**Abstract:**
Sparse coding has been shown to be relevant to both variational and Bayesian approaches. The regularization parameter in variational image restoration is intrinsically connected with the shape parameters of sparse coefficients’ distributions in Bayesian methods. How to tune those parameters in a principled yet spatially adaptive fashion turns out a challenging problem especially for the class of nonlocal image models. In this work, we propose a Bayesian structured sparse coding (BSSC) framework to address this issue - more specifically, a nonlocal extension of Gaussian scale mixture (GSM) model is developed based on structured sparsity and its related maximum a posterior (MAP) estimation problem is tackled. It is shown that the variances of sparse coefficients (the field of scalar multipliers of Gaussian) - if treated as a latent variable - can be jointly estimated along with the unknown sparse coefficients via the method of alternating optimization. Unlike previous attacks to MAP estimation, ours leads to closed-form solutions involving iterative shrinkage/filtering only and therefore admits computationally efficient implementation. Our solution to BSSC can be viewed as an exemplar of demonstrating a new variational approach toward empirical Bayesian inference with parametric models, which has computational advantages over the existing variational Bayesian approaches. When applied to image restoration, our experimental results have shown that BSSC can both preserve the sharpness of edges and suppress undesirable artifacts. Thanks to its capability of achieving a better spatial adaptation, BSSC-based image restoration often delivers reconstructed images with higher subjective/objective qualities than other competing approaches.
Image Restoration via Bayesian Structured Sparse Coding: Where Structured Sparsity Meets Gaussian Scale Mixture

Weisheng Dong · Guangming Shi · Yi Ma · Xin Li

the date of receipt and acceptance should be inserted later

Abstract Sparse coding has been shown to be relevant to both variational and Bayesian approaches. The regularization parameter in variational image restoration is intrinsically connected with the shape parameters of sparse coefficients' distributions in Bayesian methods. How to tune those parameters in a principled yet spatially adaptive fashion turns out a challenging problem especially for the class of non-local image models. In this work, we propose a Bayesian structured sparse coding (BSSC) framework to address this issue - more specifically, a nonlocal extension of Gaussian scale mixture (GSM) model is developed based on structured sparsity and its related maximum a posterior (MAP) estimation problem is tackled. It is shown that the variances of sparse coefficients (the field of scalar multipliers of Gaussian) - if treated as a latent variable - can be jointly estimated along with the unknown sparse coefficients via the method of alternating optimization. Unlike previous attacks to MAP estimation, ours leads to closed-form solutions involving iterative shrinkage/filtering only and therefore admits computationally efficient implementation. Our solution to BSSC can be viewed as an exemplar of demonstrating a new variational approach toward empirical Bayesian inference with parametric models, which has computational advantages over the existing variational Bayesian approaches. When applied to image restoration, our experimental results have shown that BSSC can both preserve the sharpness of edges and suppress undesirable artifacts. Thanks to its capability of achieving a better spatial adaptation, BSSC-based image restoration often delivers reconstructed images with higher subjective/objective qualities than other competing approaches.

Keywords Bayesian sparse coding · Gaussian scale mixture · structured sparsity · alternative minimization · variational image restoration

1 Introduction

1.1 Background and Motivation

Sparse representation of signals/images has been widely studied by signal/image processing communities in the past decades. Historically, the idea of sparsity dated back to the coring operator invented by RCA researchers in early 1980s [1]. The birth of wavelet [2] or filter bank theory [3] or multi-resolution analysis [4] in late 1980s rapidly sparked the interest in sparse representation, which has found successful applications into image coding [5], [6], [7] and denoising [8], [9], [10]. Under the framework of sparse coding, a lot of research have been centered at two related issues: basis functions (or dictionary) and statistical modeling of sparse coefficients. Exemplar studies of the former are the construction of directional multiresolution representation (e.g., contourlet [11]) and over-complete dictionary learning from training data (e.g., K-SVD [12,13], multiscale dictionary learning [16], online dictionary learning [14] and the non-parametric Bayesian dictionary learning [15]); the latter in-
include the use of Gaussian mixture models [26,32], the variational Bayesian models [17,29], the universal models [24], and the centralized Laplacian model [34] for sparse coefficients.

More recently, a class of nonlocal image restoration techniques [19], [20], [21], [41], [22], [23], [34] have attracted increasingly more attention. The key motivation behind lies in the observation that many important image structures in natural images including edges and textures can be characterized by the abundance of self-repeating patterns. Such observation has led to the formulation of nonlocal simultaneous sparse coding [21]. Our own recent works [22], [23] has more clearly verified the potential of exploiting structured sparsity in image restoration. However, a fundamental question remains open: how to achieve (local) spatial adaptation within the framework of nonlocal image restoration? This question is related to the issue of regularization parameters in a variational setting or shape parameters in a Bayesian one; but the issue becomes even more thorny when it is tangled with nonlocal regularization/sparsity. To the best of our knowledge, how to tune those parameters in a principled manner remains an open problem (e.g., please refer to [24] and its references for a survey of recent advances).

In this work, we propose a new image representation named Bayesian structured sparse coding (BSSC) that connects Gaussian scale mixture (GSM) model with structured sparsity. Our idea is to model each sparse coefficient as a Gaussian distribution with a positive scaling variable and impose a sparse distribution prior (i.e., the Jeffrey prior [36] used in this work) over the positive scaling variables. We show that the Bayesian estimates of both sparse coefficients and scaling variables can be efficiently calculated by alternating minimization. By characterizing the set of sparse coefficients of similar patches with the same prior distribution (i.e., the same non-zero means and positive scaling variables), we can effectively exploit the local and nonlocal dependencies among the sparse coefficients, which have been shown important for image restoration applications [20–22].

Our solution to BSSC can be viewed as an exemplar of demonstrating a new variational approach toward empirical Bayesian inference with parametric models. By contrast to existing variational Bayesian approaches [17,29], the derived BSSC-based image restoration algorithms admits computationally efficient implementations. More importantly, BSSC-based image restoration is capable of both achieving local spatial adaptation and exploiting nonlocal dependencies within an image. Our experimental results have shown that BSSC can restore images whose subjective/objective qualities are often higher than other state-of-the-art methods. Visual quality improvements are attributed to better preservation of edge sharpness and suppression of undesirable artifacts for some images.

1.2 Relationship to Other Competing Approaches

The connection between sparse coding and Bayesian inference has been previously studied in sparse Bayesian learning [27], [28] and more recently in Bayesian compressive sensing [17] and latent variable Bayesian models for promoting sparsity [29]. Despite offering a generic theoretical foundation as well as promising results, the Bayesian inference techniques along this line of works often involve potentially expensive sampling (e.g., approximated solutions for some choice of prior are achieved in [29]). By contrast, our BSSC formulation is conceptually much simpler and admits analytical solutions involving iterative shrinkage/filtering operators only. The other works closely related to the proposed BSSC are the group sparse coding with a Laplacian scale mixture (LSM) [30] and the field-of-GSM [25]. In [30], the LSM model with Gamma distribution imposed over the scaling variables was used to model the sparse coefficients. Approximated estimates of the scale variables were obtained using the Expectation-Maximization (EM) algorithm. Note that the scale variables derived in LSM [30] is very similar to the weights derived in the reweighted $l_1$-norm minimization [38]. In contrast to those approaches, a GSM model with nonzero means and a noninformative sparse prior imposed over scaling variables are used to model the sparse coefficients. Approximated solutions involving iterative shrinkage/filtering operators only over scaling variables are used to model the sparse coefficients here. Instead of using the EM algorithm for an approximated solution, our BSSC offers a more generic Bayesian inference of both scaling variables and sparse coefficients via efficient alternating optimization method. In [25], the field of Gaussian scale mixture model was constructed using the product of two independent homogeneous Gaussian Markov random fields (hGMRFs) to exploit the dependencies between adjacent blocks. Despite similar motivations to exploit the dependencies between the scaling variables, the techniques used in [25] is significantly different and requires a lot more computations than our BSSC framework.

The proposed work is also related to nonparametric Bayesian dictionary learning [31] and solving inverse problems with piecewise linear estimators [32] but ours is motivated by the connection between structured sparsity and the parametric GSM model. When compared with previous works on image denoising (e.g., K-SVD denoising [13], spatially adaptive singular-value thresholding [23] and Expected Patch Log Likelihood (EPLL) [26]), BSSC targets at a more general framework of combining parametric Bayesian inference with dictionary learning. BSSC-based image deblurring has also been experimentally shown superior to existing patch-based methods (e.g., Iterative Decoupled Deblurring BM3D (IDD-BM3D) [33] and Nonlocal centralized sparse representation (NCSR) [34]) and the gain in terms of ISNR is as much as 1.5dB over IDD-BM3D [33] for some test images (e.g., butterfly image - please refer to Fig. 5).
The reminder of this paper is organized as follows. In Sec. 2, we formulate the Bayesian sparse coding problem with Gaussian scale mixture model and generalize it into BSSC by exploiting structured sparsity. In Sec. 3, we elaborate on the details of how to solve BSSC by alternative minimization and emphasize the analytical solutions for both subproblems. In Sec. 4, we study the application of BSSC into image restoration and discuss efficient implementation of BSSC-based image restoration algorithms. In Sec. 5, we report our experimental results in image denoising, image deblurring and image super-resolution as supporting evidence for the effectiveness of BSSC model. In Sec. 6, we make some conclusions about the relationship of sparse coding to image restoration as well as perspectives about the future directions.

2 Bayesian Structured Sparse Coding: Connecting Gaussian Scale Mixture Models with Structured Sparsity

2.1 Bayesian Sparse Coding with Gaussian Scale Mixture Model

The basic idea behind sparse coding is to represent a signal \( x \in \mathbb{R}^n \) (\( n \) is the size of an image patch) as the linear combination of basis vectors (dictionary elements) \( D\alpha \) where \( D \in \mathbb{R}^{n \times K}, n \leq K \) is the dictionary and the coefficients \( \alpha \in \mathbb{R}^K \) satisfies some sparsity constraint. In view of the challenge with \( l_0 \)-optimization, it has been suggested that the original nonconvex optimization is replaced by its \( l_1 \)-counterpart:

\[
\alpha = \arg\min_{\alpha} \| x - D\alpha \|_2^2 + \lambda \| \alpha \|_1, \quad (1)
\]

which is convex and easier to solve. Solving the \( l_1 \)-norm minimization problem corresponds to the MAP inference of \( \alpha \) with an identically independent distributed (i.i.d) Laplacian prior \( P(\alpha_i) = \frac{1}{2\theta^2} e^{-\frac{|\alpha_i|}{2\theta^2}} \), wherein \( \theta \) denotes the standard derivation of \( \alpha_i \). It is easy to verify that the regularization parameter should be set as \( \lambda_i = 2\sigma_n^2/\theta_i \), when the i.i.d Laplacian prior is used, where \( \sigma_n^2 \) denotes the variance of approximation errors. In practice, the variances \( \theta_i \)'s of each \( \alpha_i \) are unknown and may not be easy to accurately estimated from the observation \( x \) considering that real signal/image are non-stationary and may be degraded by noise and blur.

In this paper we propose to model the sparse coefficients \( \alpha \) with the Gaussian scale mixture (GSM) [35] model. The GSM model decomposes coefficient vector \( \alpha \) into the pointwise product of a Gaussian vector \( \beta \) and a hidden scalar multiplier \( \theta \) -i.e., \( \alpha_i = \theta_i \beta_i \), where \( \theta_i \) is the positive scaling variable with probability \( P(\theta_i) \). Conditioned on \( \theta_i \), the coefficients \( \alpha_i \) is Gaussian with standard derivation \( \theta_i \). Assuming that \( \theta_i \) are i.i.d and independent of \( \beta_i \), the GSM prior of \( \alpha \) can be expressed as

\[
P(\alpha) = \prod_i P(\alpha_i), \quad P(\alpha_i) = \int_0^\infty P(\alpha_i|\theta_i)P(\theta_i)d\theta_i. \quad (2)
\]

As a family of probabilistic distributions the GSM model can contain many kurtotic distributions (e.g., the Laplacian, Generalized Gaussian, and student’s t-distribution) given an appropriate \( P(\theta_i) \).

Note that for most of choices of \( P(\theta_i) \) there is no analytical solution of \( P(\alpha_i) \) and thus it is difficult to compute the MAP estimates of \( \alpha_i \). However, such difficulty can be avoided by joint estimation of \( (\alpha_i, \theta_i) \). For a given observation \( x = D\alpha + n \), where \( n \sim N(0, \sigma_n^2) \) denotes the additive Gaussian noise, we can formulate the following MAP estimator

\[
(\alpha, \theta) = \arg\max_{\alpha, \theta} \log P(x|\alpha, \theta)P(\alpha)P(\theta) = \arg\max_{\alpha, \theta} \log P(x|\alpha) + \log P(\alpha|\theta) + \log P(\theta).
\]

(3)

where \( P(x|\alpha) \) is the likelihood term characterized by Gaussian function with variance \( \sigma_n^2 \). The prior term \( P(\alpha|\theta) \) can be expressed as

\[
P(\alpha|\theta) = \prod_i P(\alpha_i|\theta_i) = \prod_i \frac{1}{\sqrt{2\pi} \theta_i \sigma_n^2} e^{-(\alpha_i - \mu_i)^2/2\theta_i^2}.
\]

(4)

Instead of assuming the mean \( \mu_i = 0 \), inspired by our previous work NCSR [34] here we propose to use a biased-mean \( \mu_i \) for \( \alpha_i \) (its estimation will be elaborated later).

The adoption of GSM model allows us to generalize the sparsity from statistical modeling of sparse coefficients \( \alpha \) to the specification of sparse prior \( P(\theta) \). It has been suggested in the literature that noninformative prior \( \{ P(\theta_i) \approx \frac{1}{\theta_i} \} \) - a.k.a. Jeffrey’s prior - is often the favorable choice. Therefore, we have also adopted this option in this work, which translates Eq. (3) into

\[
(\alpha, \theta) = \arg\min_{\alpha, \theta} \frac{1}{2\sigma_n^2} \| x - D\alpha \|_2^2 + \sum_i \log(\theta_i \sqrt{2\pi}) + \sum_i \frac{(\alpha_i - \mu_i)^2}{2\theta_i^2} + \sum_i \log \theta_i.
\]

(5)

where we have used \( P(\theta) = \sum_i P(\theta_i) \) under the assumption with Jeffrey’s prior. The above equation can be further simplified into the following Bayesian sparse coding (BSC) problem

\[
(\alpha, \theta) = \arg\min_{\alpha, \theta} \| x - D\alpha \|_2^2 + 4\sigma_n^2 \log \theta + \sigma_n^2 \sum_i \frac{(\alpha_i - \mu_i)^2}{\theta_i^2}.
\]

(6)
In the matrix form, we have $\alpha = \Lambda \beta$ and $\mu = \Lambda \gamma$ where $\Lambda = \text{diag} (\theta_i) \in R^{K \times K}$ is a diagonal matrix characterizing the variance field for a chosen image patch. Accordingly, Bayesian sparse coding problem in Eq. (5) can be rewritten as

\[
(\beta, \theta) = \text{argmin}_{\beta, \theta} \left\| x - D A \beta \right\|^2_F + 4 \sigma_n^2 \log \theta + \sigma_n^2 \left\| \beta - \gamma \right\|^2.
\]  

(7)

which can be solved by alternatively minimizing the objective functional with respect to $\beta$ and $\theta$.

Unlike [18] that treats the multiplier as a hidden variable and cancel it out through integration (i.e., the derivation of Bayes Least-Square estimate), we explicitly use the field of Gaussian scalar multiplier to characterize the variability and dependencies among local variances. Such BSC formulation of GSM model is appealing because it allows us to further exploit the power of GSM by connecting it with structured sparsity as we will detail next.

2.2 Exploiting Structured Sparsity for the Estimation of the Field of Scalar multipliers

A key observation behind our approach is that for a collection of similar patches, their corresponding sparse coefficients $\alpha$'s should be characterized by the same prior - i.e., the density function with the same $\theta$ and $\mu$. Therefore, if one consider the simultaneous Bayesian sparse coding of GSM models for the collection of $m$ similar patches, a structured/group sparsity based extension of Eq. (7) can be written as

\[
(B, \theta) = \text{argmin}_{B, \theta} \left\| X - D A B \right\|^2_F + 4 \sigma_n^2 \log \theta + \sigma_n^2 \left\| B - \Gamma \right\|^2_F.
\]  

(8)

where $X = [x_1, ..., x_m]$ denotes the collection of $m$ similar patches, $A = \Lambda B$ is the group representation of GSM model for sparse coefficients and their corresponding first-order and second-order statistics are characterized by $\Gamma = [\gamma_1, ..., \gamma_m] \in R^{K \times m}$ and $B = [\beta_1, ..., \beta_m] \in R^{K \times m}$ respectively, wherein $\gamma_j = \gamma, j = 1, 2, ..., m$. From a collection of $m$ similar patches, we have adopted the following strategy for estimating $\mu$

\[
\mu = \sum_{j=1}^{m} w_j \alpha_j,
\]

(9)

where $w_j \sim \text{exp}(-\| x - x_j \|^2_2/h)$ is the weighting coefficient based on patch similarity. It follows from $\mu = \Lambda \gamma$

\[
\gamma = \sum_{j=1}^{m} w_j \Lambda^{-1} \alpha_j = \sum_{j=1}^{m} w_j \beta_j,
\]

(10)

We call such new formulation in Eq. (8) Bayesian structured sparse coding (BSSC) and propose to develop computationally efficient solution to this problem in the next section. Note that here the formulation of BSSC in Eq. (8) is for a given dictionary $D$. However, the dictionary $D$ can also be optimized for a fixed pair of $(B, \theta)$ such that both dictionary learning and statistical modeling of sparse coefficients can be unified within the framework of Eq. (8).

To help readers better appreciate the benefit of BSSC against BSC, we use a toy example as shown in Fig. 1 to help our illustration. If one visualize the data matrix $X$ by reshaping each image patch $x_i$ into a column vector, one can conclude that the local and nonlocal neighborhoods of any pixel are respectively characterized by the elements of the same column and row as the given one. The original (local) GSM model [18] is based on the assumption that the local statistics of a given pixel is invariant conditioned on the local window. Even though GSM denoising does not explicitly estimate the signal variance like other empirical Bayesian methods (e.g., [9],[37]), the knowledge about the variance at a pixel location is acquired from the local neighborhood (i.e., along the row direction of matrix $X$), which corresponds to the BSC formulation of GSM model in Eq. (5). By contrast, in our new formulation inspired by structured sparsity, a more robust approach of acquiring the knowledge about the local statistics at a pixel location is through the samples from the nonlocal neighborhood (i.e., along the column direction of matrix $X$), which corresponds to the BSSC formulation of GSM model in Eq. (8). Despite a similar observation made in our previous work of bilateral variance estimation [23], here we have made one step forward by taking both first-order and second-order statistics of sparse coefficients into the newly-adopted GSM model.

### 3 Solving Bayesian Structured Sparse Coding via Alternating Minimization

In this section, we will show how to solve the optimization problem in Eq. (8) by alternatively updating the estimates of $B$ and $\theta$. The key observation lies in that the two subproblems - minimization of $B$ for a fixed $\theta$ and minimization of $\theta$ for a fixed $B$ - both can be efficiently solved. Specifically, both subproblems admits closed-form solutions when the dictionary is orthogonal.

---

1 Throughout this paper, we will use subscript/superscript to denote column/row vectors of a matrix respectively.
3.1 Solving $\theta$ for a fixed $B$

For a fixed $B$, the first subproblem simply becomes

$$\theta = \arg\min_{\theta} \|X - D\Lambda B\|_F^2 + 4\sigma_n^2 \log \theta,$$

(11)

which can be rewritten as

$$\theta = \arg\min_{\theta} \|X - \sum_{i=1}^{K} d_i \beta_i \theta_i \|_F^2 + 4\sigma_n^2 \log \theta \quad \Rightarrow \quad \theta = \arg\min_{\theta} \|\hat{x} - \hat{D}\theta\|_F^2 + 4\sigma_n^2 \log \theta,$$

(12)

where the long vector $\hat{x} \in R^{m}$ denotes the vectorization of the matrix $X$, the matrix $\hat{D} = [d_1, d_2, \cdots, d_K] \in R^{m \times K}$ whose each column $d_i$ denotes the vectorization of the rank-one matrix $d_i \beta_i$, and $\beta_i \in R^m$ denotes the $i$-th row of matrix $B$. Though $\log \theta$ is non-convex, Eq. (12) can be solved by solving a sequence of reweighted $l_1$-minimization problems [38].

However, the optimization of Eq. (11) can be much simplified when the dictionary $D$ is orthogonal (e.g., DCT or PCA basis). In the case of orthogonal dictionary, Eq. (11) can be rewritten as

$$\theta = \arg\min_{\theta} \|A - AB\|_F^2 + 4\sigma_n^2 \log \theta,$$

(13)

where we have used $X = DA$. Although $\log \theta$ is non-convex, we can efficiently solve it using a local minimization method for a local minimum. Let $f(\theta) = \sum_{i=1}^{K} \log \theta_i$. We can approximate $f(\theta)$ by its first-order Taylor expansion, i.e.,

$$f(\theta^{(k+1)}) = f(\theta^{(k)}) + < \nabla f(\theta^{(k)}), \theta - \theta^{(k)} >.$$  

(14)

where $\theta^{(k)}$ denotes the solution obtained in the $k$-th iteration. By using the fact that $\nabla f(\theta^{(k)}) = \sum_{i=1}^{K} 1/\theta_i^{(k)} + \epsilon$ ($\epsilon$ is a small positive for numerical stability) and ignoring the constants in Eq. (14), Eq. (13) can be solved by iteratively minimizing

$$\theta^{(k+1)} = \arg\min_{\theta} \|A - AB\|_F^2 + 4\sigma_n^2 \|W\|_1,$$

(15)

where $W = \text{diag}(\frac{1}{\theta_i^{(k+1)} + \epsilon})$ is the reweighting matrix that is often used in iterative reweighted $l_1$-minimization.

Since both $A$ and $W$ are diagonal, we can decompose the minimization problem in Eq. (15) into $K$ parallel scalar optimization problems which admit highly efficient implementation. Let $\alpha^i \in R^{1 \times m}$ and $\beta^i \in R^{1 \times m}$ denote the $i$-th row of matrix $A \in R^{n \times m}$ and $B \in R^{n \times m}$, respectively. Eq. (15) can be rewritten as

$$\theta^{(k+1)} = \arg\min_{\theta} \sum_{i=1}^{K} \|\alpha^i \|_T - (\beta^i)T \theta_i \|_2^2 + 4\sigma_n^2 \sum_{i=1}^{K} \frac{\theta_i}{\theta_i^{(k)} + \epsilon},$$

(16)

which can be conveniently decomposed into a sequence of independent scalar optimization problems

$$\theta_i^{(k+1)} = \arg\min_{\theta_i} \|\alpha^i \|_T - (\beta^i)T \theta_i \|_2^2 + 4\sigma_n^2 \frac{\theta_i}{\theta_i^{(k)} + \epsilon},$$

(17)

Now one can see this is standard $l_2$-$l_1$ optimization problem whose closed-form solution is given by

$$\theta_i^{(k+1)} = \frac{1}{(\beta^i)T \tau} [\beta^i (\alpha^i)^T - \tau]_+,,$$

(18)

where the threshold $\tau = \frac{4\sigma_n^2}{\theta_i^{(k)} + \epsilon}$ and $[\cdot]_+$ denotes the soft shrinkage operator.
3.2 Solving $B$ for a fixed $\theta$

The second subproblem is in fact easier to solve than the first one. It takes the following form

$$ B = \arg \min_{B} ||X - DA_{B}||_F^2 + \sigma_n^2 ||B - I||_F^2. $$

(19)

Since both terms are $l_2$, the closed-form solution to Eq. (19) is essentially a Wiener filtering

$$ B = (\hat{D}^T \hat{D} + \sigma_n^2 I)^{-1} (\hat{D}^T X + I). $$

(20)

where $D = DA$. Note that when $D$ is orthogonal, Eq. (20) can be further simplified into

$$ B = (A^T \Lambda + \sigma_n^2 I)^{-1} (A^T \Lambda + I). $$

(21)

where $A^T \Lambda + \sigma_n^2 I$ is a diagonal matrix and therefore its inverse can be easily computed.

By alternatively solving both subproblems of Eqs. (11) and (19) for the estimates of $\Lambda$ and $B$, the image data matrix $X$ can then be reconstructed as

$$ \hat{X} = DA_{\hat{B}}, $$

(22)

where $\hat{B}$ and $\hat{B}$ denotes the final estimates of $\hat{B}$ and $\hat{B}$.

4 Application of Bayesian Structured Sparse Coding into Image Restoration

In the previous sections, we have seen how to solve BSSC problem for a single image data matrix $X$ (a collection of image patches similar to a chosen exemplar). In this section, we generalize such formulation to whole-image reconstruction and study the applications of BSSC into image restoration including image denoising, image deblurring and image super-resolution. The standard image degradation model is used here: $y = Hx + w$ where $x \in R^N, y \in R^M$ denotes the original and degraded images respectively, $H \in R^{N \times M}$ is the degradation matrix and $w$ is additive white Gaussian noise observing $N(0, \sigma^2_n)$. The whole-image reconstruction problem can be expressed as

$$ (x, \{B_l\}, \{\theta_l\}) = \arg \min_{x,\{B_l\},\{\theta_l\}} ||y - Hx||_2^2 $$

$$ + \sum_{l=1}^L (\eta) ||{\hat{R}} l x - DA l B l ||_F^2 $$

$$ + \sigma_n^2 ||B - I||_F^2 + 4\sigma_n^2 log \theta_l. $$

(23)

where $\hat{R} l x = [R_{1l} x, R_{2l} x, \cdots, R_{m_l} x] \in R^{n \times m}$ denotes the data matrix formed by the group of image patches similar to the $l$-th exemplar patch $x_l$ (including $x_l$ itself). $R_l \in R^{n \times N}$ denotes the matrix extracting the $l$-$th$ patch $x_l$ from $x$, and $L$ is the total number of exemplars extracted from the reconstructed image $x$. Invoking the principle of alternative optimization again, we propose to solve the whole-image reconstruction problem in Eq. (23) by alternating the solutions to the following two subproblems:

4.1 Solving $x$ for a fixed $\{B_l\}, \{\theta_l\}$

Let $X_l = DA_l B_l$. When $\{B_l\}$ and $\{\theta_l\}$ are fixed, so is $\{X_l\}$. Therefore, Eq. (23) reduces to the following $l_2$-optimization problem

$$ x = \arg \min_{x} ||y - Hx||_2^2 + \sum_{l=1}^L \eta ||{\hat{R}} l x - X_l||_F^2. $$

(24)

which admits the following closed-form solution

$$ x = (H^T H + \eta \sum_{l=1}^L \hat{R}_l^T \hat{R}_l)^{-1} (H^T y + \eta \sum_{l=1}^L \hat{R}_l^T \hat{X}_l). $$

(25)

where we use $\hat{R}_l \hat{X}_l = \sum_{j=1}^m R_{lj} \hat{x}_j$ and $\hat{R}_l \hat{X}_l \hat{X}_l = \sum_{j=1}^m R_{lj} \hat{x}_j$. $\hat{x}_l$ denotes the $j$-$th$ column of matrix $\hat{X}_l$. Note that for image deblurring application where $H = I$ the matrix to be inverted in Eq. (25) is diagonal, and its inverse can be solved easily. Actually, similar to the K-SVD approach Eq. (25) can be computed by weighted averaging each reconstructed patches sets $\hat{X}_l$. For image deblurring and super-resolution applications, Eq. (25) can be computed by using a conjugate gradient (CG) algorithm.

4.2 Solving $\{B_l\}, \{\theta_l\}$ for a fixed $x$

When $x$ is fixed, the first term in Eq. (23) go away and the subproblem boils down to a sequence of patch-level BSSC problems formed for each exemplar - i.e.,

$$ \{B_l, \theta_l\} = \arg \min_{B_l, \theta_l} ||X_l - DA_l B_l||_F^2 + \frac{\sigma_n^2}{\eta} ||B_l - I||_F^2 + \frac{4\sigma_n^2}{\eta} log \theta_l. $$

(26)

where we use $X_l = \hat{R}_l x$. This is exactly the problem we have studied in the previous section.

One important issue of the BSSC-based image restoration is the selection of the dictionary. To adapt to the local image structures, instead of learning an over-complete dictionary for each dataset $X_l$ as in [21], here similar to NCSR [34] we learn the principle component analysis (PCA) based dictionary for each dataset. The use of the orthogonal dictionary much simplifies the Bayesian inference of BSSC. Putting things together, a complete image restoration based on BSSC can be summarized as follows.
Algorithm 1. BSSC-based Image Restoration

\begin{itemize}
    \item Initialization:
        \begin{enumerate}
            \item set the initial estimate as $\hat{x} = y$ for image denoising and deblurring; or initialize $\hat{x}$ by bicubic interpolation for image super-resolution;
            \item Set parameters $\eta$;
            \item Obtain data matrices $\{X_l\}$’s from $\hat{x}$ (though kNN search) for each exemplar and compute the PCA basis $\{D_l\}$ for each $X_l$.
        \end{enumerate}
    \item Outer loop (solve Eq. (23) by alternative optimization):
        \begin{enumerate}
            \item Iterate on $k = 1, 2, \cdots, k_{\text{max}}$
            \item (a) Image-to-patch transformation: obtain data matrices $\{X_l\}$’s for each exemplar;
            \item (b) Estimate biased means $\mu$ using Eq. (9) for each $X_l$;
            \item (c) Inner loop (solve Eq. (26) for each data $X_l$): iterate on $J = 1, 2, \cdots, J$:
                \begin{enumerate}
                    \item update $\theta_l$ for fixed $B_l$ using Eq. (18);
                    \item update $B_l$ for fixed $\theta_l$ using Eq. (21);
                    \item Reconstruct $X_l$’s from $\theta_l$ and $B_l$ using Eq. (22);
                    \item If $\text{mod}(k, k_0) = 0$, update the PCA basis $\{D_l\}$ for each $X_l$;
                    \item Patch-to-image transformation: obtain reconstructed $\hat{x}^{(k+1)}$ from $\{X_l\}$’s by solving Eq. (25);
                \end{enumerate}
                \item (d) Reconstruct $X_l$’s from $\theta_l$ and $B_l$ using Eq. (22);
                \item (e) If $\text{mod}(k, k_0) = 0$, update the PCA basis $\{D_l\}$ for each $X_l$;
                \item (f) Patch-to-image transformation: obtain reconstructed $\hat{x}^{(k+1)}$ from $\{X_l\}$’s by solving Eq. (25);
        \end{enumerate}
\end{itemize}

In Algorithm 1 we update $D_l$ in every $k_0$ to save computational complexity. We also found that Algorithm 1 empirically converges even when the inner loop executes only one iteration (i.e., $J = 1$). We note that the above algorithm can lead to a variety of implementations depending the choice of degradation matrix $H$. When $H$ is the identity matrix, Algorithm 1 is an image denoising algorithm using iterative regularization technique [39]. When $H$ is a blur matrix or reduced blur matrix, Eq. (23) becomes the standard formulation of non-blind image deblurring or image super-resolution problem. The capability of capturing rapidly-changing statistics in natural images - e.g., through the use of GSM - can make patch-based nonlocal image models even more powerful.

5 Experimental Results

In this section, we report our experimental results of applying BSSC-based image restoration into image denoising, image deblurring and super-resolution. The experimental setup of this work is similar to that in our previous work on NCSR [34]. The basic parameter setting of BSSC is as follows: patch size $6 \times 6$, number of similar blocks $K = 44$; $k_{\text{max}} = 14$, $k_0 = 1$ for image denoising, and $k_{\text{max}} = 450$, $k_0 = 40$ for image deblurring and super-resolution.

To evaluate the quality of restored images, both PSNR and SSIM [40] metrics are used. However, due to limited page space, we can only show part of the experimental results in this paper. More detailed comparisons and complete experimental results are available at the following website: http://www.csee.wvu.edu/~xinl/source.html. The source codes of this paper will be made publicly available after the publication of this paper.

5.1 Image denoising

We have compared BSSC-based image denoising method against three current state-of-the-art methods including BM3D Image Denoising with Shape-Adaptive PCA (BM3D-SAPCA) [41] (it is an enhanced version of BM3D denoising [20] in which local spatial adaptation is achieved by shape-adaptive PCA), learned simultaneous sparse coding (LSSC) [21] and nonlocally centralized sparse representation (NCSR) denoising [34]. As can be seen from Table I, the proposed BSSC has achieved highly competitive denoising performance to other leading algorithms. For the collection of 12 test images, BM3D-SAPCA and BSSC are mostly the best two performing methods - on the average, BSSC falls behind BM3D-SAPCA by less than 0.2dB for three out of six noise levels but deliver at least comparable for the other three. We note that the complexity of BM3D-SAPCA is much higher than that of the original BM3D; by contrast, our pure Matlab implementation of BSSC algorithm (without any C-coded optimization) still runs reasonably fast. It takes around 20 seconds to denoise a $256 \times 256$ image on a PC with an Intel i7-2600 processor at 3.4GHz.

Figs. 2 and 3 include the visual comparison of denoising results for two typical images (lena and house) at moderate ($\sigma_w = 20$) and heavy ($\sigma_w = 100$) noise levels respectively. It can be observed from Fig. 2 that BM3D-SAPCA and BSSC seem to deliver the best visual quality at the moderate noise level; by contrast, restored images by LSSC and NCSR both suffer from noticeable artifacts especially around the smooth areas close to the hat. When the noise contamination is severe, the superiority of BSSC to other competing approaches is easier to justify - as can be seen from Fig. 3, BSSC achieves the most visually pleasant restoration of the house image especially when one inspects the zoomed portions of roof regions closely.

5.2 Image deblurring

We have also compared BSSC-based image deblurring and three other competing approaches in the literature: constrained total variation image deblurring (denoted by FISTA), Iterative Decoupled Deblurring BM3D (IDD-BM3D) [33] and nonlocally centralized sparse representation (NCSR) denoising [34]. Note that the IDD-BM3D and NCSR are two recently developed state-of-the-art non-blind image deblurring approaches. In our comparative study, two commonly-used
### 5.3 Image superresolution

In our study on image super-resolution, simulated LR images are acquired from first applying a $7 \times 7$ uniform blur to the HR image, then down-sampling the blurred image by a factor of 3 along each dimension, and finally adding white Gaussian noise with $\sigma_n^2 = 25$ to the LR images. For color images, we work with the luminance channel only; simple bicubic interpolation method is applied to the upsampling of chrominance channels. Table III includes the PSNR/SSIM comparison for a set of 10 test images among four competing approaches. It can be seen that BSSC outperforms others in most situations (8 out of 10). Visual quality comparison as shown in Figs. 8 and 7 also justifies the superiority of BSSC to other SR techniques.

### 6 Conclusions

In this paper, we proposed a new framework named Bayesian structured sparse coding (BSSC) that connects structured sparsity with Gaussian scale mixture for image restoration.
Fig. 2 Denoising performance comparison on the Lena image with moderate noise corruption. (a) Original image; (b) Noisy image ($\sigma_n = 20$); denoised images by (c) BM3D-SAPCA [41] (PSNR=33.20 dB, SSIM=0.8803); (d) LSSC [21] (PSNR=32.88 dB, SSIM=0.8742); (e) NCSR [34] (PSNR=32.92 dB, SSIM=0.8760); (f) Proposed BSSC (PSNR=33.08, SSIM=0.8787).

BSSC model attempts to characterize both the biased-mean (like in NCSR) and spatially-varying variance (like in GSM) of sparse coefficients. It is shown that the BSSC problem, thanks to the power of alternating direction method of multipliers - can be decomposed into two subproblems both of which admit closed-form solutions when orthogonal basis is used. When applied to image restoration, BSSC leads to computationally efficient algorithms involving iterative shrinkage/filtering only. Our solution to BSSC can be viewed as an exemplar of demonstrating a new variational approach toward empirical Bayesian inference with parametric models. Extensive experimental results have shown that BSSC can both preserve the sharpness of edges and suppress undesirable artifacts more effectively than other competing approaches. This work clearly shows the importance of spatial adaptation regardless the underlying image model is local or nonlocal; in fact, local variations and nonlocal invariance are two sides of the same coin - one has to take both of them into account during the art of image modeling.

In addition to image restoration, BSSC can also be further studied along the line of dictionary learning. In our current implementation, we use PCA basis for its facilitating the derivation of analytical solutions. For non-unitary dictionary, we can solve the BSSC problem by reducing it to iterative reweighted $l_1$-minimization problem [38]. It is also possible to incorporate dictionary $D$ into the optimization problem formulated in Eq. (5); and from this perspective, we can view BSSD as a Bayesian generalization of K-SVD algorithm. Joint optimization of dictionary and sparse coefficients is a more difficult problem and deserves more study. Finally, it is interesting to explore the relationship of BSSC to recent advances in Bayesian nonparametrics [45],[31]. Parametric or nonparametric, we think it will eventually boils down to the capability of the model in striking an optimal tradeoff between local and nonlocal dependencies within image signals.

References
Fig. 3 Denoising performance comparison on the House image with strong noise corruption. (a) Original image; (b) Noisy image ($\sigma_n = 100$); denoised images by (c) BM3D-SAPCA [41] (PSNR=35.20 dB, SSIM=0.6767); (d) LSSC [21] (PSNR=25.63 dB, SSIM=0.7389); (e) NCSR [34] (PSNR=25.65 dB, SSIM=0.7434); (f) Proposed BSSC (PSNR=26.70 dB, SSIM=0.7430).

### Table 2 PSNR (dB) and SSIM results of the deblurred images.

<table>
<thead>
<tr>
<th>Images</th>
<th>Butterfly</th>
<th>Boats</th>
<th>C. Man</th>
<th>Starfish</th>
<th>Parrot</th>
<th>Lena</th>
<th>Barbara</th>
<th>Peppers</th>
<th>Leaves</th>
<th>House</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDD-BM3D [33]</td>
<td>0.9058</td>
<td>0.8355</td>
<td>0.8278</td>
<td>0.8200</td>
<td>0.8750</td>
<td>0.8274</td>
<td>0.7440</td>
<td>0.8143</td>
<td>0.8024</td>
<td>0.8400</td>
<td>0.8400</td>
</tr>
<tr>
<td>NCSR [34]</td>
<td>0.9216</td>
<td>0.8820</td>
<td>0.8580</td>
<td>0.8640</td>
<td>0.9041</td>
<td>0.8654</td>
<td>0.8225</td>
<td>0.8422</td>
<td>0.9418</td>
<td>0.8756</td>
<td>0.8780</td>
</tr>
<tr>
<td>Proposed BSSC</td>
<td>0.9273</td>
<td>0.8810</td>
<td>0.8574</td>
<td>0.8807</td>
<td>0.9103</td>
<td>0.8666</td>
<td>0.8255</td>
<td>0.8402</td>
<td>0.9485</td>
<td>0.8755</td>
<td>0.8814</td>
</tr>
</tbody>
</table>

- Gaussian blur with standard deviation $\sigma = \sqrt{2}$

### Table 3 PSNR (dB) and SSIM results (luminance components) of the reconstructed HR images.

<table>
<thead>
<tr>
<th>Images</th>
<th>Butterfly</th>
<th>Parrot</th>
<th>Plants</th>
<th>Hat</th>
<th>flower</th>
<th>Racoon</th>
<th>Bike</th>
<th>Patheon</th>
<th>Girl</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV [43]</td>
<td>26.56</td>
<td>27.85</td>
<td>0.8797</td>
<td>29.20</td>
<td>27.51</td>
<td>27.54</td>
<td>23.66</td>
<td>26.00</td>
<td>31.24</td>
<td>27.88</td>
</tr>
<tr>
<td>Sparsity [44]</td>
<td>0.9000</td>
<td>0.8900</td>
<td>0.8909</td>
<td>0.8843</td>
<td>0.8148</td>
<td>0.7070</td>
<td>0.7582</td>
<td>0.7232</td>
<td>0.7880</td>
<td>0.8121</td>
</tr>
<tr>
<td>NCSR [34]</td>
<td>28.70</td>
<td>28.70</td>
<td>31.55</td>
<td>29.63</td>
<td>27.87</td>
<td>28.51</td>
<td>23.23</td>
<td>26.27</td>
<td>32.87</td>
<td>28.15</td>
</tr>
<tr>
<td>Proposed BSSC</td>
<td>28.10</td>
<td>30.50</td>
<td>34.00</td>
<td>31.27</td>
<td>29.50</td>
<td>29.28</td>
<td>24.74</td>
<td>27.19</td>
<td>33.65</td>
<td>29.80</td>
</tr>
</tbody>
</table>

- Noisless

<table>
<thead>
<tr>
<th>Images</th>
<th>Butterfly</th>
<th>Parrot</th>
<th>Plants</th>
<th>Hat</th>
<th>flower</th>
<th>Racoon</th>
<th>Bike</th>
<th>Patheon</th>
<th>Girl</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparsity [44]</td>
<td>0.8477</td>
<td>0.8139</td>
<td>0.8047</td>
<td>0.7017</td>
<td>0.7557</td>
<td>0.6632</td>
<td>0.7131</td>
<td>0.6697</td>
<td>0.7291</td>
<td>0.7519</td>
</tr>
<tr>
<td>NCSR [34]</td>
<td>23.61</td>
<td>27.15</td>
<td>29.57</td>
<td>28.31</td>
<td>26.60</td>
<td>27.22</td>
<td>23.45</td>
<td>25.40</td>
<td>30.71</td>
<td>26.78</td>
</tr>
<tr>
<td>Proposed BSSC</td>
<td>0.7552</td>
<td>0.7738</td>
<td>0.7700</td>
<td>0.7212</td>
<td>0.7052</td>
<td>0.6422</td>
<td>0.6477</td>
<td>0.6205</td>
<td>0.7051</td>
<td>0.7043</td>
</tr>
</tbody>
</table>

- Noisy


Fig. 4 Deblurring performance comparison on the Starfish image. (a) Original image; (b) Noisy and blurred image ($9 \times 9$ uniform blur, $\sigma_n = \sqrt{2}$); deblurred images by (c) FISTA [42] (PSNR=27.75 dB, SSIM=0.8200); (d) IDD-BM3D [33] (PSNR=29.48 dB, SSIM=0.8640); (e) NCSR [34] (PSNR=30.28 dB, SSIM=0.8807); (f) Proposed BSSC (PSNR=30.58 dB, SSIM=0.8862).


Fig. 5 Deblurring performance comparison on the Butterfly image. (a) Original image; (b) Noisy and blurred image ($9 \times 9$ uniform blur, $\sigma_n = \sqrt{2}$); deblurred images by (c) FISTA [42] (PSNR=28.37 dB, SSIM=0.9058); (d) IDD-BM3D [33] (PSNR=29.21 dB, SSIM=0.9216); (e) NCSR [34] (PSNR=29.68 dB, SSIM=0.9273); (f) Proposed BSSC (PSNR=30.45 dB, SSIM=0.9377).
Fig. 6 Deblurring performance comparison on the Barbara image. (a) Original image; (b) Noisy and blurred image (Gaussian blur, $\sigma_n = \sqrt{2}$); deblurred images by (c) FISTA [42] (PSNR=25.03 dB, SSIM=0.7377); (d) IDD-BM3D [33] (PSNR=27.19 dB, SSIM=0.8231); (e) NCSR [34] (PSNR=27.91 dB, SSIM=0.8304); (f) Proposed BSSC (PSNR=28.42 dB, SSIM=0.8462).
Fig. 7 Image super-resolution performance comparison on the *Plant* image (scaling factor 3, $\sigma_n = 0$). (a) Original image; (b) Low-resolution image; reconstructed images by (c) TV [43] (PSNR=31.34 dB, SSIM=0.8797); (d) Sparsity-based [44] (PSNR=31.55 dB, SSIM=0.8964); (e) NCSR [34] (PSNR=34.00 dB, SSIM=0.9369); (f) Proposed BSSC (PSNR=34.33 dB, SSIM=0.9236).
Fig. 8 Image super-resolution performance comparison on the *Hat* image (scaling factor 3, \( \sigma_n = 5 \)). (a) Original image; (b) Low-resolution image; reconstructed images by (c) TV [43] (PSNR=28.13 dB, SSIM=0.7701); (d) Sparsity-based [44] (PSNR=28.31 dB, SSIM=0.7212); (e) NCSR [34] (PSNR=29.94 dB, SSIM=0.8238); (f) Proposed BSSC (PSNR=\textbf{30.21} dB, SSIM=\textbf{0.8354}).
Image Restoration via Bayesian Structured Sparse Coding: Where Structured Sparsity Meets Gaussian Scale Mixture

Weisheng Dong · Guangming Shi · Yi Ma · Xin Li

Abstract Sparse coding has been shown to be relevant to both variational and Bayesian approaches. The regularization parameter in variational image restoration is intrinsically connected with the shape parameters of sparse coefficients' distributions in Bayesian methods. How to tune those parameters in a principled yet spatially adaptive fashion turns out a challenging problem especially for the class of nonlocal image models. In this work, we propose a Bayesian structured sparse coding (BSSC) framework to address this issue - more specifically, a nonlocal extension of Gaussian scale mixture (GSM) model is developed based on structured sparsity and its related maximum a posterior (MAP) estimation problem is tackled. It is shown that the variances of sparse coefficients (the field of scalar multipliers of Gaussian) - if treated as a latent variable - can be jointly estimated along with the unknown sparse coefficients via the method of alternating optimization. Unlike previous attacks to MAP estimation, ours leads to closed-form solutions involving iterative shrinkage/filtering only and therefore admits computationally efficient implementation. Our solution to BSSC can be viewed as an exemplar of demonstrating a new variational approach toward empirical Bayesian inference with parametric models, which has computational advantages over the existing variational Bayesian approaches. When applied to image restoration, our experimental results have shown that BSSC can both preserve the sharpness of edges and suppress undesirable artifacts. Thanks to its capability of achieving a better spatial adaptation, BSSC-based image restoration often delivers reconstructed images with higher subjective/objective qualities than other competing approaches.

Keywords Bayesian sparse coding · Gaussian scale mixture · structured sparsity · alternative minimization · variational image restoration

1 Introduction

1.1 Background and Motivation

Sparse representation of signals/images has been widely studied by signal/image processing communities in the past decades. Historically, the idea of sparsity dated back to the coring operator invented by RCA researchers in early 1980s [1]. The birth of wavelet [2] or filter bank theory [3] or multi-resolution analysis [4] in late 1980s rapidly sparked the interest in sparse representation, which has found successful applications into image coding [5], [6], [7] and denoising [8], [9], [10]. Under the framework of sparse coding, a lot of research have been centered at two related issues: basis functions (or dictionary) and statistical modeling of sparse coefficients. Exemplar studies of the former are the construction of directional multiresolution representation (e.g., contourlet [11]) and over-complete dictionary learning from training data (e.g., K-SVD [12,13], multiscale dictionary learning [16], online dictionary learning [14] and the non-parametric Bayesian dictionary learning [15]); the latter in-
include the use of Gaussian mixture models [26, 32], the variational Bayesian models [17, 29], the universal models [24], and the centralized Laplacian model [34] for sparse coefficients.

More recently, a class of nonlocal image restoration techniques [19], [20], [21], [41], [22], [23, 34] have attracted increasingly more attention. The key motivation behind lies in the observation that many important image structures in natural images including edges and textures can be characterized by the abundance of self-repeating patterns. Such observation has led to the formulation of nonlocal simultaneous sparse coding [21]. Our own recent works [22], [23] has more clearly verified the potential of exploiting structured sparsity in image restoration. However, a fundamental question remains open: how to achieve (local) spatial adaptation within the framework of nonlocal image restoration? This question is related to the issue of regularization parameters in a variational setting or shape parameters in a Bayesian one; but the issue becomes even more thorny when it is tangled with nonlocal regularization/sparsity. To the best of our knowledge, how to tune those parameters in a principled manner remains an open problem (e.g., please refer to [24] and its references for a survey of recent advances).

In this work, we propose a new image representation named Bayesian structured sparse coding (BSSC) that connects Gaussian scale mixture (GSM) model with structured sparsity. Our idea is to model each sparse coefficient as a Gaussian distribution with a positive scaling variable and impose a sparse distribution prior (i.e., the Jeffrey prior [36] used in this work) over the positive scaling variables. We show that the Bayesian estimates of both sparse coefficients and scaling variables can be efficiently calculated by alternating minimization. By characterizing the set of sparse coefficients of similar patches with the same prior distribution (i.e., the same non-zero means and positive scaling variables), we can effectively exploit the local and nonlocal dependencies among the sparse coefficients, which have been shown important for image restoration applications [20–22]. Our solution to BSSC can be viewed as an exemplar of demonstrating a new variational approach toward empirical Bayesian inference with parametric models. By contrast to existing variational Bayesian approaches [17, 29], the derived BSSC-based image restoration algorithms admits computationally efficient implementations. More importantly, BSSC-based image restoration is capable of both achieving local spatial adaptation and exploiting nonlocal dependencies within an image. Our experimental results have shown that BSSC can restore images whose subjective/objective qualities are often higher than other state-of-the-art methods. Visual quality improvements are attributed to better preservation of edge sharpness and suppression of undesirable artifacts for some images.

1.2 Relationship to Other Competing Approaches

The connection between sparse coding and Bayesian inference has been previously studied in sparse Bayesian learning [27], [28] and more recently in Bayesian compressive sensing [17] and latent variable Bayesian models for promoting sparsity [29]. Despite offering a generic theoretical foundation as well as promising results, the Bayesian inference techniques along this line of works often involve potentially expensive sampling (e.g., approximated solutions for some choice of prior are achieved in [29]). By contrast, our BSSC formulation is conceptually much simpler and admits analytical solutions involving iterative shrinkage/filtering operators only. The other works closely related to the proposed BSSC are the group sparse coding with a Laplacian scale mixture (LSM) [30] and the field-of-GSM [25]. In [30], the LSM model with Gamma distribution imposed over the scaling variables was used to model the sparse coefficients. Approximated estimates of the scale variables were obtained using the Expectation-Maximization (EM) algorithm. Note that the scale variables derived in LSM [30] is very similar to the weights derived in the reweighted $l_1$-norm minimization [38]. In contrast to those approaches, a GSM model with nonzero means and a noninformative sparse prior imposed over scaling variables are used to model the sparse coefficients. Approximated estimates of the scale variables were obtained using the Expectation-Maximization (EM) algorithm. Note that the scale variables derived in LSM [30] is very similar to the weights derived in the reweighted $l_1$-norm minimization [38]. In contrast to those approaches, a GSM model with nonzero means and a noninformative sparse prior imposed over scaling variables are used to model the sparse coefficients. Instead of using the EM algorithm for an approximated solution, our BSSC offers a more generic Bayesian inference of both scaling variables and sparse coefficients via efficient alternating optimization method. In [25], the field of Gaussian scale mixture model was constructed using the product of two independent homogeneous Gaussian Markov random fields (hGMRFs) to exploit the dependencies between adjacent blocks. Despite similar motivations to exploit the dependencies between the scaling variables, the techniques used in [25] is significantly different and requires a lot more computations than our BSSC framework.

The proposed work is also related to nonparametric Bayesian dictionary learning [31] and solving inverse problems with piecewise linear estimators [32] but ours is motivated by the connection between structured sparsity and the parametric GSM model. When compared with previous works on image denoising (e.g., K-SVD denoising [13], spatially adaptive singular-value thresholding [23] and Expected Patch Log Likelihood (EPLL) [26]), BSSC targets at a more general framework of combining parametric Bayesian inference with dictionary learning. BSSC-based image deblurring has also been experimentally shown superior to existing patch-based methods (e.g., Iterative Decoupled Deblurring BM3D (IDD-BM3D) [33] and Nonlocal centralized sparse representation (NCSR) [34]) and the gain in terms of ISNR is as much as 1.5dB over IDD-BM3D [33] for some test images (e.g., butterfly image - please refer to Fig. 5).
2 Bayesian Structured Sparse Coding: Connecting Gaussian Scale Mixture Models with Structured Sparsity

2.1 Bayesian Sparse Coding with Gaussian Scale Mixture Model

The basic idea behind sparse coding is to represent a signal \( x \in R^n \) (\( n \) is the size of an image patch) as the linear combination of basis vectors (dictionary elements) \( D\alpha \) where \( D \in R^{n\times K}, n \leq K \) is the dictionary and the coefficients \( \alpha \in R^K \) satisfies some sparsity constraint. In view of the challenge with \( l_0 \)-optimization, it has been suggested that the original nonconvex optimization is replaced by its \( l_1 \)-counterpart:

\[
\alpha = \arg\min_{\alpha} ||x - D\alpha||^2_2 + \lambda||\alpha||_1, \tag{1}
\]

which is convex and easier to solve. Solving the \( l_1 \)-norm minimization problem corresponds to the MAP inference of \( \alpha \) with an identically independent distributed (i.i.d) Laplace prior \( P(\alpha_i) = \frac{1}{2\sigma_n^2}e^{-\frac{|\alpha_i|}{\sigma_n}} \), wherein \( \theta_i \) denotes the standard derivation of \( \alpha_i \). It is easy to verify that the regularization parameter should be set as \( \lambda_i = 2\sigma_n^2/\theta_i \), when the i.i.d Laplacian prior is used, where \( \sigma_n^2 \) denotes the variance of approximation errors. In practice, the variances \( \theta_i \)'s of each \( \alpha_i \) are unknown and may not be easily accurately estimated from the observation \( x \) considering that real signal/image are non-stationary and may be degraded by noise and blur.

In this paper we propose to model the sparse coefficients \( \alpha \) with the Gaussian scale mixture (GSM) [35] model. The GSM model decomposes coefficient vector \( \alpha \) into the point-wise product of a Gaussian vector \( \beta \) and a hidden scalar multiplier \( \theta \) -i.e., \( \alpha_i = \theta_i\beta_i \), where \( \theta_i \) is the positive scaling variable with probability \( P(\theta_i) \). Conditioned on \( \theta_i \), the coefficients \( \alpha_i \) is Gaussian with standard derivation \( \theta_i \). Assuming that \( \theta_i \) are i.i.d and independent of \( \beta_i \), the GSM prior of \( \alpha \) can be expressed as

\[
P(\alpha) = \prod_i P(\alpha_i), \quad P(\alpha) = \int_0^{\infty} P(\alpha_i|\theta_i)P(\theta_i)d\theta_i. \tag{2}
\]

As a family of probabilistic distributions the GSM model can contain many kurtotic distributions (e.g., the Laplacian, Generalized Gaussian, and student’s t-distribution) given an appropriate \( P(\theta_i) \).

Note that for most of choices of \( P(\theta_i) \) there is no analytical expression of \( P(\alpha_i) \) and thus it is difficult to compute the MAP estimates of \( \alpha_i \). However, such difficulty can be avoided by joint estimation of \( (\alpha_i, \theta_i) \). For a given observation \( x = D\alpha + n \), where \( n \sim N(0, \sigma_n^2) \) denotes the additive Gaussian noise, we can formulate the following MAP estimator

\[
(\alpha, \theta) = \arg\max_{\alpha, \theta} \log P(x|\alpha)P(\alpha, \theta) \tag{3}
\]

where \( P(x|\alpha) \) is the likelihood term characterized by Gaussian function with variance \( \sigma_n^2 \). The prior term \( P(\alpha, \theta) \) can be expressed as

\[
P(\alpha, \theta) = \prod_i P(\alpha_i|\theta_i) = \prod_i \frac{1}{\theta_i\sqrt{2\pi}}exp\left(-\frac{(\alpha_i - \mu_i)^2}{2\theta_i^2}\right). \tag{4}
\]

Instead of assuming the mean \( \mu_i = 0 \), inspired by our previous work NCSR [34] here we propose to use a biased-mean \( \mu_i \) for \( \alpha_i \) (its estimation will be elaborated later).

The adoption of GSM model allows us to generalize the sparsity from statistical modeling of sparse coefficients \( \alpha \) to the specification of sparse prior \( P(\theta) \). It has been suggested in the literature that noninformative prior [36] \( P(\theta_i) \approx \frac{1}{\theta_i} \) - a.k.a. Jeffrey’s prior - is often the favorable choice. Therefore, we have also adopted this option in this work, which translates Eq. (3) into

\[
(\alpha, \theta) = \arg\min_{\alpha, \theta} \frac{1}{2\sigma_n^2}||x - D\alpha||^2_2 + \sum_i \log(\theta_i\sqrt{2\pi}) + \sum_i \frac{(\alpha_i - \mu_i)^2}{2\theta_i^2} + \sum_i \log(\theta_i). \tag{5}
\]

where we have used \( P(\theta) = \sum_i P(\theta_i) \) under the assumption with Jeffrey’s prior. The above equation can be further simplified into the following Bayesian sparse coding (BSC) problem

\[
(\alpha, \theta) = \arg\min_{\alpha, \theta} ||x - D\alpha||^2_2 + 4\sigma_n^2\log\theta + \sigma_n^2 \sum_i \frac{(\alpha_i - \mu_i)^2}{\theta_i^2}. \tag{6}
\]
In the matrix form, we have $\alpha = A\beta$ and $\mu = A\gamma$ where
$A = \text{diag}(\theta_i) \in \mathbb{R}^{K \times K}$ is a diagonal matrix characterizing

the variance field for a chosen image patch. Accordingly, the Bayesian sparse coding problem in Eq. (5) can be rewritten as

$$
(\beta, \theta) = \underset{\beta, \theta}{\text{argmin}} \|x - DA\beta\|_F^2 + 4\sigma_n^2 \log |\theta| + \sigma_n^2 \|\beta - \gamma\|_2^2.
$$

(7)

which can be solved by alternatively minimizing the objective functional with respect to $\beta$ and $\theta$.

Unlike [18] that treats the multiplier as a hidden variable and cancel it out through integration (i.e., the derivation of Bayes Least-Square estimate), we explicitly use the field of Gaussian scalar multiplier to characterize the variability and dependencies among local variances. Such BSC formulation of GSM model is appealing because it allows us to further exploit the power of GSM by connecting it with structured sparsity as we will detail next.

2.2 Exploiting Structured Sparsity for the Estimation of the Field of Scalar multipliers

A key observation behind our approach is that for a collection of similar patches, their corresponding sparse coefficients $\alpha$’s should be characterized by the same prior - i.e., the density function with the same $\theta$ and $\mu$. Therefore, if one consider the simultaneous Bayesian sparse coding of GSM models for the collection of $m$ similar patches, a structured/group sparsity based extension of Eq. (7) can be written as

$$
(B, \theta) = \underset{B, \theta}{\text{argmin}} \|X - DB\beta\|_F^2 + 4\sigma_n^2 \log |\theta| + \sigma_n^2 \|B - \Gamma\|_F^2.
$$

(8)

where $X = [x_1, \ldots, x_m]$ denotes the collection of $m$ similar patches, $A = AB$ is the group representation of GSM model for sparse coefficients and their corresponding first-order and second-order statistics are characterized by $\Gamma = [\gamma_1, \ldots, \gamma_m] \in \mathbb{R}^{K \times m}$ and $B = [\beta_1, \ldots, \beta_m] \in \mathbb{R}^{K \times m}$ respectively, wherein $\gamma_j = \gamma, j = 1, 2, \cdots, m$. From a collection of $m$ similar patches, we have adopted the following strategy for estimating $\mu$

$$
\mu = \sum_{j=1}^{m} w_j \alpha_j,
$$

(9)

where $w_j \sim exp(-\|x - x_j\|_2^2/h))$ is the weighting coefficient based on patch similarity. It follows from $\mu = A\gamma$

$$
\gamma = \sum_{j=1}^{m} w_j A^{-1} \alpha_j = \sum_{j=1}^{m} w_j \beta_j,
$$

(10)

We call such new formulation in Eq.(8) Bayesian structured sparse coding (BSSC) and propose to develop computationally efficient solution to this problem in the next section. Note that here the formulation of BSSC in Eq. (8) is for a given dictionary $D$. However, the dictionary $D$ can also be optimized for a fixed pair of $(B, \theta)$ such that both dictionary learning and statistical modeling of sparse coefficients can be unified within the framework of Eq. (8).

To help readers better appreciate the benefit of BSSC against BSC, we use a toy example as shown in Fig. 1 to help our illustration. If one visualize the data matrix $X$ by reshaping each image patch $x_i$ into a column vector, one can conclude that the local and nonlocal neighborhoods of any pixel are respectively characterized by the elements of the same column and row as the given one. The original (local) GSM model [18] is based on the assumption that the local statistics of a given pixel is invariant conditioned on the local window. Even though GSM denoising does not explicitly estimate the signal variance like other empirical Bayesian methods (e.g., [9],[37]), the knowledge about the variance at a pixel location is acquired from the local neighborhood (i.e., along the row direction of matrix $X$), which corresponds to the BSC formulation of GSM model in Eq. (5). By contrast, in our new formulation inspired by structured sparsity, a more robust approach of acquiring the knowledge about the local statistics at a pixel location is through the samples from the nonlocal neighborhood (i.e., along the column direction of matrix $X$), which corresponds to the BSSC formulation of GSM model in Eq. (8). Despite a similar observation made in our previous work of bilateral variance estimation [23], here we have made one step forward by taking both first-order and second-order statistics of sparse coefficients into the newly-adopted GSM model.

3 Solving Bayesian Structured Sparse Coding via Alternating Minimization

In this section, we will show how to solve the optimization problem in Eq. (8) by alternatively updating the estimates of $B$ and $\theta$. The key observation lies in that the two subproblems - minimization of $B$ for a fixed $\theta$ and minimization of $\theta$ for a fixed $B$ - both can be efficiently solved. Specifically, both subproblems admits closed-form solutions when the dictionary is orthogonal.
3.1 Solving $\theta$ for a fixed $B$

For a fixed $B$, the first subproblem simply becomes

$$\theta = \arg\min_{\theta} \|X - DB\|_F^2 + 4\sigma_n^2 \log \theta, \quad (11)$$

which can be rewritten as

$$\theta = \arg\min_{\theta} \|X - \sum_{i=1}^{K} d_i \beta_i^t \|_F^2 + 4\sigma_n^2 \log \theta$$

$$= \arg\min_{\theta} \|\tilde{x} - \hat{D}\theta\|_2^2 + 4\sigma_n^2 \log \theta, \quad (12)$$

where the long vector $\tilde{x} \in \mathbb{R}^{n \times m}$ denotes the vectorization of the matrix $X$, the matrix $\hat{D} = [d_1, d_2, \ldots, d_K] \in \mathbb{R}^{n \times K}$ whose each column $d_i$ denotes the vectorization of the rank-one matrix $d_i \beta_i^t$, and $\beta_i^t \in \mathbb{R}^{n \times m}$ denote the $i$-th row of matrix $B$. Though $\log \theta$ is non-convex, Eq. (12) can be solved by solving a sequence of reweighted $l_1$-minimization problems [38].

However, the optimization of Eq. (11) can be much simplified when the dictionary $D$ is orthogonal (e.g., DCT or PCA basis). In the case of orthogonal dictionary, Eq. (11) can be rewritten as

$$\theta = \arg\min_{\theta} \|A - AB\|_F^2 + 4\sigma_n^2 \log \theta, \quad (13)$$

where we have used $X = DA$. Although $\log \theta$ is non-convex, we can efficiently solve it using a local minimization method for a local minimum. Let $f(\theta) = \sum_{i=1}^{K} \log \theta_i$. We can approximate $f(\theta)$ by its first-order Taylor expansion, i.e.,

$$f(\theta^{(k+1)}) = f(\theta^{(k)}) + \nabla f(\theta^{(k)}) \cdot (\theta - \theta^{(k)}) = f(\theta^{(k)}) - \nabla f(\theta^{(k)}) ||\theta - \theta^{(k)}||_2 + \epsilon. \quad (14)$$

where $\epsilon$ is a small positive for numerical stability and ignoring the constants in Eq. (14), Eq. (13) can be solved by iteratively minimizing

$$\theta^{(k+1)} = \arg\min_{\theta} \|A - AB\|_F^2 + 4\sigma_n^2 \|W\|_1. \quad (15)$$

where $W = \text{diag}(\frac{1}{\beta^{(k)}_i + \epsilon})$ is the reweighting matrix that is often used in iterative reweighted $l_1$-minimization.

Since both $A$ and $W$ are diagonal, we can decompose the minimization problem in Eq. (15) into $K$ parallel scalar optimization problems which admit highly efficient implementation. Let $\alpha^t \in \mathbb{R}^{1 \times m}$ and $\beta^t \in \mathbb{R}^{1 \times m}$ denote the $i$-th row of matrix $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times m}$, respectively. Eq. (15) can be rewritten as

$$\theta_i^{(k+1)} = \arg\min_{\theta_i} \sum_{i=1}^{K} ||(\alpha^t)^T - (\beta^t)^T \theta_i||_2^2 + 4\sigma_n^2 \sum_{i=1}^{K} \frac{\theta_i}{\theta_i^{(k)} + \epsilon} + \epsilon \quad (16)$$

which can be conveniently decomposed into a sequence of independent scalar optimization problems

$$\theta_i^{(k+1)} = \arg\min_{\theta_i} \sum_{i=1}^{K} ||(\alpha^t)^T - (\beta^t)^T \theta_i||_2^2 + 4\sigma_n^2 \frac{\theta_i}{\theta_i^{(k)} + \epsilon} + \epsilon \quad (17)$$

Now one can see this is standard $l_2$-$l_1$ optimization problem whose closed-form solution is given by

$$\theta_i^{(k+1)} = \frac{1}{\beta^t(\beta^t)^T} [\beta^t(\alpha^t)^T - \tau], \quad (18)$$

where the threshold $\tau = \frac{4\sigma_n^2}{\theta_i^{(k)} + \epsilon}$ and $[\cdot]_+$ denotes the soft shrinkage operator.
3.2 Solving \( B \) for a fixed \( \theta \)

The second subproblem is in fact easier to solve than the first one. It takes the following form

\[
B = \arg \min \| X - D \Lambda B \|_F^2 + \sigma_n^2 \| B - \Gamma \|_F^2. \tag{19}
\]

Since both terms are \( l_2 \), the closed-form solution to Eq. (19) is essentially a Wiener filtering

\[
B = (\hat{D}^T \hat{D} + \sigma_n^2 I)^{-1} (\hat{D}^T X + \Gamma), \tag{20}
\]

where \( \hat{D} = D \Lambda \). Note that when \( D \) is orthogonal, Eq. (20) can be further simplified into

\[
B = (A^T \Lambda + \sigma_n^2 I)^{-1} (A^T A + \Gamma). \tag{21}
\]

where \( \Lambda^T A + \sigma_n^2 I \) is a diagonal matrix and therefore its inverse can be easily computed.

By alternatively solving both subproblems of Eqs. (11) and (19) for the estimates of \( \Lambda \) and \( B \), the image data matrix \( X \) can then be reconstructed as

\[
\hat{X} = D \hat{\Lambda} \hat{B}, \tag{22}
\]

where \( \hat{\Lambda} \) and \( \hat{B} \) denotes the final estimates of \( \Lambda \) and \( B \).

4 Application of Bayesian Structured Sparse Coding into Image Restoration

In the previous sections, we have seen how to solve BSSC problem for a single image data matrix \( X \) (a collection of image patches similar to a chosen exemplar). In this section, we generalize such formulation to whole-image reconstruction and study the applications of BSSC into image restoration including image denoising, image deblurring and image superresolution. The standard image degradation model is used here: \( y = Hx + w \) where \( x \in \mathbb{R}^N \), \( y \in \mathbb{R}^M \) denotes the original and degraded images respectively, \( H \in \mathbb{R}^{N \times M} \) is the degradation matrix and \( w \) is additive white Gaussian noise observing \( N(0, \sigma_n^2) \). The whole-image reconstruction problem can be expressed as

\[
(x, \{B_l\}, \{\theta_l\}) = \arg \min_{x, \{B_l\}, \{\theta_l\}} \| y - Hx \|_2^2
+ \sum_{l=1}^L (\eta_l \| \hat{R}_l x - D \Lambda_l B_l \|_F^2
+ \sigma_n^2 \| B_l - \Gamma \|_F^2 + 4\sigma_n^2 \log \theta_l). \tag{23}
\]

where \( \hat{R}_l x = [R_{1l} x, R_{2l} x, \cdots, R_{ml} x] \in \mathbb{R}^{m \times M} \) denotes the data matrix formed by the group of image patches similar to the \( l \)-th exemplar patch \( x_l \) (including \( x_l \) itself), \( R_l \in \mathbb{R}^{m \times N} \) denotes the matrix extracting the \( l \)-th patch \( x_l \) from \( x \), and \( L \) is the total number of exemplars extracted from the reconstructed image \( x \). Invoking the principle of alternative optimization again, we propose to solve the whole-image reconstruction problem in Eq. (23) by alternating the solutions to the following two subproblems:

4.1 Solving \( x \) for a fixed \( \{B_l\}, \{\theta_l\} \)

Let \( \hat{X}_l = D \Lambda_l B_l \). When \( \{B_l\} \) and \( \{\theta_l\} \) are fixed, so is \( \{\hat{X}_l\} \). Therefore, Eq. (23) reduces to the following \( l_2 \)-optimization problem

\[
x = \arg \min_{x} \| y - Hx \|_2^2 + \sum_{l=1}^L \eta_l \| R_l x - \hat{X}_l \|_F^2. \tag{24}
\]

which admits the following closed-form solution

\[
x = (H^T H + \eta \sum_{l=1}^L \hat{R}_l^T \hat{R}_l)^{-1} (H^T y + \eta \sum_{l=1}^L \hat{R}_l^T \hat{X}_l). \tag{25}
\]

where we use \( \hat{R}_l \) defined as above. Note that for image denoising application where \( H = I \) the matrix to be inverted in Eq. (25) is diagonal, and its inverse can be solved easily. Actually, similar to the K-SVD approach Eq. (25) can be computed by weighted averaging each reconstructed patches sets \( \hat{X}_l \). For image deblurring and super-resolution applications, Eq. (25) can be computed by using a conjugate gradient (CG) algorithm.

4.2 Solving \( \{B_l\}, \{\theta_l\} \) for a fixed \( x \)

When \( x \) is fixed, the first term in Eq. (23) go away and the subproblem boils down to a sequence of patch-level BSSC problems formed for each exemplar - i.e.,

\[
(B_l, \theta_l) = \arg \min_{B_l, \theta_l} \| X_l - D \Lambda_l B_l \|_F^2 + \sigma_n^2 \| B_l - \Gamma \|_F^2 + 4\sigma_n^2 \log \theta_l. \tag{26}
\]

where we use \( X_l = \hat{R}_l x \). This is exactly the problem we have studied in the previous section.

One important issue of the BSSC-based image restoration is the selection of the dictionary. To adapt to the local image structures, instead of learning an over-complete dictionary for each dataset \( X_l \), as in [21], here similar to NCSR [34] we learn the principle component analysis (PCA) based dictionary for each exemplar. The use of the orthogonal dictionary much simplifies the Bayesian inference of BSSC. Putting things together, a complete image restoration based on BSSC can be summarized as follows.
Algorithm 1. BSSC-based Image Restoration

- Initialization:
  (a) set the initial estimate as \( \hat{x} = y \) for image denoising and deblurring; or initialize \( \hat{x} \) by bicubic interpolation for image super-resolution;
  (b) Set parameters \( \eta \);
  (c) Obtain data matrices \( \{X_l\} \)'s from \( \hat{x} \) (though kNN search) for each exemplar and compute the PCA basis \( \{D_l\} \) for each \( X_l \).
- Outer loop (solve Eq. (23) by alternative optimization):
  Iterate on \( k = 1, 2, \cdots , k_{\text{max}} \):
  (a) Image-to-patch transformation: obtain data matrices \( \{X_l\} \)'s for each exemplar;
  (b) Estimate biased means \( \mu \) using Eq. (9) for each \( X_l \);
  (c) Inner loop (solve Eq. (26) for each data \( X_l \)): iterate on \( J = 1, 2, \cdots , J_l \):
    (I) update \( \theta_l \) for fixed \( B_l \) using Eq. (18);
    (II) update \( B_l \) for fixed \( \theta_l \) using Eq. (21);
    (d) Reconstruct \( X_l \)'s from \( \theta_l \) and \( B_l \) using Eq. (22);
    (e) If \( \text{mod}(k, k_0) = 0 \), update the PCA basis \( \{D_l\} \) for each \( X_l \);
    (f) Patch-to-image transformation: obtain reconstructed \( \hat{x}^{(k+1)} \) from \( \{X_l\} \)'s by solving Eq. (25);
- Output: \( \hat{x}^{(k+1)} \).

In Algorithm 1 we update \( D_l \) in every \( k_0 \) to save computational complexity. We also found that Algorithm 1 empirically converges even when the inner loop executes only one iteration (i.e., \( J = 1 \)). We note that the above algorithm can lead to a variety of implementations depending the choice of degradation matrix \( H \). When \( H \) is the identity matrix, Algorithm 1 is an image denoising algorithm using iterative regularization technique [39]. When \( H \) is a blur matrix or reduced blur matrix, Eq. (23) becomes the standard formulation of non-blind image deblurring or image super-resolution problem. The capability of capturing rapidly-changing statistics in natural images - e.g., through the use of GSM - can make patch-based nonlocal image models even more powerful.

5 Experimental Results

In this section, we report our experimental results of applying BSSC-based image restoration into image denoising, image deblurring and super-resolution. The experimental setup of this work is similar to that in our previous work on NCSR [34]. The basic parameter setting of BSSC is as follows: patch size - 6 x 6, number of similar blocks - \( K = 44; k_{\text{max}} = 14, k_0 = 1 \) for image denoising, and \( k_{\text{max}} = 450, k_0 = 40 \) for image deblurring and super-resolution.

To evaluate the quality of restored images, both PSNR and SSIM [40] metrics are used. However, due to limited page space, we can only show part of the experimental results in this paper. More detailed comparisons and complete experimental results are available at the following website: http://www.csee.wvu.edu/~xinl/source.html. The source codes of this paper will be made publicly available after the publication of this paper.

5.1 Image denoising

We have compared BSSC-based image denoising method against three current state-of-the-art methods including BM3D Image Denoising with Shape-Adaptive PCA (BM3D-SAPCA) [41] (it is an enhanced version of BM3D denoising [20] in which local spatial adaptation is achieved by shape-adaptive PCA), learned simultaneous sparse coding (LSSC) [21] and nonlocally centralized sparse representation (NCSR) denoising [34]. As can be seen from Table I, the proposed BSSC has achieved highly competitive denoising performance to other leading algorithms. For the collection of 12 test images, BM3D-SAPCA and BSSC are mostly the best two performing methods - on the average, BSSC falls behind BM3D-SAPCA by less than 0.2 dB for three out of six noise levels but deliver at least comparable for the other three. We note that the complexity of BM3D-SAPCA is much higher than that of the original BM3D; by contrast, our pure Matlab implementation of BSSC algorithm (without any C-coded optimization) still runs reasonably fast. It takes around 20 seconds to denoise a 256 x 256 image on a PC with an Intel i7-2600 processor at 3.4GHz.

Figs. 2 and 3 include the visual comparison of denoising results for two typical images (lena and house) at moderate (\( \sigma_w = 20 \)) and heavy (\( \sigma_w = 100 \)) noise levels respectively. It can be observed from Fig. 2 that BM3D-SAPCA and BSSC seem to deliver the best visual quality at the moderate noise level; by contrast, restored images by LSSC and NCSR both suffer from noticeable artifacts especially around the smooth areas close to the hat. When the noise contamination is severe, the superiority of BSSC to other competing approaches is easier to justify - as can be seen from Fig. 3. BSSC achieves the most visually pleasant restoration of the house image especially when one inspect the zoomed portions of roof regions closely.

5.2 Image deblurring

We have also compared BSSC-based image deblurring and three other competing approaches in the literature: constrained total variation image deblurring (denoted by FISTA), Iterative Decoupled Deblurring BM3D (IDD-BM3D) [33] and nonlocally centralized sparse representation (NCSR) denoising [34]. Note that the IDD-BM3D and NCSR are two recently developed state-of-the-art non-blind image deblurring approaches. In our comparative study, two commonly-used
Table I The PSNR (dB) results by different denoising methods. In each cell, the results of the four denoising methods are reported in the following order: top-left-BM3D-SAPCA [41]; top-right-LSSC [21]; bottom-left-NCSR [34]; bottom-right-proposed BSSC

<table>
<thead>
<tr>
<th>σ_n</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>38.86</td>
<td>39.68</td>
<td>36.07</td>
<td>35.88</td>
<td>34.43</td>
<td>34.14</td>
</tr>
<tr>
<td>Monarch</td>
<td>38.70</td>
<td>38.85</td>
<td>35.81</td>
<td>35.96</td>
<td>34.09</td>
<td>34.23</td>
</tr>
<tr>
<td>Barbara</td>
<td>38.69</td>
<td>38.53</td>
<td>34.74</td>
<td>34.48</td>
<td>32.46</td>
<td>32.15</td>
</tr>
<tr>
<td>Boat</td>
<td>38.49</td>
<td>38.74</td>
<td>35.57</td>
<td>34.82</td>
<td>32.34</td>
<td>32.52</td>
</tr>
<tr>
<td>C. Man</td>
<td>38.38</td>
<td>38.44</td>
<td>35.07</td>
<td>34.95</td>
<td>33.27</td>
<td>32.96</td>
</tr>
<tr>
<td>Couple</td>
<td>38.36</td>
<td>38.65</td>
<td>34.98</td>
<td>35.27</td>
<td>33.02</td>
<td>33.32</td>
</tr>
<tr>
<td>F. Print</td>
<td>37.50</td>
<td>37.23</td>
<td>34.10</td>
<td>33.99</td>
<td>32.29</td>
<td>32.17</td>
</tr>
<tr>
<td>Hill</td>
<td>37.35</td>
<td>37.42</td>
<td>33.90</td>
<td>33.95</td>
<td>32.03</td>
<td>32.11</td>
</tr>
<tr>
<td>House</td>
<td>38.54</td>
<td>38.24</td>
<td>34.52</td>
<td>34.14</td>
<td>32.31</td>
<td>31.96</td>
</tr>
<tr>
<td>Man</td>
<td>38.17</td>
<td>38.39</td>
<td>34.12</td>
<td>34.28</td>
<td>31.99</td>
<td>32.03</td>
</tr>
<tr>
<td>Peppers</td>
<td>37.60</td>
<td>37.41</td>
<td>34.13</td>
<td>33.96</td>
<td>32.20</td>
<td>32.06</td>
</tr>
<tr>
<td>Straw</td>
<td>36.67</td>
<td>36.71</td>
<td>32.65</td>
<td>32.57</td>
<td>30.46</td>
<td>30.31</td>
</tr>
<tr>
<td>Average</td>
<td>37.51</td>
<td>37.16</td>
<td>33.84</td>
<td>33.68</td>
<td>32.06</td>
<td>31.89</td>
</tr>
</tbody>
</table>

5.3 Image superresolution

In our study on image super-resolution, simulated LR images are acquired from first applying a $7 \times 7$ uniform blur to the HR image, then down-sampling the blurred image by a factor of 3 along each dimension, and finally adding white Gaussian noise with $\sigma_n^2 = 25$ to the LR images. For color images, we work with the luminance channel only; simple bicubic interpolation method is applied to the upsampling of chrominance channels. Table III includes the PSNR/SSIM comparison for a set of 10 test images among four competing approaches. It can be seen that BSSC outperforms others in most situations (8 out of 10). Visual quality comparison as shown in Figs. 8 and 7 also justifies the superiority of BSSC to other SR techniques.

6 Conclusions

In this paper, we proposed a new framework named Bayesian structured sparse coding (BSSC) that connects structured sparsity with Gaussian scale mixture for image restoration.
BSSC model attempts to characterize both the biased-mean (like in NCSR) and spatially-varying variance (like in GSM) of sparse coefficients. It is shown that the BSSC problem, thanks to the power of alternating direction method of multipliers - can be decomposed into two subproblems both of which admit closed-form solutions when orthogonal basis is used. When applied to image restoration, BSSC leads to computationally efficient algorithms involving iterative shrinkage/filtering only. Our solution to BSSC can be viewed as an exemplar of demonstrating a new variational approach toward empirical Bayesian inference with parametric models. Extensive experimental results have shown that BSSC can both preserve the sharpness of edges and suppress undesirable artifacts more effectively than other competing approaches. This work clearly shows the importance of spatial adaptation regardless the underlying image model is local or nonlocal; in fact, local variations and nonlocal invariance are two sides of the same coin - one has to take both of them into account during the art of image modeling.

In addition to image restoration, BSSC can also be further studied along the line of dictionary learning. In our current implementation, we use PCA basis for its facilitating the derivation of analytical solutions. For non-unitary dictionary, we can solve the BSSC problem by reducing it to iterative reweighted $l_1$-minimization problem [38]. It is also possible to incorporate dictionary $D$ into the optimization problem formulated in Eq. (5); and from this perspective, we can view BSSD as a Bayesian generalization of K-SVD algorithm. Joint optimization of dictionary and sparse coefficients is a more difficult problem and deserves more study. Finally, it is interesting to explore the relationship of BSSC to recent advances in Bayesian nonparametrics [45],[31]. Parametric or nonparametric, we think it will eventually boils down to the capability of the model in striking an optimal tradeoff between local and nonlocal dependencies within image signals.

References

Fig. 3 Denoising performance comparison on the House image with strong noise corruption. (a) Original image; (b) Noisy image ($\sigma_n = 100$); denoised images by (c) BM3D-SAPCA [41] (PSNR=35.20 dB, SSIM=0.6767); (d) LSSC [21] (PSNR=25.63 dB, SSIM=0.7389); (e) NCSR [34] (PSNR=25.65 dB, SSIM=0.7434); (f) Proposed BSSC (PSNR=26.70, SSIM=0.7430).


### Table 2

<table>
<thead>
<tr>
<th>Images</th>
<th>Butterfly</th>
<th>Boats</th>
<th>C. Man</th>
<th>Starfish</th>
<th>Parrot</th>
<th>Lena</th>
<th>Barbara</th>
<th>Peppers</th>
<th>Leaves</th>
<th>House</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDD-BM3D [33]</td>
<td>29.21</td>
<td>31.20</td>
<td>28.56</td>
<td>29.48</td>
<td>31.06</td>
<td>29.70</td>
<td>27.98</td>
<td>29.62</td>
<td>29.38</td>
<td>34.44</td>
<td>30.06</td>
</tr>
<tr>
<td>NCSR [34]</td>
<td>0.9216</td>
<td>0.8820</td>
<td>0.8580</td>
<td>0.8640</td>
<td>0.9041</td>
<td>0.8654</td>
<td>0.8225</td>
<td>0.8422</td>
<td>0.9418</td>
<td>0.8766</td>
<td>0.8780</td>
</tr>
<tr>
<td></td>
<td>29.68</td>
<td>31.08</td>
<td>28.62</td>
<td>30.28</td>
<td>31.95</td>
<td>29.96</td>
<td>28.10</td>
<td>29.66</td>
<td>29.98</td>
<td>34.31</td>
<td>30.36</td>
</tr>
<tr>
<td>Proposed BSSC</td>
<td>0.9273</td>
<td>0.8810</td>
<td>0.8574</td>
<td>0.8807</td>
<td>0.9103</td>
<td>0.8666</td>
<td>0.8255</td>
<td>0.8402</td>
<td>0.9485</td>
<td>0.8755</td>
<td>0.8814</td>
</tr>
<tr>
<td></td>
<td>0.9377</td>
<td>0.8918</td>
<td>0.8669</td>
<td>0.8862</td>
<td>0.9145</td>
<td>0.8783</td>
<td>0.8465</td>
<td>0.8491</td>
<td>0.9582</td>
<td>0.8748</td>
<td>0.8904</td>
</tr>
</tbody>
</table>

**Table 3**

<table>
<thead>
<tr>
<th>Images</th>
<th>Butterfly</th>
<th>Parrot</th>
<th>Plants</th>
<th>Hat</th>
<th>flower</th>
<th>Racoon</th>
<th>Bike</th>
<th>Pathogen</th>
<th>Girl</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FISTA [42]</td>
<td>26.56</td>
<td>27.85</td>
<td>0.8797</td>
<td>29.20</td>
<td>27.51</td>
<td>27.54</td>
<td>23.66</td>
<td>26.00</td>
<td>31.24</td>
<td>27.88</td>
</tr>
<tr>
<td>IDD-BM3D [33]</td>
<td>0.9000</td>
<td>0.8900</td>
<td>0.8099</td>
<td>0.8483</td>
<td>0.8148</td>
<td>0.7070</td>
<td>0.7582</td>
<td>0.7232</td>
<td>0.7880</td>
<td>0.8121</td>
</tr>
<tr>
<td></td>
<td>24.70</td>
<td>28.70</td>
<td>31.55</td>
<td>29.63</td>
<td>27.87</td>
<td>28.51</td>
<td>23.23</td>
<td>26.27</td>
<td>32.87</td>
<td>28.15</td>
</tr>
<tr>
<td></td>
<td>0.8170</td>
<td>0.8323</td>
<td>0.8715</td>
<td>0.8288</td>
<td>0.7964</td>
<td>0.7273</td>
<td>0.7212</td>
<td>0.7025</td>
<td>0.8016</td>
<td>0.7998</td>
</tr>
<tr>
<td></td>
<td>0.9156</td>
<td>0.9144</td>
<td>0.9180</td>
<td>0.8699</td>
<td>0.8558</td>
<td>0.7706</td>
<td>0.8027</td>
<td>0.7506</td>
<td>0.8273</td>
<td>0.8472</td>
</tr>
<tr>
<td>NCSR [34]</td>
<td>28.10</td>
<td>30.50</td>
<td>34.00</td>
<td>31.27</td>
<td>29.50</td>
<td>29.28</td>
<td>24.74</td>
<td>27.19</td>
<td>33.65</td>
<td>29.80</td>
</tr>
<tr>
<td></td>
<td>0.9156</td>
<td>0.9144</td>
<td>0.9180</td>
<td>0.8699</td>
<td>0.8558</td>
<td>0.7706</td>
<td>0.8027</td>
<td>0.7506</td>
<td>0.8273</td>
<td>0.8472</td>
</tr>
<tr>
<td>Proposed BSSC</td>
<td>28.45</td>
<td>30.65</td>
<td>34.33</td>
<td>31.51</td>
<td>29.73</td>
<td>29.38</td>
<td>24.77</td>
<td>27.37</td>
<td>33.65</td>
<td>29.97</td>
</tr>
<tr>
<td></td>
<td>0.9272</td>
<td>0.9190</td>
<td>0.9236</td>
<td>0.8753</td>
<td>0.8638</td>
<td>0.7669</td>
<td>0.8062</td>
<td>0.7556</td>
<td>0.8236</td>
<td>0.8512</td>
</tr>
</tbody>
</table>


Fig. 4 Deblurring performance comparison on the Starfish image. (a) Original image; (b) Noisy and blurred image ($9 \times 9$ uniform blur, $\sigma_n = \sqrt{2}$); deblurred images by (c) FISTA [42] (PSNR=27.75 dB, SSIM=0.8200); (d) IDD-BM3D [33] (PSNR=29.48 dB, SSIM=0.8640); (e) NCSR [34] (PSNR=30.28 dB, SSIM=0.8807); (f) Proposed BSSC (PSNR=30.58 dB, SSIM=0.8862).
Fig. 5 Deblurring performance comparison on the Butterfly image. (a) Original image; (b) Noisy and blurred image ($9 \times 9$ uniform blur, $\sigma_n = \sqrt{2}$); deblurred images by (c) FISTA [42] (PSNR=28.37 dB, SSIM=0.9058); (d) IDD-BM3D [33] (PSNR=29.21 dB, SSIM=0.9216); (e) NCSR [34] (PSNR=29.68 dB, SSIM=0.9273); (f) Proposed BSSC (PSNR=30.45 dB, SSIM=0.9377).
Fig. 6 Deblurring performance comparison on the Barbara image. (a) Original image; (b) Noisy and blurred image (Gaussian blur, $\sigma_n = \sqrt{2}$); deblurred images by (c) FISTA [42] (PSNR=25.03 dB, SSIM=0.7377); (d) IDD-BM3D [33] (PSNR=27.19 dB, SSIM=0.8231); (e) NCSR [34] (PSNR=27.91 dB, SSIM=0.8304); (f) Proposed BSSC (PSNR=28.42 dB, SSIM=0.8462).
Fig. 7 Image super-resolution performance comparison on the \textit{Plant} image (scaling factor 3, $\sigma_n = 0$). (a) Original image; (b) Low-resolution image; reconstructed images by (c) TV [43] (PSNR=31.34 dB, SSIM=0.8797); (d) Sparsity-based [44] (PSNR=31.55 dB, SSIM=0.8964); (e) NCSR [34] (PSNR=34.00 dB, SSIM=0.9369); (f) Proposed BSSC (PSNR=34.33 dB, SSIM=0.9236).
Fig. 8 Image super-resolution performance comparison on the Hat image (scaling factor 3, $\sigma_n = 5$). (a) Original image; (b) Low-resolution image; reconstructed images by (c) TV [43] (PSNR=28.13 dB, SSIM=0.7701); (d) Sparsity-based [44] (PSNR=28.31 dB, SSIM=0.7212); (e) NCSR [34] (PSNR=29.94 dB, SSIM=0.8238); (f) Proposed BSSC (PSNR=**30.21 dB**, SSIM=**0.8354**).