality reduction." [Online]. Available: https://arxiv.org/abs/1606.04478

[35] W. E. Zhang, M. Tan, Q. Z. Sheng, L. Yao, and Q. Shi, "Efficient orthogonal non-negative matrix factorization over Stiefel manifold," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 1743–1752.

[36] F. Pompili, N. Gillis, P.-A. Absil, and F. Glineur, "Two algorithms for orthogonal nonnegative matrix factorization with application to clustering," *Neurocomputing*, vol. 141, pp. 15–25, Oct. 2014.

[37] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 1828–1832.

[38] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[39] D.-G. Kim and Z. H. Shamsi, "Enhanced residual noise estimation of low rank approximation for image denoising," *Neurocomputing*, vol. 293, pp. 1–11, Jun. 2018.
[40] X. Huang, L. Chen, J. Tian, and X. Zhang, "Blind image noise level estimation using texture-based eigenvalue analysis," *Multimedia Tools Appl.*, vol. 75, no. 5, pp. 2713–2724, 2016.
[41] X. Liu, M. Tanaka, and M. Okutomi, "Single-image noise level estimation for blind denoising," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5226–5237, Dec. 2013.
[42] W. Liu and W. Lin, "Additive white Gaussian noise level estimation in SVD domain for images," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 872–883, Mar. 2013.
[43] D. Zoran and Y. Weiss, "Scale invariance and noise in natural images," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2209–2216.
[44] H. Jeffreys, *The Theory of Probability*. Oxford, U.K.: Oxford Univ. Press, 1998.
[45] J. Winn and C. M. Bishop, "Variational message passing," *J. Mach. Learn. Res.*, vol. 6, pp. 661–694, Apr. 2005.
[46] J. W. Miskin, "Ensemble learning for independent component analysis," Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 2000.
[47] Z. Ghahramani and M. J. Beal, "Variational inference for Bayesian mixtures of factor analysers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 449–455.
[48] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[49] C. G. Khatri and K. V. Mardia, "The von Mises–Fisher matrix distribution in orientation statistics," *J. Roy. Stat. Soc. B (Methodol.)*, vol. 39, no. 1, pp. 95–106, 1977.
[50] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, no. 6, pp. 1135–1151, 1981.
[51] L. Zhang, W. Dong, D. Zhang, and G. Shi, "Two-stage image denoising by principal component analysis with local pixel grouping," *Pattern Recognit.*, vol. 43, no. 4, pp. 1531–1549, 2010.
[52] T. Dai *et al.*, "A generic denoising framework via guided principal component analysis," *J. Vis. Commun. Image Represent.*, vol. 48, pp. 340–352, Oct. 2017.
[53] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). Berlin, Germany: Springer-Verlag, 2006.
[54] D. Van De Ville and M. Kocher, "SURE-based non-local means," *IEEE Signal Process. Lett.*, vol. 16, no. 11, pp. 973–976, Nov. 2009.
[55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[56] M. Hashemi and S. Beheshti, "Adaptive noise variance estimation in BayesShrink," *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 12–15, Jan. 2010.
[57] T. Furnival, R. K. Leary, and P. A. Midgley, "Denoising time-resolved microscopy image sequences with singular value thresholding," *Ultramicroscopy*, vol. 178, pp. 112–124, Jul. 2017.
[58] P. Gopalan, F. Ruiz, R. Ranganath, and D. Blei, "Bayesian nonparametric poisson factorization for recommendation systems," *J. Mach. Learn. Res.*, vol. 33, pp. 275–283, 2014.
[59] D. M. Blei, P. R. Cook, and M. Hoffman, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 439–446.
[60] F. Luisier, T. Blu, and M. Unser, "Image denoising in mixed Poisson–Gaussian noise," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 696–708, Mar. 2011.



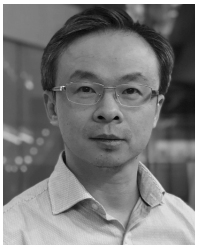
Caoyuan Li received the bachelor's degree in software engineering from Tongji University, Shanghai, China, in 2013. He is currently pursuing the the dual Ph.D. degree with the Beijing Institute of Technology and the University of Technology Sydney. His main interests include machine learning, image processing, and data mining.



Hong-Bo Xie (M'12) is currently a Vice-Chancellor's Senior Research Fellow with the ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology, Brisbane, Australia. He has published over 70 peer-reviewed journal and conference papers. His research interests include machine learning, signal and image processing and nonlinear time series analysis and their applications in biomedical engineering.



Xuhui Fan received the bachelor's degree in mathematical statistics from the University of Science and Technology of China, Hefei, China, in 2010, and the Ph.D. degree in computer science from the University of Technology Sydney, Chippendale, NSW, Australia, in 2015. His current research interests include stochastic random partition and the Bayesian nonparametrics.



Richard Yi Da Xu is currently an Associate Professor with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia. He has authored about 50 papers, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Pattern Recognition*, *ACM Transactions on Knowledge Discovery from Data*, the Association for the Advancement of Artificial Intelligence, and the International Conference on Image Processing. His current research interests include machine learning, computer vision, and statistical data mining.



Sabine Van Huffel (F'09) was a Guest Professor with Stanford University, CA, USA, and Uppsala University, Sweden, and a Visiting Fellow and Scientist with the University of Minnesota, MN, USA. She has been a Full Professor with the Department of Electrical Engineering, KU Leuven, since 2002. She has the following publications record: two monographs, about 304 articles in peer reviewed international journals, about 257 conference papers, four edited books, seven edited journal special issues, and 21 book chapters. Her research topics are fundamental/theoretical as well as application oriented and are performed in the domain of (multi)linear algebra, (non)linear signal analysis, classification, and system identification with a special focus on the development of numerically reliable and robust algorithms for improving medical diagnostics. She was elected as a fellow of the Royal Flemish Academy of Belgium for Sciences and the Arts in 2017.



Scott A. Sisson is currently a Professor and the Head of Statistics of the School of Mathematics and Statistics, University of New South Wales. He is also a past President of the Statistical Society of Australia. He is also the Chair of the Australian Chapter of the International Society of Bayesian Analysis (ISBA). His current research interests include the Bayesian statistics and computation, big data, extreme value theory, and applied science. He was an Australian Research Council Queen Elizabeth II Research Fellow. He is a member of the Australian National Committee for Mathematical Sciences. He is internationally recognized for his work in computational and Bayesian statistics, in particular for developing inferential techniques for computationally intractable models and challenging data. He received the Moran Medal from the Australian Academy of Science, the G. N. Alexander Medal from the Engineers Australia, and the J. G. Russell Award from the Australian Academy of Science.



Kerrie Mengersen is currently a Full Professor with the School of Mathematical Sciences, Queensland University of Technology, Australia. She has over 300 refereed journal publications, over 40 keynote and invited international conference presentations, and attraction of over 30 large research grants. She was elected as a fellow of the Australian Academy of Science in 2018. In 2016, QUT awarded her with the title of Distinguished Professor, in recognition of her outstanding achievements, both nationally and internationally, in mathematics and statistical research. She was the National President of the Statistical Society of Australia from 2011 to 2012 and the President of the International Society for Bayesian Analysis from 2015 to 2017.